Informatics in Education, 2016, Vol. 15, No. 1, 85–102

DOI: 10.15388/infedu.2016.05

Data Mining of Undergraduate Course Evaluations

Yuheng Helen JIANG, Sohail Syed JAVAAD, Lukasz GOLAB
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
e-mail: {y29jiang, sjavaad, lgolab}@uwaterloo.ca

Received: March 2015

Abstract. In this paper, we take a new look at the problem of analyzing course evaluations. We examine ten years of undergraduate course evaluations from a large Engineering faculty. To the best of our knowledge, our data set is an order of magnitude larger than those used by previous work on this topic, at over 250,000 student evaluations of over 5,000 courses taught by over 2,000 distinct instructors. We build linear regression models to study the factors affecting course and instructor appraisals, and we perform a novel information-theoretic study to determine when some classmates rate a course and/or its instructor highly but others poorly. In addition to confirming the results of previous regression studies, we report a number of new observations that can help improve teaching and course quality.

Keywords: course evaluation, entropy, regression.

1. Introduction

Course evaluations, which typically include questions about the course and the instructor and an overall appraisal, are widely used to monitor and improve teaching quality (Cronbach, 1963, Kulik, 2001, Nasser and Fresko, 2002). There is a lengthy literature on this topic, including survey design (Aleamoni, 1978, Bangert, 2004, Bangert, 2006, Cashin, 1995), and data-driven studies on which course and instructor attributes affect the overall appraisal (Badur and Mardikyan, 2011, Feldman, 1976, Bedard and Kuhn, 2008, Feldman, 1978, Marsh, 1982, Marsh, 1980, Cohen, 1981, Feldman, 2007). Multiple studies, usually on data sets containing evaluations of up to several hundred classes, have shown that high ratings are given to small classes, upper-year courses, and instructors who are experienced, enthusiastic, well-organized, and respond to questions effectively.

In this paper, we take a detailed new look at the problem of analyzing course evaluation data. We have obtained a data set from the Engineering faculty of a large Canadian university that, to the best of our knowledge, is an order of magnitude larger that those analyzed in previous work: nearly 264,000 student evaluations of 5,740 undergraduate courses taught by 2,140 distinct instructors from 2003 till 2012. The evaluations include

instructor-oriented and course-oriented questions, as well as separate overall course and instructor ratings. Our analysis includes linear regression models for predicting students' satisfaction with the course and the instructor, and a novel information-theoretic study to characterize courses whose ratings have high entropy, i.e., those which some classmates rate highly and some poorly. In addition to confirming the results of previous studies done on smaller data sets, we make a number of novel observations.

According to our regression results, *teaching quality* is highly influenced by the instructor's attitude towards teaching and visual presentation, but not affected by the availability of the instructor outside the classroom. In terms of *course appraisals*, students like challenging courses and those in which tests are a good reflection of the course material, but attributes such as the course workload and usefulness of textbooks and regularly scheduled tutorials have little effect on the course rating. Some of these findings may be Engineering-specific while others are a sign of the times: students nowadays are accustomed to finding information online, so textbooks and instructor office hours are not as critical as they once were. Interestingly, morning classes tend to be rated higher and more uniformly than evening classes.

Since our data set includes separate teaching and course appraisals, we also examine the similarities and differences between them. The overall teaching quality is easier to predict (using instructor-oriented attributes from the evaluation questionnaire) than the overall course quality (using course-oriented attributes). In fact, overall teaching quality is a good predictor of the overall course rating, reinforcing the influence of good teaching on course quality. Additionally, when an instructor teaches the *same* course for the second time, the teaching and course ratings significantly increase compared to the first time, followed by holding steady until the ninth time, and decreasing after the tenth time.

Our entropy study indicates that classmates agree more on teaching quality than the overall satisfaction with the course; in particular, classmates may have different opinions about the usefulness of textbooks and tutorials, but they tend to uniformly rate the instructor's oral presentation skills. Furthermore, the best predictor of the entropy of the overall course appraisal is the entropy of the perceived value of tests and assignments.

Also, evaluations of highly-rated courses and instructors have low entropy (most students give high ratings), whereas those of courses and instructors with low ratings have higher entropy (most students give low ratings, but some still give high ratings). As expected, the larger the class, the higher the entropy of the overall teaching quality.

Limitations: The findings of this study are based on the evaluation questionnaires. The questions in the questionnaire are designed by school and the validations of the questions are not within our scope. Furthermore, information such as student marks is not available, therefore it is not possible to correlate course performance and instructors' performance with student grades, yet students may rate their instructors or course based on the grade that they receive.

Related Work: To the best of our knowledge, no previous work considered the distribution of ratings in addition to average scores, or compared the differences between predicting teaching quality and course quality for the same set of courses. In terms of

variability of evaluations, Feldman (Feldman, 1977) studied the differences between multiple data sets from different institutions, while our entropy study characterizes the variability of classmates' opinions within the same course. In terms of linear regression studies, we obtained similar results to Feldman (Feldman, 1976), Thomas and Galambos (Thomas and Galambos, 2004), Rodriguez *et al.* (Rodríguez and Benassi, 2014), and Onwuegbuzie *et al.*(Onwuegbuzie *et al.*, 2007) (instructor's attitude), Feldman (Feldman, 1976) and Goldstein and Benassi (Goldstein and Benassi, 2006) (organization and clarity af3 fect the overall teaching evaluation); Bedard & Kuhn (Bedard and Kuhn, 2008) (teaching quality tends to be higher in smaller classes); Feldman (Feldman, 1978) and Marsh (Marsh, 1980) (upper-year courses are rated higher); Marsh (Marsh, 1982) (experienced instructors are rated higher); Kek and Stow (Kek and Stow, 2009) (classes with large sizes are rated lower) and Badur & Mardikyan (Badur and Mardikyan, 2011) (courses with high attendance are rated higher). For course-related attributes, other studies also found that the usefulness of assignments and tests is important (Feldman, 1976, Feldman, 2007, Marsh, 1980, Marsh, 1982); however, unlike (Feldman, 1976, Feldman, 2007), we found no correlation between textbook quality and course appraisal. The novel aspects of our regression analysis include detailed explanations of the effect of course level and teaching experience (in addition to the overall appraisal, which specific teaching or course related attributes are affected?), and a new study of the effect of teaching the same course multiple times on the teaching evaluation. Finally, we did not have access to grades, and therefore were unable to correlate course evaluations with student performance, as was done in (Cohen, 1981, Feldman, 2007, Thomas and Galambos, 2004).

Roadmap: Section 2 discusses our data set; Section 3 presents our analysis of teaching and course appraisals; Section 4 discusses our entropy analysis of teaching and course quality; and Section 5 concludes the paper.

2. Data

Table 1 lists the seventeen questions on our evaluation forms; we will refer to them by their abbreviations (e.g., Q1), and we will use the terms evaluation, survey and questionnaire interchangeably. Q1 through Q9 refer to teaching attributes and Q11 through Q16 refer to course attributes. Q10 and Q17 are the overall appraisals. Each question has five possible answers from A (best) to E (worst). For each question, we have the frequencies of each possible answer and an average, which is computed as follows: an A-response is assigned 100, B-response 75, C-response 50, D-response 25 and E-response zero. We also have the course code (from which we can tell whether it is a 1st, 2nd, 3rd or 4th year course), semester, and an anonymized instructor ID.

Additionally, we obtained the following attributes from online course calendars: class size, course type (compulsory or elective), time of lecture (we define morning classes as those which start before 10:00, day classes as those which start between 10:00 and 17:00, and evening classes as those which start after 17:00), and the number of lectures

Table 1
Questions on course evaluation form

Q1	Instructor's organization and clarity
Q2	Instructor's response to questions
Q3	Instructor's oral presentation
Q4	Instructor's visual presentation
Q5	Instructor's availability and approachability outside of class
Q6	Instructor's level of explanation
Q7	Instructor's encouragement to think independently
Q8	Instructor's attitude towards teaching
Q9	Professor-class relationship
Q10	**Overall appraisal of teaching quality**
Q11	Difficulty of concepts covered
Q12	Workload required to complete this course
Q13	Usefulness of textbooks
Q14	Contribution of assignments to understanding of concepts
Q15	How well tests reflect the course material
Q16	Value of tutorials
Q17	**Overall appraisal of the course**

per week (one three-hour lecture, two 90-minute lectures or three one-hour lectures). Finally, we derived the following attributes for each course offering: teaching experience of the instructor (total number of times he or she taught in the past), attendance (the number of evaluations received divided by course enrolment[1]), and *specific* teaching experience (the number of times this instructor has taught this particular course). We are not aware on any previous work that investigated the effect of specific teaching experience on course evaluations.

Following previous work (Feldman, 2007), we remove evaluations with fewer than 15 responses before doing any analysis. This leaves 257,612 evaluations of 5,150 courses taught by 2,112 distinct instructors. 29 percent of the courses are 1st-year, 27 percent 2nd-year, 22 percent 3rd-year and 23 percent 4th-year. 59 percent are compulsory (core) and 41 optional (elective). Fig. 1 shows box plots of the average scores of each question. Q10 (overall teaching quality) has a mean of 76 and a standard deviation of 16, while Q17 (overall course appraisal) has a mean of 70 and a standard deviation of 14.3. From the teaching-related attributes, oral presentation (Q3), attitude towards teaching (Q8) and professor-class relationship (Q9) have the highest average scores, while encouragement to think independently (Q7) has a noticeably lower average. From the course-related attributes, value of assignments (Q14) and tests (Q15) have high average scores, while usefulness of textbooks (Q13) and tutorials (Q16) are lower. Finally, as Table 2 shows, teaching and course appraisal averages have increased slightly from year to year since 2002, but we did not find these differences to be statistically significant according to Tukey's Honestly Significant Difference (HSD) test (Abdi and Williams, 2010).

[1] This is the attendance on the day of the course evaluation, which we assume to be the average attendance throughout the course.

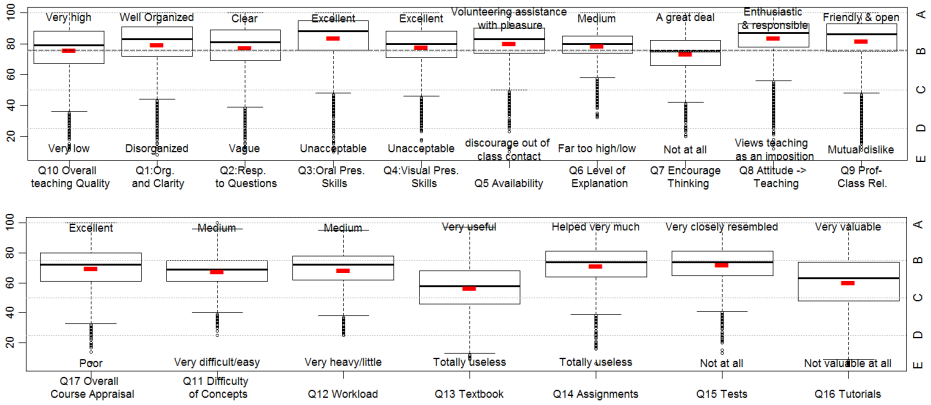

Fig. 1. Box plots for each question on the course evaluation (teaching-related questions on the top, course-related questions on the bottom).

Table 2
Average teaching (Q10) and course quality (Q17) by year from 2003 to 2012

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Q10	74.15	75.14	75.42	75.64	75.44	75.92	76.20	76.21	76.01	75.78
Q17	67.95	68.99	69.48	69.69	69.73	70.01	70.21	70.53	70.18	70.47

3. Predicting Teaching and Course Appraisals

We start with bivariate regression and detailed analysis of the effect of individual attributes on the overall teaching (Q10) and course (Q17) appraisals. Then, we build multivariate linear regression models to predict Q10 and Q17. We use the WEKA software (Hall *et al.*, 2009) and report the Root Mean Squared Error (RMSE) using 10-fold cross-validation. We then test two dimension-reduction techniques: WEKA's subset-evaluation and Principal Component Analysis (PCA), and compare the RMSE to that of linear regression on all the available attributes. Finally, we comment on the difference between predicting Q10 and Q17.

3.1. *Effect of Individual Attributes on Teaching and Course Quality*

Teaching quality (Q10) is positively correlated with each of Q1 through Q9, with Q8 (attitude) and Q4 (visual presentation skills) being strongly correlated, and Q3 (oral presentation skills) least correlated. The overall course appraisal (Q17) is positively correlated with each of Q11 through Q16, with Q15 (how well tests reflect course material) being strongly correlated, and Q12 (course workload) and Q16 (value of tutorials) least correlated. Furthermore, class size is negatively correlated with Q10 and Q17, while atten-

dance, course level and teaching experience are positively correlated. Electives are rated higher than compulsory courses. Finally, semester (fall/winter/spring) and the number of classes per week do not significantly affect teaching or course quality.

Previous work (Feldman, 1978) found that lecture times have no effects on course evaluations, but we observed morning classes to be rated higher than evening classes in our data set, and verified it to be statistically significant as per Tukey's HSD test. Upon further inspection, we observed that evening classes tend to be large and they tend to receive lower scores on organization and clarity (Q1), visual presentation (Q4) and value of assignments (Q14). This may be because students sitting in the back of a large classroom have difficulty following the instructor; moreover, large classes may have simple assignments that can be graded quickly. Also, we hypothesize that students who regularly attend early morning classes are good students who are more likely to give through course evaluations. On the other hand, many students are on campus in the evening to do their homework or to work on group projects. Since they are on campus anyway, they may attend evening classes but continue their homework in class rather than pay attention, contributing to lower satisfaction with the course.

Zooming in on the effect of class size, we found that large classes get worse oral presentation (Q3) and level of explanations (Q6) ratings, which makes sense since it may be difficult for the instructor to find the right level of explanation that will satisfy all students in a large class. On the other hand, the value of tutorials (Q16) increases, perhaps because instructors teaching large classes make an effort to organize effective tutorials, knowing that they may not be able to answer everyone's questions in class.

We obtained interesting results regarding *specific* teaching experience. 41 percent of our data consist of courses that are taught for the first time, and only 0.4 percent are taught by the same instructor more than 12 times. Fig. 2 shows box plots of the average Q10 (top) and Q17 (bottom) scores versus instructor's specific experience with the

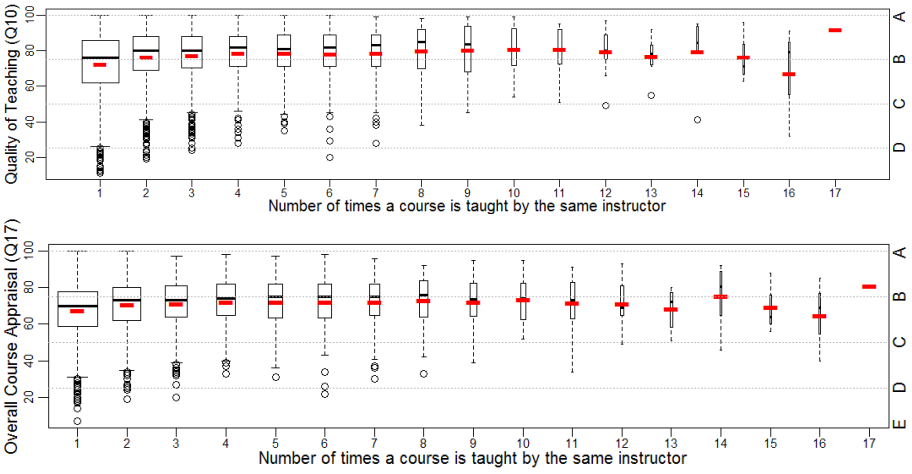

Fig. 2. Box plots for Q10 (top) and Q17 (bottom) for different numbers an instructor teaches the same course.

particular course. Instructors who teach a course for the first time clearly obtain lower ratings, which increase noticeably when they teach the same course the second time. The ratings then appear to plateau and start going down after the tenth time. Upon further inspection, we found that the scores for response to questions (Q2) and value of tests (Q15) increase with specific teaching experience, meaning that, over time, instructors get better at designing tests that accurately reflect the course material.

On the other hand, teaching appraisals tend to increase steadily with overall teaching experience. As with specific teaching experience, response to questions (Q2) improves; additionally, so do oral presentation skills (Q3), encouragement to think (Q7) and attitude towards teaching (Q8).

Interestingly, while 4th-year courses are rated significantly higher than lower-level courses (as confirmed by Tukey's HSD test), the second-highest evaluations come from first-year courses, with second and third-year courses having the lowest ratings. In particular, first and 4th year courses tend to have higher scores on organization and clarity (Q1), response to questions (Q2), and visual presentation skills (Q4). Many upper-year courses are technical electives taught by subject matter experts, which may explain their higher ratings. Additionally, universities often pay special attention to first-year courses (e.g., they may be taught by dedicated instructors with years of experience). However, more attention should be paid to improve mid-level courses.

Finally, we note that elective courses have higher ratings than core courses on the majority of attributes, including Q10 and Q17, but they score lower on workload (Q12), i.e., electives have less workload, and value of assignments (Q14), test (Q15) and tutorials (Q16). One possible explanation for this is that students have high expectations and high level of interest in elective courses, but they find the content and the tests and assignments too easy. They still rate elective courses highly, but might enjoy them even more if they were more challenging.

3.2. *Multivariate Regression*

Table 3 shows the results (regression coefficients and P-values) of linear regression to predict teaching quality (Q10) using teaching-related attributes (Q1–9) and to predict course quality (Q17) using course-related attributes (Q11–16). The RMSE values are 3.39 for Q10 and 9.77 for Q17. For teaching quality (Q10), organization and clarity (Q1) and response to questions (Q2) have the largest coefficients, whereas availability outside of class (Q5) has the smallest coefficient. For the course appraisal (Q17), the variables with the largest coefficients are how well tests reflect the course material (Q15) and the value of assignments (Q14), while the variables with the smallest coefficients are usefulness of tutorials (Q16) and textbooks (Q13). Note the semantics of the question about workload: higher score means less work in the course. Thus, the fact that the coefficient of the workload variable is negative means that Engineering students tend to rate more demanding courses higher.

Next, we build a new set of linear regression models, this time using all the questions from the course evaluations (both teaching and course related) to predict Q10 and Q17.

Table 3
Multivariate regression results

Attributes	Coefficient	P-value
Predicting Teaching Quality (Q10) Using Teaching-Related Attributes		
(Intercept)	–23.99	< 2e – 16
Q1 OrganizationClarity	0.327	< 2e – 16
Q2 ResponseToQuestions	0.271	< 2e – 16
Q3 OralPresentationSkills	0.029	7e – 09
Q4 VisualPresentationSkills	0.070	2e – 15
Q5 Availability	0.012	0.0678
Q6 LevelOfExplanation	0.068	< 2e – 16
Q7 EncourageThinking	0.195	< 2e – 16
Q8 AttitudeTowardsTeaching	0.105	< 2e – 16
Q9 ProfClassRelation	0.190	< 2e – 16
Predicting Course Quality (Q17) Using Course-Related Attributes		
(Intercept)	–4.473	0.000235
Q11 DifficultyOfConcepts	0.286	< 2e – 16
Q12 Workload	–0.115	< 2e – 16
Q13 Textbook	0.064	5.79e – 13
Q14 Assignments	0.369	< 2e – 16
Q15 Tests	0.429	<2e–16
Q16 Tutorials	0.031	0.000336

For Q10, the RMSE drops very slightly from 3.39 to 3.23, while for Q17, the RMSE drops significantly from 9.77 to 5.18. This indicates that the overall course appraisal highly depends on teaching-related attributes, not just on the course-related attributes. Finally, we added other variables to the regression models, including class size, attendance, course level and type, instructor experience, instructor specific experience, term of year and time of lecture. The RMSE was 3.2 for Q10 and 5.14 for Q17, i.e., the improvements were very small.

3.3. *Dimensionality Reduction*

We now find a small subset of good features to predict the Q10 and Q17 scores. We use two feature-reduction techniques: WEKA's *CfsSubsetEval* algorithm, which selects non-redundant features with high predictive power and low inter-correlation (Hall, 1998), and Principal Component Analysis (PCA), which constructs new features that are linear combinations of existing ones. The idea is to see if a smaller subset of features can give a linear regression model whose RMSE is nearly as low as that of the full model. Results are summarized in Table 4. The first column describes the attributes used in the linear regression model, and the second and third columns show the corresponding RMSE for predicting Q10 and Q17, respectively. The first three rows correspond to the results we described earlier: using Q1–Q9 to predict Q10 and Q11–Q16 to predict Q17 (referred to as "Related

Table 4
Summary of multivariate regression results and dimensionality reduction

	Q10 RMSE	Q17 RMSE
Related survey attributes	3.39	9.77
All survey attributes	3.22	5.18
All survey attributes + other attributes	3.20	5.14
CfsSubsetEval	3.31	5.43
PCA	3.68	5.49

survey attributes"), using Q1–Q9 and Q11–Q16 to predict Q10 and Q17 (referred to as "All survey attributes"), and adding other attributes such as course level, attendance, etc.

The *CfsSubsetEval* algorithm selected the following eight features as the best features for predicting teaching quality (Q10): instructor's organization and clarity (Q1), response to questions (Q2), visual presentation skills (Q4), encouraging students to think (Q7), attitude towards teaching (Q8), professor-class relationship (Q9), how well tests reflect the course material (Q15), and attendance. Thus, six attributes came from teaching-related questions, while one (Q15) was course-related and one (attendance) was a derived attribute from other data. As Table 4 shows, using these eight features to build a linear regression model gave a RMSE that was not much higher than using all available features. For predicting the overall course appraisal (Q17), the eight best features were: instructor's response to questions (Q2), visual presentation skills (Q4), encouraging students to think (Q7), professor-class relationship (Q9), difficulty level (Q11), usefulness of assignments (Q14), how well tests reflect the course material (Q15), and attendance.

Interestingly, six of these are the same as those used for Q10. Again, the RMSE of a linear regression model using only these eight features was not much higher than using all available features.

Next, we use WEKA's PCA algorithm to create different numbers of linear combinations of features, from one to eight, and use them to build linear regression models for predicting Q10 and Q17. We observed that the RMSE kept dropping significantly up to five principal components, and much less so for six or more components. Thus, for this particular task, five principal components appear to give the best tradeoff between the number of features and their predictive power. They are listed in Table 5; note that the first component is a linear combination of attributes that we previously showed to be highly correlated with teaching and course quality. As shown in Table 4, the RMSE of the linear regression model with five linear combinations of features obtained via PCA is not much higher than the one that uses the eight features suggested by *CfsSubsetEval*.

3.4. *Predicting Teaching vs. Course Quality*

An interesting aspect of our course evaluations is that they contain a separate teaching quality rating (Q10) and an overall course quality rating (Q17) rather than just a single overall satisfaction rating. In this section, we contrast these two variables. First, in Fig. 3

Table 5
Five principal components extracted from all survey attributes and other attributes

1	0.35 ResponseToQuestions + 0.35ProfClassRelationship + 0.34Attitude + 0.33 OrganizationClarity + 0.33 VisualPresentationSkills
2	0.47 CourseLevel – 0.43 ClassSize + 0.41 CourseType – 0.32 LecturesPerWeek – 0.28 UsefulnessOfTutorials
3	– 0.51 Workload – 0.49 Difficulty – 0.36 LevelOfExplanations + 0.3 EncouragesThinking – 0.28 Tests
4	0.49 UsefulnessOfTutorials + 0.45 UsefulnessOfAssignments – 0.32 ClassSize + 0.32 UsefulnessOfTextbook + 0.3 Attendance
5	– 0.58 Semester – 0.56 Attendance – 0.41 TimeOfClass + 0.24 LecturesPerWeek + 0.23 CourseLevel

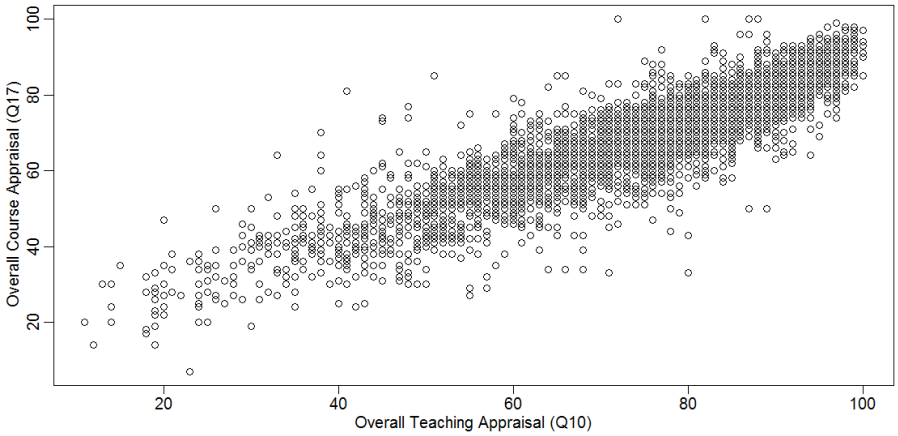

Fig. 3. Scatter plot of teaching quality (Q10) vs overall course appraisal (Q17).

we show a scatter plot of the Q17 scores on the y-axis and the Q10 scores on the x-axis. The relationship is clear, but there are a few outliers, mainly corresponding to highly-rated instructors teaching a poorly-rated course. Upon further inspection, we found that these outliers are hands-on design workshop courses, and the reason for their low course rating was a very high workload.

We also make two observations based on the results in Table 4. First, the RMSE numbers indicate that the overall course appraisal is harder to predict that the overall teaching appraisal. This makes sense since specific teaching skills such as organization, clarity and enthusiasm are easier to evaluate. Since there are only six course-oriented questions and nine teaching-oriented ones, one suggestion is to add more course-oriented question to the evaluation form, for example, the students' prior interest in the subject, whether course was graded fairly and in a timely manner, or whether students feel that the course content is practical and up-to-date (some of these questions appear on the course evaluation forms of other institutions).

The other observation is that the RMSE of the linear regression model for predicting Q17 drops significantly after adding teaching-oriented attributes. In fact, we also tested

a linear regression model for predicting Q17 only based on Q10, and obtained a RMSE of 6.14[2]. That is, the overall teaching quality is a better predictor of the course appraisal than the six course-related attributes. Furthermore, based on the results using *CfsSubsetEval* described earlier, there is a common set of six attributes (some course-related, some teaching-related, plus attendance) that are highly correlated with both Q10 and Q17. Again, this indicates that the overall course appraisal is highly influenced by the quality of the instructor.

4. Entropy Analysis

In this section, rather than predicting the teaching quality and course appraisal scores, we investigate the distribution of the answers to Q10 and Q17. For each course offering, we compute the entropy of each of the 17 questions as follows. Let p_A, p_B, p_C, p_D and p_E be the relative fractions of the students who chose options A, B, C, D and E, respectively. Then the entropy is

$$-p_A \log_2 p_A - p_B \log_2 p_B - p_C \log_2 p_C - p_D \log_2 p_D - p_E \log_2 p_E \ .$$

Higher entropy means that there is more variability in the responses among the students in a given class. For example, a course appraisal with 10 percent A's, 20 percent B's, 40 percent C's, 15 percent D's and 5 percent E's has higher entropy and more variability than one with 70 percent A's and 30 percent B's.

We start with box plots of the entropy of each survey question in Fig. 4. Aggregated over all the courses in our data set, the average entropy of Q17, at 1.63, is higher than that of the Q10, at 1.47. According to the t-test, this difference is statistically significant. This indicates that classmates agree more on the teaching quality than the overall course

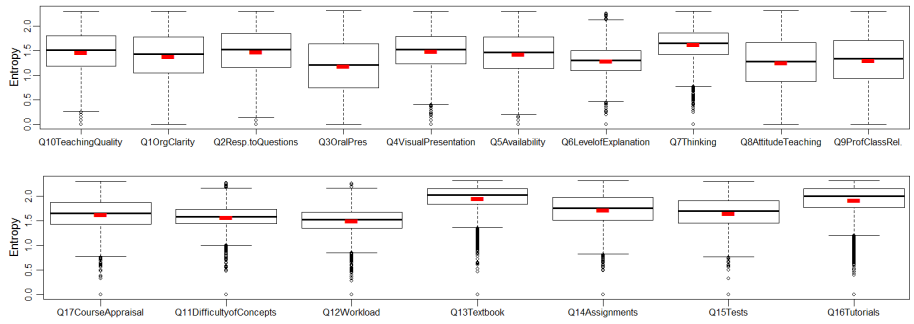

Fig. 4. Box plots for the entropy of each survey question.

[2] Conversely, predicting Q10 only based on Q17 gave a linear regression model with a RMSE of 6.96, meaning that it is more accurate to predict teaching quality based on teaching-related survey questions rather than on the overall course appraisal alone.

quality. This observation is consistent with our earlier observation that teaching quality (Q10) is easier to predict (using the available attributes) than the overall course appraisal (Q17).

Of the teaching-related questions, quality of oral presentation (Q3) has the lowest entropy of 1.13, which makes sense: good or bad speakers are uniformly perceived as such. Encouragement to think independently (Q7) has the highest entropy, which also makes sense since different students may be interested in different topics or aspects of a course. Of the course-related questions, usefulness of textbooks (Q13) and usefulness of tutorials (Q16) clearly have the highest entropy, of 1.95 and 1.92, respectively. This is likely due to the different learning styles of different students: some learn on their own and/or from lectures, while others need a good textbook or effective tutorials. Workload (Q12) has the lowest entropy, which makes sense: e.g., a heavy course is perceived as heavy by the majority of students.

4.1. *Predicting the Entropy of Q10 and Q17*

We now turn our attention to predicting the entropy of Q10 and Q17 using linear regression. We compute the RMSE of three models, similar to those used in the Section 3, but using entropy, not average score, as features. First, we predict the entropy of Q10 and Q17 using only the entropy of the teaching or course-related survey attributes, respectively ("Related survey attributes"). Next, we use the entropy of all survey attributes ("All survey attributes"), followed by adding the values of other attributes such as class size, instructor experience, etc. Results are summarized in Table 6 and discussed below.

The entropy of teaching quality ratings (Q10) is explained by the entropy of the teaching-related survey questions (Q1–Q9); adding other attributes to the model does not improve the RMSE. The entropy of response to questions (Q2) and organization and clarity (Q1) had the largest regression coefficients of 0.28 and 0.27, respectively, whereas entropy of oral presentation (Q3) had the smallest coefficient of –0.03. We conclude that classmates disagree on the overall teaching quality largely because they disagree on the organization and clarity of the instructor or his or her effectiveness in responding to questions.

The entropy of the overall course appraisal (Q17) can be explained by the entropy of all the survey questions, both teaching-related and course-related (using only the course-related questions has a higher RMSE, once again confirming the fact that teaching qua-

Table 6
Multivariate regression results for the entropy of Q10 and Q17

	Q10 RMSE	Q17 RMSE
Related survey attributes	0.15	0.24
All survey attributes	0.15	0.19
All survey attributes + other attributes	0.15	0.19

lity significantly influences the overall course appraisal). In particular, the entropy of usefulness of assignments (Q14) had the largest regression coefficient of 0.35, whereas the entropy of usefulness of tutorials (Q16) had the smallest coefficient of 0.01. This suggests that if classmates disagree on the overall course appraisal, they do so because some enjoyed working on the assignments but others did not. On the other hand, disagreement in the rating of tutorials does not lead to disagreement in the overall rating of the course. One possible explanation for this result is that students who do not find tutorials useful may choose not attend them, and if they like other aspects of the course, they will still rate it highly.

Fig. 5 shows a scatter plot of the entropy of Q10 and entropy of Q17, which appear to be linearly correlated. There are a few outliers with a Q10 entropy between 1 and 2, but zero entropy of Q17. Upon further inspection, we discovered that these courses are unanimously rated A thanks to their excellent tests (Q15) and assignments (Q15) (also unanimously rated A). Thus, despite some natural variability in their teaching quality scores, their overall appraisal was unanimously excellent.

4.2. *Effect of Other Attributes on Entropy of Q10 and Q17*

As for the other attributes besides the survey questions, the number of lectures per week is not significantly correlated with the entropy of Q10 or Q17. On the other hand, the type of course matters in an interesting way: optional courses tend to have higher entropy of teaching quality, but lower entropy of course quality. One way to explain this is as follows. Students who sign up for an optional course are interested in the material and may rate the course uniformly well, regardless of how the course actually turns out. At the same time, some of these students may rate the instructor more highly than they normally would have, just because they liked the topic of the course, while others may rate the instructor normally.

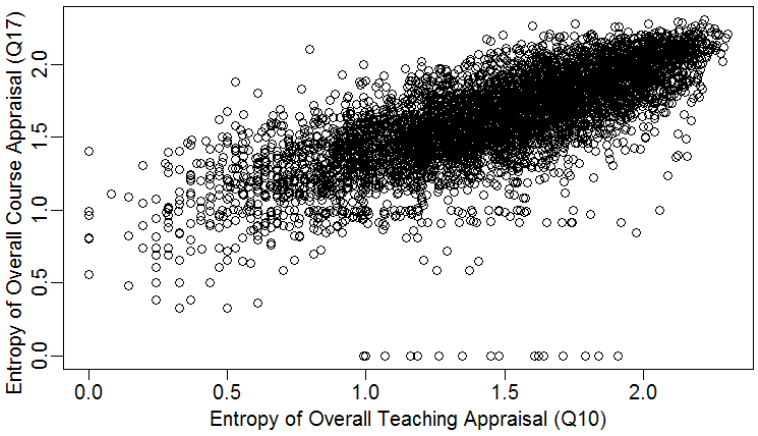

Fig. 5. Scatter plot of entropy of teaching quality (Q10) vs entropy of overall course appraisal (Q17).

In terms of the time of lecture, evening classes have higher entropy of their appraisals. As we mentioned earlier, some students who attend evening classes do not pay attention to the instruction and may sit in the back of the classroom and do their homework. Such students may give lower ratings than those who pay attention. On the other hand, students who make an effort to wake up early and attend morning classes tend to pay attention and provide more consistent feedback.

Teaching experience is slightly negatively correlated with the entropy of teaching and course scores, meaning that experienced instructors tend to be rated more uniformly. Furthermore, the entropy of Q10 is higher for instructors teaching a particular course for the first time and for those who taught the same course more than ten times. This is likely because the overall appraisal is lower in these cases, and lower appraisals tend to also have higher entropy (as we will discuss in Section 4.3).

Class size is positively correlated only with the entropy of teaching quality, while attendance is negatively related. This makes sense, as having more students naturally leads to more diverse opinions of teaching quality.

Finally, in terms of the course level, the entropy of the overall course appraisal is lower in first year, and then it increases significantly in the second and third years, and drops in the fourth year. The increase from first year might be because as students take more courses, they develop a better idea of what they like and do not like in a course, and as a result they express stronger opinions on their evaluations. The fourth-year drop is likely due to the fact that many fourth-year courses are optional, which, as discussed above, have lower course appraisal entropy.

4.3. *Detailed Analysis of the Distribution of Responses to Q10 and Q17*

Recall that the course evaluations studied in this paper have five possible answers for each question, ranging from A (best) to E (worst). Note that entropy analysis does not fully capture the polarity of opinions expressed by different students in the same class. For example, a course appraisal with 50 percent A's and 50 percent B's (and no other ratings) has the same entropy as an appraisal with 50 percent A's and 50 percent E's (and no other ratings). Clearly, the latter is more "controversial" as some students love it and others hate it. Motivated by this observation, we now further investigate how the responses to Q10 and Q17 are distributed over the five possible options.

We begin by plotting the entropy of Q10 and Q17 versus the Q10 and Q17 scores in Fig. 6. Highly-rated courses have low entropy – mostly A's and perhaps a few B's. Poorly-rated courses have high entropy, meaning that they may have a non-zero number of all five possible responses. This suggests that good courses and instructors are rated highly by the majority of students, but mediocre ones may be rated highly or poorly, depending on the student.

To validate this hypothesis, Fig. 7 plots the relative percentage of teaching and course appraisals with "no gaps" for different ranges of the Q10 and Q17 scores. We informally define a teaching or course appraisal (Q10 or Q17) with no gaps as one that has at least one of every possible option (A through E). Intuitively, courses with no gaps elicit the most variable opinions, ranging from best (A) to worst (E).

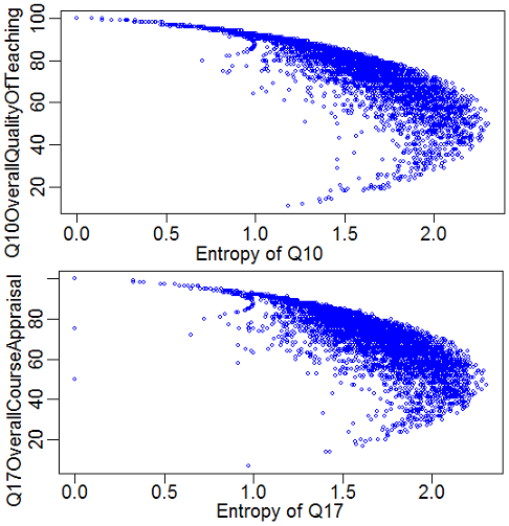

Fig. 6. Scatter plot of Q10 (top) and Q17 (bottom) scores versus their entropy.

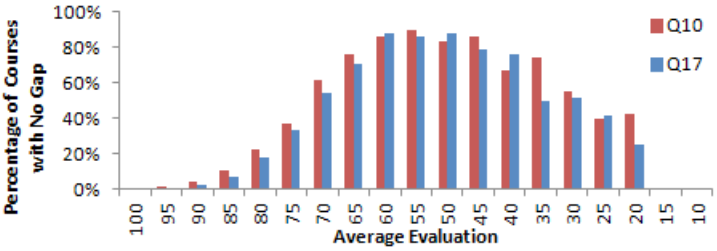

Fig. 7. Relative percentage of Q10/Q17 appraisals with no gaps (i.e., at least one A, B, C, D and E response) for different ranges of the Q10/Q17 averages.

As Fig. 7 shows, many courses rated between 50 and 60 contain no gaps, meaning that the average appraisal is a C, but there is also at least one A, B, D and E. In other words, these poorly-rated courses are rated very highly and very poorly by at least some students. More surprisingly, even some courses rated as poorly as 20 have no gaps (some students liked them), as do some courses rated as highly as 80 (some students hated them). One possible explanation for the former is that some students in bad courses may not take the evaluations seriously and they will simply choose the first answer for every question – which happens to be A – so they can complete the questionnaire as soon as possible and leave. If true, this means that the real average appraisal of such courses is even lower than reported. For the latter, we hypothesize that even highly-rated courses may have a handful of unhappy students for various reasons.

Finally, we zoom in on courses with the most extreme distributions of Q10 and Q17 ratings. There are no courses whose appraisals only contain A's and E's, and no other ratings in between. However, there are 13 courses whose teaching appraisals only have A's,

B's and E's, and no C's and D's. The teaching quality scores of these 13 courses range from 76 to 96. Thus, these are courses that obtained mostly A and B ratings, with only a few E's. Digging deeper, we noticed that the lowest-rated questions for these courses are encouraging to think independently (Q7) and how well test reflect the course material (Q15); both of these contained many D's and E's. We hypothesize that these courses had good instructors but poorly-designed tests (or perhaps unfairly-graded tests that did not reward independent thinking); most students rated the instructor highly despite the problems with tests, but a few may have found these problems so serious that they felt the instructor deserves to be rated poorly.

5. Conclusions

In this paper, we presented our methodology and results of analyzing a large set of undergraduate course evaluations from an Engineering faculty of a major Canadian university. Our regression analysis in Section 3 revealed similar results to those obtained in previous work (using smaller data sets from other institutions), and new insights into the learning strategies of students nowadays, the effect of lecture times, teaching the same course multiple times, course year and course type. We also presented a novel information-theoretic study on the distribution of responses to the course evaluation questions, which suggested the reasons why classmates may rate a given course and instructor differently, and discovered that some bad courses are still rated highly by some students.

Based on our analysis, in order to improve the teaching quality, instructors should consider enhance their attitude, and organization visual presentation skills. They want to make sure that they respond questions well and clearly. In order to improve the course quality, instructors may want to design tests and assignments such that they are closely related to the course material. Due to the low evaluations on the usefulness of textbooks and tutorials, instructors may consider improving the quality of textbook and tutorials. From a institution's perspective, it should try to schedule more morning classes with smaller number of students since evening classes and classes with large size receive worse evaluations. Furthermore, an instructor may consider discontinue teaching a course after the tenth time due to the declining ratings.

We have two directions in mind for future work. One is to validate the hypothesis we made in Section 4.3 regarding the reason why poorly-rated courses often have some students that rate them highly (the highest rating appears as the first option). One possible experiment is to reverse the order of possible options on a random sample of evaluations; another possibility is to compare our results to those from a different institution whose evaluation questionnaires have the possible options listed in reverse order of ours, if one exists. The other future work direction is to analyze the free-text comments that students make in their course evaluations to help understand specific reasons behind students' ratings. One problem is that these comments are not recorded by the Faculty, and therefore instructors would have to voluntarily share them, which may result in a biased sample.

References

Abdi, H., Williams, L.J. (2010). Tukey's honestly significant difference (HSD) test. *Encyclopedia of Research Design*, 1–5.

Aleamoni, L.M. (1978). Development and factorial validation of the Arizona course/instructor evaluation questionnaire. *Educational and Psychological Measurement*, 38(4), 1063–1067.

Badur, B., Mardikyan, S. (2011). Analyzing teaching performance of instructors using data mining techniques. *Informatics in Education*, 10(2), 245–257.

Bangert, A.W. (2004). The seven principles of good practice: a framework for evaluating on-line teaching. *The Internet and Higher Education*, 7(3), 217–232.

Bangert, A.W. (2006). The development of an instrument for assessing online teaching effectiveness. *Journal of Educational Computing Research*, 35(3), 227–244.

Bedard, K., Kuhn, P. (2008). Where class size really matters: class size and student ratings of instructor effectiveness. *Economics of Education Review*, 27(3), 253–265.

Cashin, W.E. (1995). Student ratings of teaching: the research revisited. idea paper no. 32. *Kansas State University, Center for Faculty Evaluation and Development*.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Review of Educational Research*, 51(3), 281–309.

Cronbach, L. (1963). Course improvements through evaluation. *The Teachers College Record*, 64(8), 672–672.

Feldman, K.A. (1976). The superior college teacher from the students' view. *Research in Higher Education*, 5(3), 243–288.

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses: a review and analysis. *Research in Higher Education*, 6(3), 223–274.

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: what we know and what we don't. *Research in Higher Education*, 9(3), 199–242.

Feldman, K.A. (2007). Identifying exemplary teachers and teaching: evidence from student ratings. In: *The scholarship of teaching and learning in higher education: An evidence-based perspective*. 93–143.

Goldstein, G., Benassi, V. (2006). Students' and instructors' beliefs about excellent lecturers and discussion leaders. *Research in Higher Education*, 47(6), 685–707.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. (2009). The weka data mining software: an update. *SIGKDD Explorations*, 11(1), 10–18.

Hall, M.A. (1998). *Correlation-Based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, New Zealand.

Kek, M., Stow, S. (2009). What makes students happy? Factors influencing student engagement using student evaluation data. In: *ALTC First Year Experience Curriculum Design Symposium 2009: FYE Showcase Abstracts*. 59–63.

Kulik, J.A. (2001). Student ratings: validity, utility, and controversy. *New directions for institutional research*, 2001(109), 9–25.

Marsh, H.W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal*, 17(2), 219–237.

Marsh, H.W. (1982). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement*, 6(1), 47–59.

Nasser, F., Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, 27(2), 187–198.

Onwuegbuzie, A., Witcher, A., Collins, K., Filer, J., Wiedmaier, C., Moore, C. (2007). Students perceptions of characteristics of effective college teachers: a validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, 44(1), 113–160.

Rodríguez, A., Joan-Lluís, C., Víctor M., G. (2014). Teaching performance: determinants of the student assessment. *Academia Revista Latinoamericana de Administración*, 27(3), 402–418.

Thomas, E., Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, 45(3), 251–269.

Y. Jiang is currently an associate data scientist at Capital One Canada. She obtained a B.A.Sc. degree in Management Engineering in 2013 and a M.A.Sc. degree in Management Sciences in 2015 from the University of Waterloo. Her research interests include data analysis and mining in education and energy systems.

S.S Javaad is currently the Head of Payment Systems Department at the State Bank of Pakistan. He obtained a M.A.Sc. degree in Management Sciences from the University of Waterloo in 2013.

L. Golab is an Assistant Professor in the Department of Management Sciences at the University of Waterloo, and a Canada Research Chair. He obtained a B.Sc. in Computer Science from the University of Toronto in 2001 and a Ph.D. in Computer Science from the University of Waterloo in 2006, winning the Alumni Gold Medal for top Ph.D. graduate. His research interests include data analytics and database systems.