International Journal of Engineering Education Vol. 31, No. 6(A), pp. 1476–1490, 2015
Printed in Great Britain

0949-149X/91 $3.00+0.00

Comparison of Student Evaluation of Teaching Results when Stratified by Protocol, Course Content, and Course Structure*

CHRISTOPHER R. DENNISON, ROBERT BUTZ, R. SHAWN FUHRER and JASON P. CAREY
Department of Mechanical Engineering, University of Alberta, Edmonton AB, Canada. E-mail: cdenniso@ualberta.ca and jpcarey@ualberta.ca

Focusing on the mechanical engineering undergraduate program at the University of Alberta, this study attempts to quantify biases in student evaluation of teaching (SET) results that could be attributed to SET protocol, course content, and course delivery mode. SET results were compiled for five academic years of paper based SET evaluation and one semester of online SET evaluation. 20 core undergraduate courses were included; class size from 70–130; 35 professors. Statistical analysis included compilation of frequency histograms, determination of means and standard deviations, and rank-sum tests for significant differences based on aggregated data for several stratifications. Results showed significantly reduced response rate for online SET when compared to paper; ratings of professor evaluation were not different. No significant differences were found when results were compared on the basis of course content or delivery mode. Our aggregated data showed SET protocol lead to lower response rate, but not significant differences in instructor evaluation. Course content and delivery mode did not manifest in significant changes in SET results. Typical variability in instructor rating was 0.4/5.0 considering all data. Administrators and senior faculty should be aware of these results when ascertaining instructor performance. Although focused on one department, the study is a first step in a larger evaluation of SET in engineering. The study identified key variables that must be further evaluated.

Keywords: universal student rating of instruction; student evaluation of teaching; engineering; education; measurement

1. Introduction

Student evaluation of teaching (SET) has been used as a metric to arguably evaluate instructor effectiveness since the 1920s. SET is typically used in decisions regarding yearly evaluation and for tenure and promotion decisions. Partially due to the central role SET plays in assessment and promotion, it is one of the most researched topics in personnel evaluation. Principal foci in SET research are concerned with: validity of results; factors influencing bias; and correlations between student grades and instructor rating. In an overall sense, central questions that previous research seeks to answer is whether metrics associated with SET are appropriate measures of teaching effectiveness and whether SET actually leads to improved teaching and quality of graduates [1, 2].

Recurring areas of research in SET include: (1) administration of evaluations: anonymity, timing, instructor presence; (2) class characteristics: size, selectivity; (3) instructor characteristics: gender, etc.; (4) student characteristics: age etc.; and (5) reaction to the use of evaluations [3]. Among these themes, and important from the perspective of the researcher and institutions, questions surrounding validity of results and bias in evaluations are often investigated [1, 4]. While the opinions to these questions are varied, a consistent conclusion drawn is that when "properly" designed (psychometrically valid), administered, and interpreted, SET results can be reliable measures to indicate teaching quality [5]. To properly administer and interpret/apply SET, a clear understanding of biases in results, among other parameters, is important. The research on factors of SET bias suggests that many factors can influence SET results. A recent review from the University of Alberta suggests that the mechanics of administering SET can influence bias and that whenever possible, consistency in administration should be insured [1]. This suggests that changes in SET protocol could represent a source of bias and therefore should be considered when interpreting results.

In the fall semester of 2013, the University of Alberta discontinued administering paper-based SET in favor of an online (web-based) protocol. The historic paper-based SET protocol involved distribution, in class time, of paper surveys to students and was financially costly. In the online protocol, students complete the survey via their university email account (outside of class time). The transition from paper-based and in class SET to online provides a unique opportunity to investigate changes in SET response rate and ratings of overall instructor effectiveness that could be attributed to change in protocol and that could suggest protocol-related bias. Furthermore, the authors

* Accepted 15 July 2015.

took this opportunity to examine SET results from their home department and quantify differences in SET results when stratified along course content (solid mechanics, thermo-fluids etc.) and delivery method (laboratory, traditional lecture and project based design).

Following the 2013 online-based SET evaluations, the Faculty of Engineering returned to two consecutive terms of paper-based SET evaluations; however, in fall 2014, the University of Alberta, mandated that all future SET evaluations be online.

Our overall objective was to quantify changes in response rate and instructor effectiveness from data and structure research questions for future work; this is of primary importance since the University has decided to endorse online SET as the only acceptable process. In this paper, we compare response rate and universal student rating of instruction (USRI) score for question 221 (Q221, overall effectiveness of instructor) for the past five years to response rate and Q221 score for the first web-based SET (fall 2013) in various contexts. Due to the characteristics of our available data-set, we cannot perform statistical tests in all cases and therefore our results and discourse will be context specific, but will focus on the overall questions for our home department:

1. Is there a difference in USRI scores and response rates between paper- and web-based assessments? Is there a correlation between response rate and USRI score?
2. Are there differences in USRI scores and response rates between different course types (labs, design, lecture based, miscellaneous)?
3. Does the subject matter (solid mechanics, thermo-fluids, etc) influence USRI scores and response rates?
4. Are there differences in USRI scores and response rates between summer and traditional fall/winter terms?
5. What is the variability in course USRI scores and response rates? Does this vary yearly?
6. How do variability in USRI results on a per participant basis compare to variability in aggregated data?

2. Materials and methods

2.1 Study design

This study focused on all core courses in the Mechanical Engineering Department at the University of Alberta, that all students must take as part of their undergraduate degree, coded: MecE 200, 230, 250, 260, 265, 300, 301, 330 or 331, 340, 360, 362, 370 or 371, 380, 390, 403, 451, 460, 463. Description of these courses can be found in the UofA calendar[1]. The preceding list includes 13 lecture based courses (127 SET datasets; comprising 115 paper based and 12 online based), 5 design courses (68 SET datasets; comprising 63 paper based and 5 online based), and 2 laboratory courses (21 SET datasets; comprising 19 paper based and 2 online based).

SET data collected was the response rate (% of total class that responded to questionnaire) and question 221 score (Q221 score) of the questionnaire: ''overall I find the instructor excellent''. Students completing the survey have five options based on a Likert scale for agreeing/disagreeing that their instructor was excellent:

- 1 for ''Strongly disagree''
- 2 for ''Disagree''
- 3 for ''Neither agree or disagree''
- 4 for ''Agree''
- 5 for ''strongly agree''

Data from the paper-based SET reviews from the past 5 years (winter 2008 to summer 2013), was collected and compared with results of the Fall 2013 term web-based SET results. In total, 197 paper-based SET datasets and 19 online-based SET datasets were included in this work.

Ethics approval was received from the University of Alberta Ethics board (Study ID Pro00045934) for this study; participant anonymity was preserved by coding the data *a priori* as well as blinding the corresponding authors to all participant information and course identifiers. The remaining authors (Fuhrer and Butz) are graduate trainees in Mechanical Engineering and as such have open access to all SET data at the University of Alberta. Therefore, Fuhrer and Butz compiled all data, per ethics approval, blinded data, and generated all statistics described in this work.

2.2 Participants and data

All faculty-based instructors in the department were invited to participate, 35 participated in this study. Data does not include short-term contract instructors. The department represents the gamut of junior to senior faculty, at all ranks, represents a number of ethnicities, but only includes two female professors; Mechanical Engineering is known to have the poor gender balance [6]. The data includes responses for one long serving Faculty Service Officer[2]. The department is fairly junior with 11 Assistant Professors, 14 Associate Professors, all but two, hired in the last 10 years and progressing normally through

[1] http://www.registrarsoffice.ualberta.ca/Registration-and-Courses/Courses-Listings.aspx

[2] At UofA Faculty Service Officers are members of the faculty but whose job description focuses principally in support of teaching and/or research and/or service to the community.

the ranks, and 17 Professors. Because of the three program streams in our department, (traditional, cooperative education program (co-op) I and co-op II) class sizes range from 70 to 130.

2.3 Data analysis

Data compilation, blinding of data, and presentation in this work:

All data presented is randomly assigned a unique identifier to prevent participant recognition. Participant identifiers were assigned by creating an alphabetized list of the participants arranged in ascending order by first name. Subsequently using the Excel 2010 Data Analysis package to generate a list of random numbers from 1–1000 with a uniform distribution. The list was then sorted by the random number column in ascending order, and in this configuration whole numbers from 1–35 were assigned to each participant in ascending order.

Every SET dataset was assigned a unique identification number by sorting the 216 datasets by the unique participant identifier numbers in ascending numerical order, and again using the Excel Data Analysis package, a list of random numbers were generated from 1–1000 with a uniform distribution. The database was then sorted by the random number list in descending order and assigned whole numbers from 1–216 in ascending order.

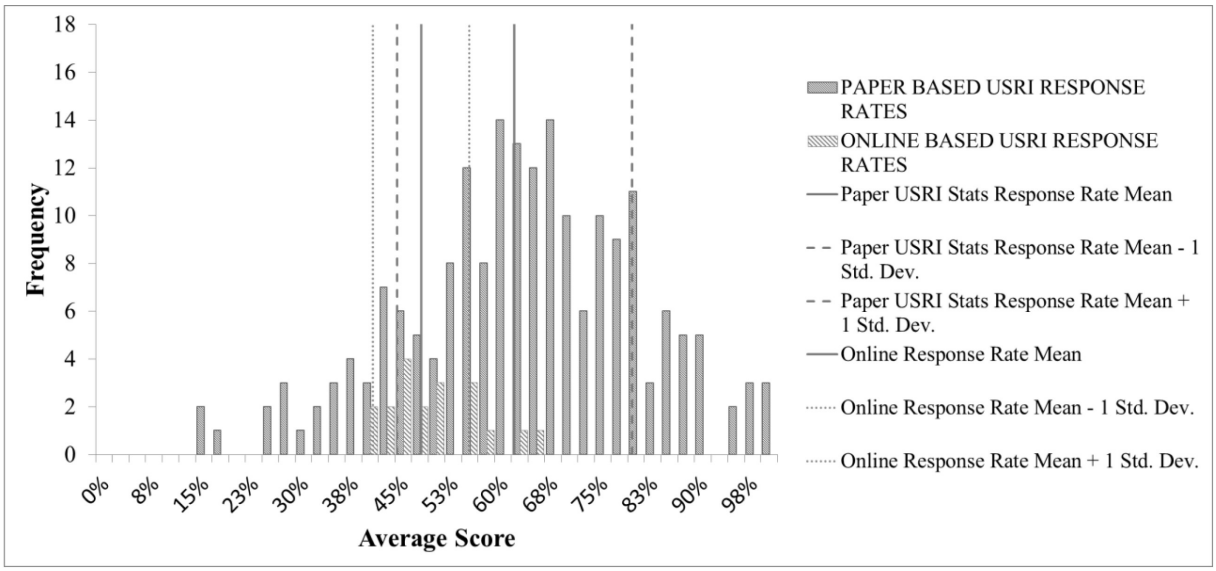


Fig. 1. Histogram for response rate. Paper-based data and online are both shown. Solid vertical lines indicate means for paper and online data. Vertical dashed lines indicate first standard deviation.

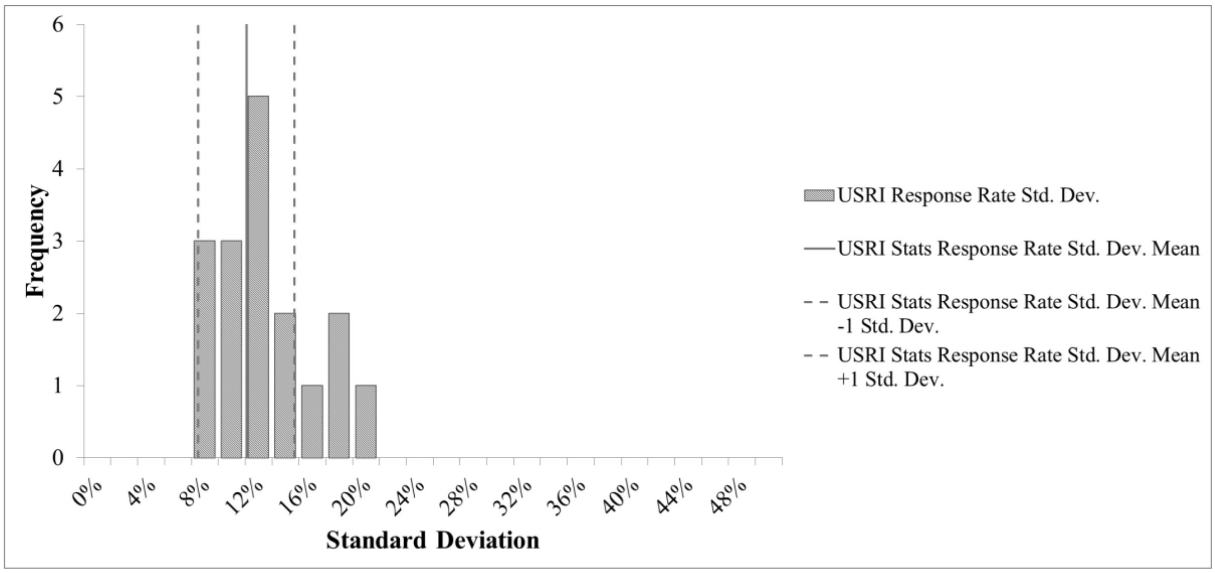


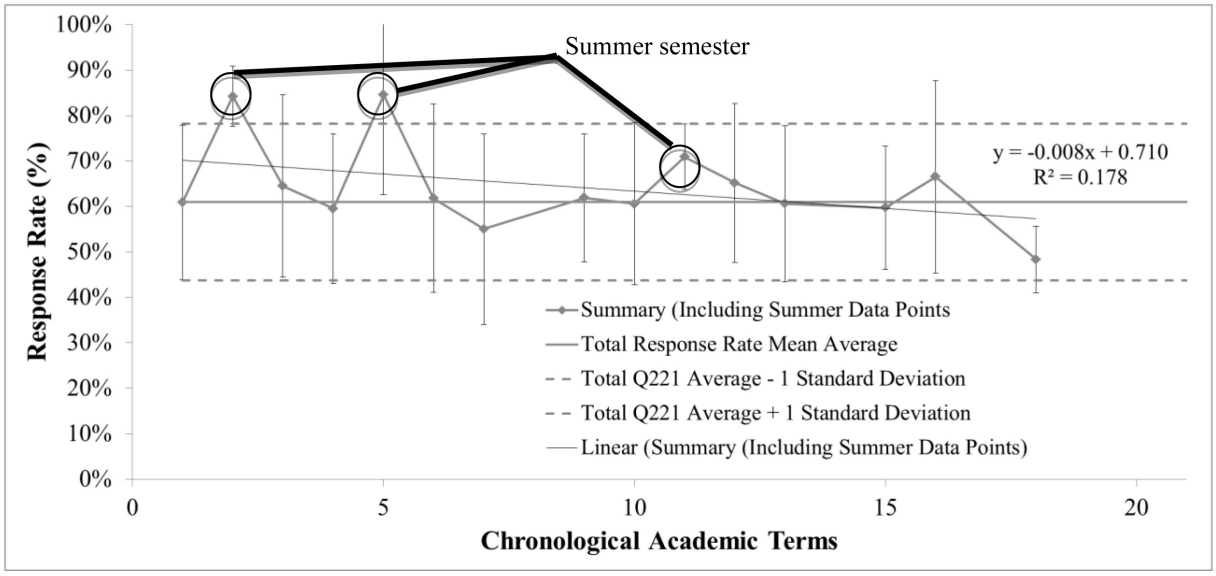
Fig. 2. Histogram showing frequency of the standard deviation in response rate for all courses. Bar heights for histogram based on all course data (both online and paper-based SET).

For individual course code numbers, the list of courses included in this study was first assigned the correct course code letter, acronyms are:

- DC: Design courses
- Misc: Numerical Methods, and Measurements
- SM: Solid mechanics
- TC: Technical Communications
- TF: Thermo-fluids

Courses were then sorted by course number, and a random number set was generated with a uniform distribution from 1–500. The list was then sorted first by course code letter (in ascending order), then by random number (in descending order). Courses were then assigned a whole numbers in ascending order for the given course code.

In some cases multiple courses were grouped under one identifier. This was because the courses focus on similar subject matter and the courses are part of recent departmental course reorganizations within the 5 year review. The courses in these groups were assigned their identifiers by the same method as described above, however after the first course in the group was assigned its identification number, the rest of the courses in that group were assigned that same number, and thereby they were removed from being assigned a different ID number when continuing to assign numbers to the remainder of the course offerings for that given course code.



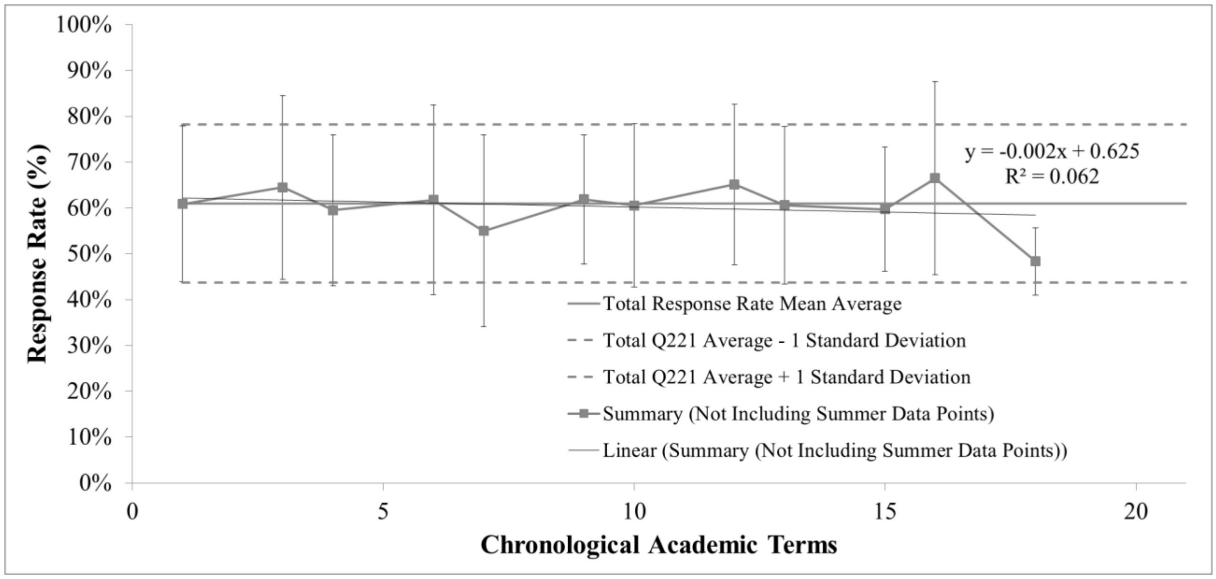


Fig. 3. (top) mean response rate (solid points) and standard deviation (error bar) for chronological semester. (bottom) as above, but with data for summer semesters omitted.

Statistics:
Descriptive statistics (mean and standard deviations) were computed for all paper-based data for each course. In some cases, these statistics were computed based on aggregated data for course-type (e.g. solid mechanics—SM, thermo-fluids—TF). Rank sum tests (sometimes referred to as Mann-Whitney U-tests) were also performed to test for significant differences in results when stratified by protocol, course content, course delivery method etc.

3. Results

3.1 Response rate

Figure 1 is a histogram for response rate, for both online-SET and paper-based SET data. As shown, the mean response rate for online and paper is 47% and 62%, respectively. The online response rates indicate a standard deviation of 8%, while the paper data indicates 20%.

Figure 2 is a histogram showing frequency of standard deviation in response rate (for all data), and indicates the standard deviation ranged from 8% to 22%. The mean deviation (considering all course data) is 12% (Fig. 2).

Figure 3 shows the average response rate data, ordered chronologically, for the past 5 years. In the plots, each data point is the mean value of the all the data points for each given term. Error bars on each data point are ±1 standard deviation. The mean of all data points (inclusive of online and paper-based data, all 218 points without summer terms) and ±1 standard deviation are shown as horizontal lines. These graphs do not distinguish between online and paper collection for overall mean; however, term 18, the last data point to the right, is the semester where USRI was administered online. The top plot includes summer terms. Three summer terms (chronological data points 8, 14, 17) have no data as the courses were all taught by contract instructors; thus only data points 2, 5 and 11 are summer terms (as indicated). It is obvious that these two points are outside the standard deviation for all data points (the average response rates exceed 80%). Removing summer terms (bottom plot) there is a consistent average response rate of approximately 60%. The one online data point (point 18), is lower, but within the standard deviation of the data set when considered as a whole. Linear regression of these data indicates nominally zero slope, suggesting there is no systematic increase or decrease in response rate over the past 5 years. Coefficients of determination indicate that a linear fit to the data is poor, with only 6% of the variation in data explained by a linear trend (bottom plot, summer data omitted). Taken together, these facts suggest that average response rate has remained constant over the preceding 5 years.

Figure 4 shows that in our data there is no significant difference in response rate data when stratified by course type (design, laboratory and traditional lecture). Although all error bars in the

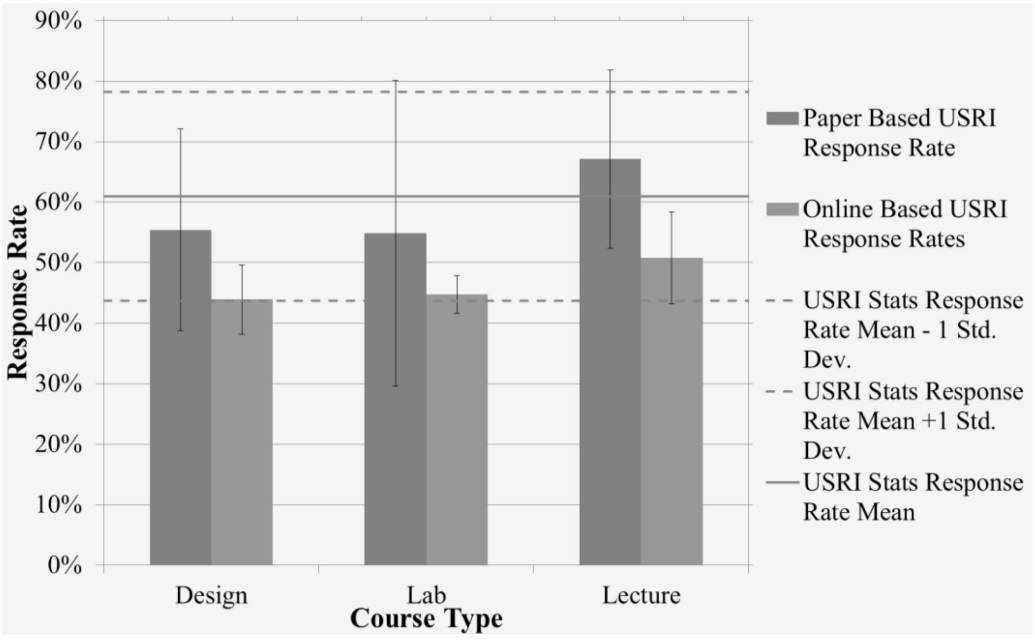


Fig. 4. Mean response height (bar chart) and standard deviation in response (error bars) stratified by course type and SET questionnaire format (paper or online). Solid horizontal line is mean response rate considering all data together, and dashed horizontal lines indicate first standard deviation in response rate (all data).

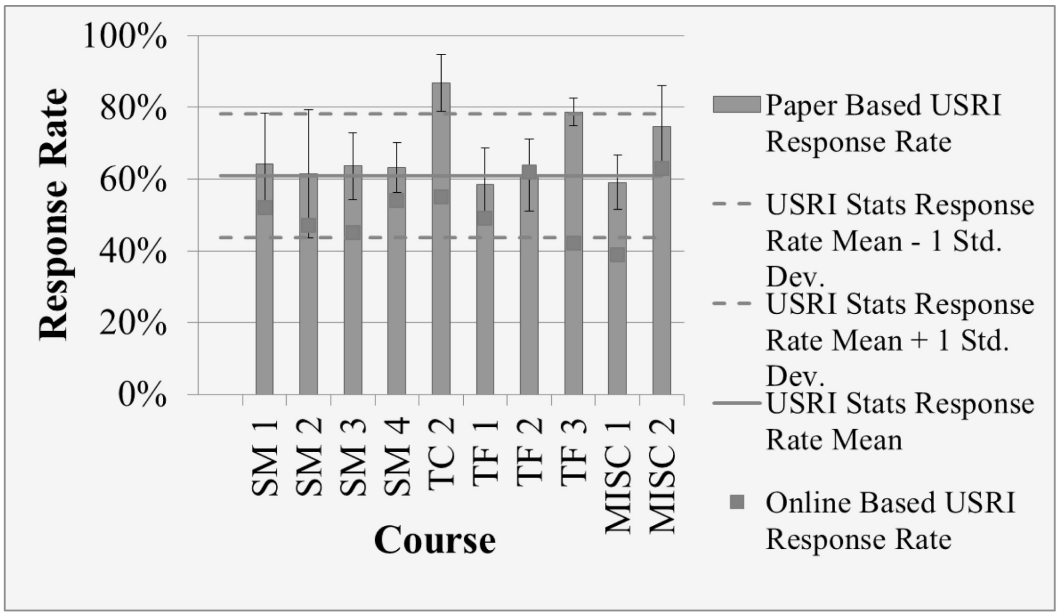


Fig. 5. Mean response height (bar chart) and standard deviation in response (error bars) by primary course focus for primarily lecture courses (SM—solid mechanics; TC—technical communication; TF—thermofluids; MISC—technical electives and other applied courses covering instrumentation and applications). Solid squares overlapping bars are response rate for the single online SET data point that corresponds to each course. Solid horizontal line is mean response rate considering all data together, and dashed horizontal lines indicate first standard deviation in response rate (all data).

Figure 4 overlap and therefore statistical significance cannot be asserted, traditional lecture based courses, on average, had the highest response rates.

Figure 5 presents response rate data (lecture courses only) stratified by the course focus (e.g. solid mechanics—SM etc.). In general, there is no systematic or statistical difference in the mean response rates. The mean paper responses for the majority of the courses, but not TC2, TF3 and MISC 2, have response rates that are comparable with the above-mentioned average of 60%. Online response rates (solid squares overlapping bars) are all lower than the average paper responses with the exception of course TF2.

3.2 Scores of universal student ratings of instructions (USRI)

Figure 6 is a histogram of median Q221 response for both paper and online protocols. As shown the mean Q221 is greater with the online protocol (4.5

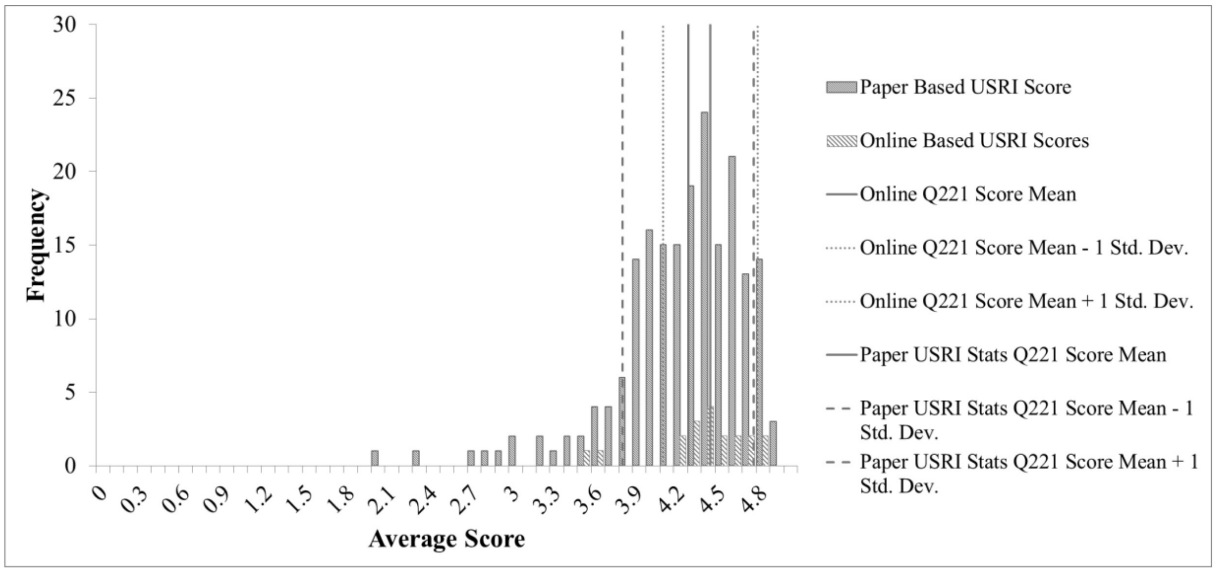


Fig. 6. Histogram for median score on Q221 (“overall instructor is excellent?”) question. Paper-based data (greyed) and online (hatched) are both shown. Solid vertical lines indicate means for paper and online data. Vertical dashed lines indicate first standard deviation.

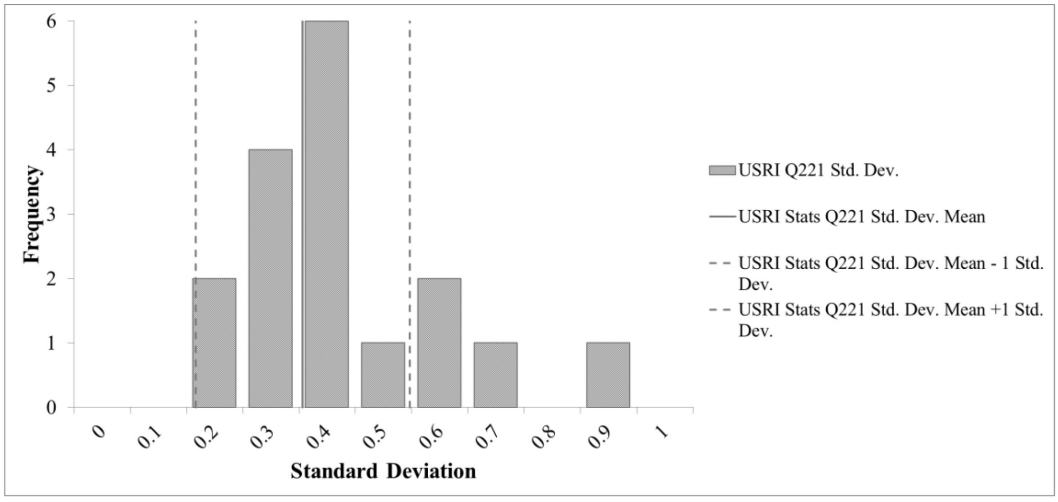


Fig. 7. Histogram for standard deviation in Q221 score for all courses considering all data. Solid vertical lines indicates mean. Vertical dashed lines indicate first standard deviation.

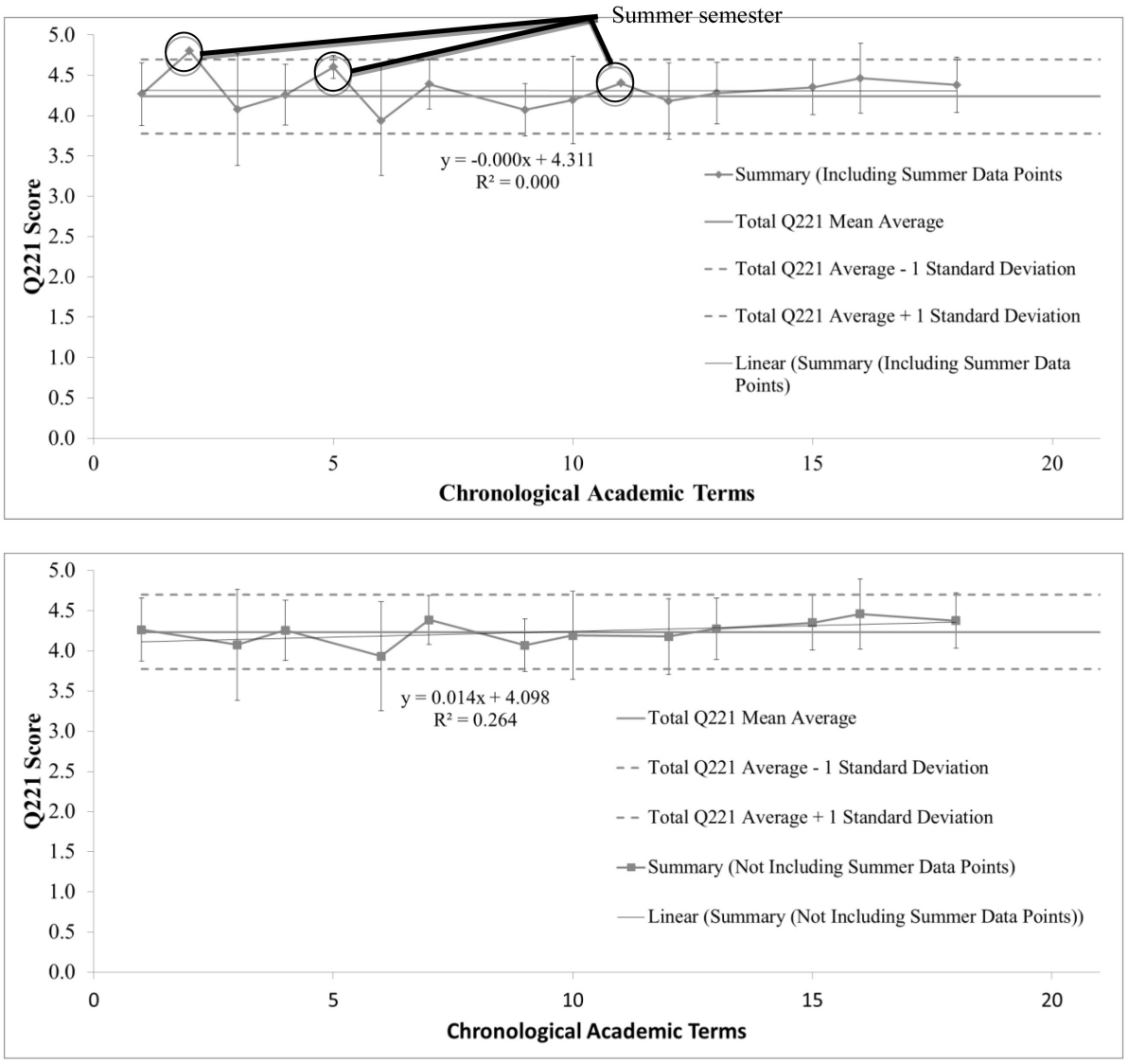


Fig. 8. (top) mean Q221 score (solid points) and standard deviation (error bar) for chronological semester; (bottom) as above, but with data for summer semesters omitted.

for online versus 4.3 for paper). The standard deviation in Q221 is 0.4 for the online protocol, while it is 0.5 for the paper protocol. Figure 7 shows a histogram of standard deviations in Q221 considering all courses and both online and paper protocols. On average, deviation in Q221 was 0.4, with a first standard deviation of 0.2.

Figure 8 shows Q221 scores in chronological order over the preceding five years. In the plots below, each data point is the mean value of the all the data points for each given term. Error bars on each data point are ± 1 standard deviation. The average mean of all data points (all 218 points without summer terms) and ± 1 standard deviation are shown as horizontal lines. These graphs do not distinguish between online and paper collection for overall mean; however, term 18, the last data point to the right, are for the semester where USRI was administered online. The top plot includes summer terms. Three summer terms (data points 8, 14, 17) have no data as the courses were all taught by contract instructors; thus only data points 2 and 5 are summer terms. These two points (4.6 and 4.8, respectively) are greater than all other points. Linear regression shows a nominally zero slope for the chronological Q221 data, and coefficient of determination that is nominally zero, further indicating no systematic linear increase or decrease in Q221.

Removing summer terms (bottom plot, Fig. 8) results in an increasing regression line for chronological Q221 scores. Taken together, the modest slope (0.02) and the coefficient of determination (0.26), which suggests only 26% of the increase is systematic, make it difficult to assert that instructor quality or student perception of instructor effectiveness are increasing

Figure 9 shows mean Q221 stratified by course type (design, lecture and lab). The figure shows that the there is no significant Q221 score differences between course types. Online evaluations led to higher average results.

Figure 10 shows Q221 score stratified by course focus (solid mechanics—SM; thermofluids—TF etc.). Of the lecture based courses examined, only SM2 and MISC 2 had mean Q221 scores below 4.0. Only course TF2 had an online Q221 score below the average and outside the standard deviation of the paper evaluation. All other online scores were equal or above the average as well as within the standard deviations.

3.3 Q221 versus Response rate

Figure 11 shows Q221 score versus response rate, for both paper and online protocol data. Regression of paper and online data indicates that a linear fit to the data is poor. Less than 1% of variation of Q221 is explained by increase in response rate for paper data, while for online data 27% is explained.

3.4 Participant variation

Figure 12 and Table 1 show how instructor data can change over time. We limited our analysis to participants who have more than 9 data points over the five year time frame. Within this subset of partici-

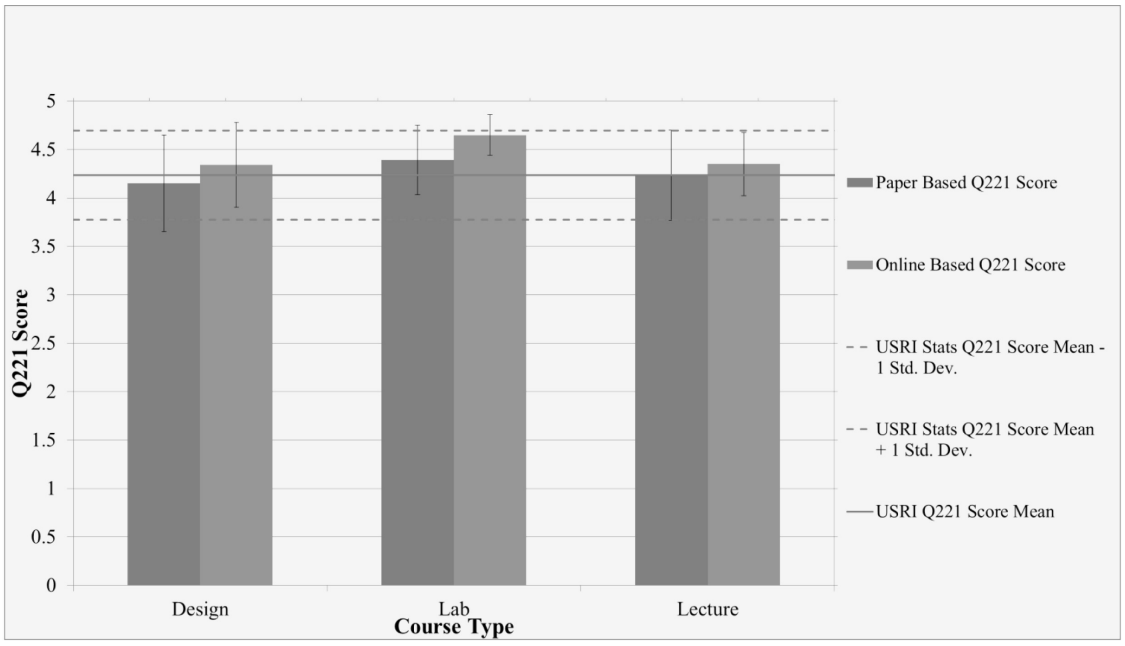


Fig. 9. Mean Q221 score stratified based on course type: design, lab and lecture. Error bars are first standard deviation. Solid horizontal line indicates overall mean for all courses, horizontal dashed lines indicate standard deviation based on all courses.

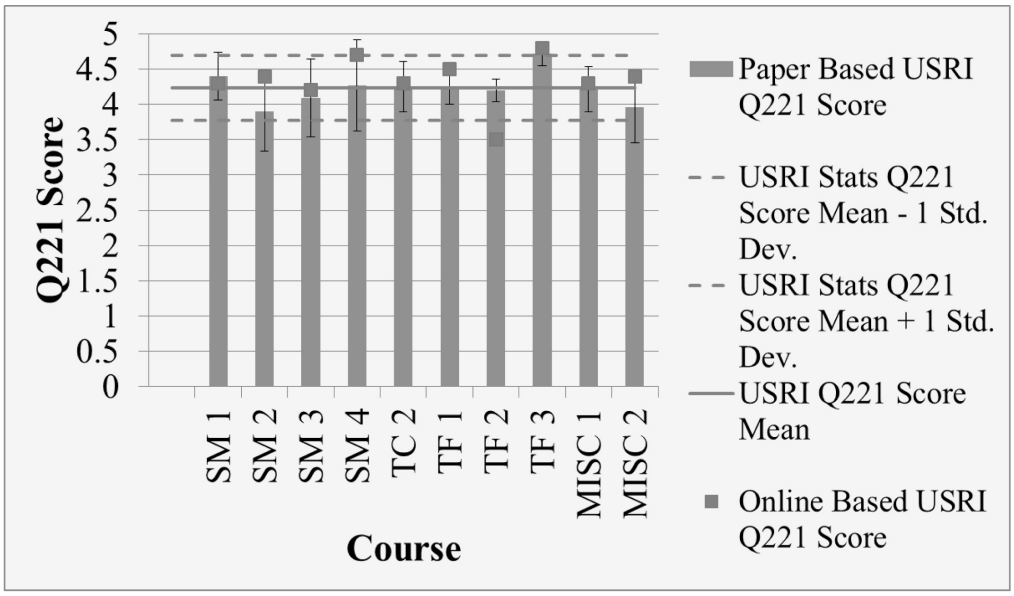


Fig. 10. Mean Q221 score (bar height) and standard deviation in response (error bars) by primary course focus for primarily lecture courses (SM—solid mechanics; TC—technical communication; TF—thermofluids; MISC—technical electives and other applied courses covering instrumentation and applications). Solid squares overlapping bars are response rate for the single online SET data point that corresponds to each course. Solid horizontal line is mean response rate considering all data together, and dashed horizontal lines indicate first standard deviation in response rate (all data).

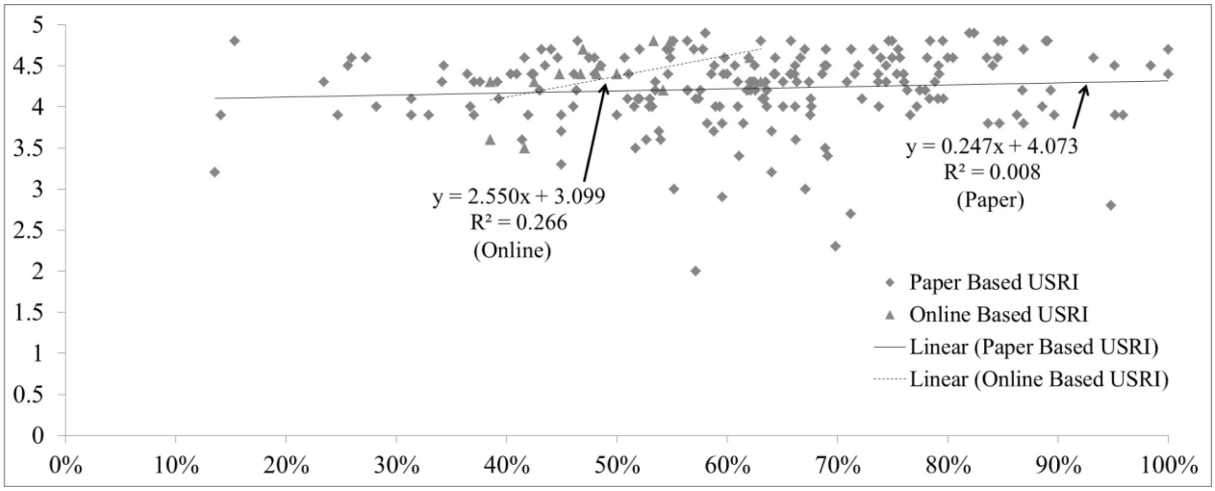


Fig. 11. Q221 score versus response rate for both paper protocol data and online protocol. Solid line is best fit regression line for paper protocol data, while dashed is for online data.

pants, we selected the participants with the lowest and highest standard deviations in Q221 scores, and two participants selected at random. The error bars shown in this graph are for a term where the participant had more than one course in a selected term. The data line itself is the averaged value of all data points for the participant in each term. If the participant only has one data point in the term, there is no error bar. The entire data set (including online data points) Q221 score mean is shown as a solid grey line, with dashed grey lines above and below to show average $\pm$ 1 standard deviation. This graph does not distinguish between online and paper responses.

The individual participant mean and standard deviation results are shown in Table 1.

Table 1. Q221 results for the highest, lowest and two random particiants

Participant	Mean Q221 Score	STDEV Q221 Score
4	4.49	0.18
9	4.43	0.29
17	3.74	0.79
35	4.22	0.26

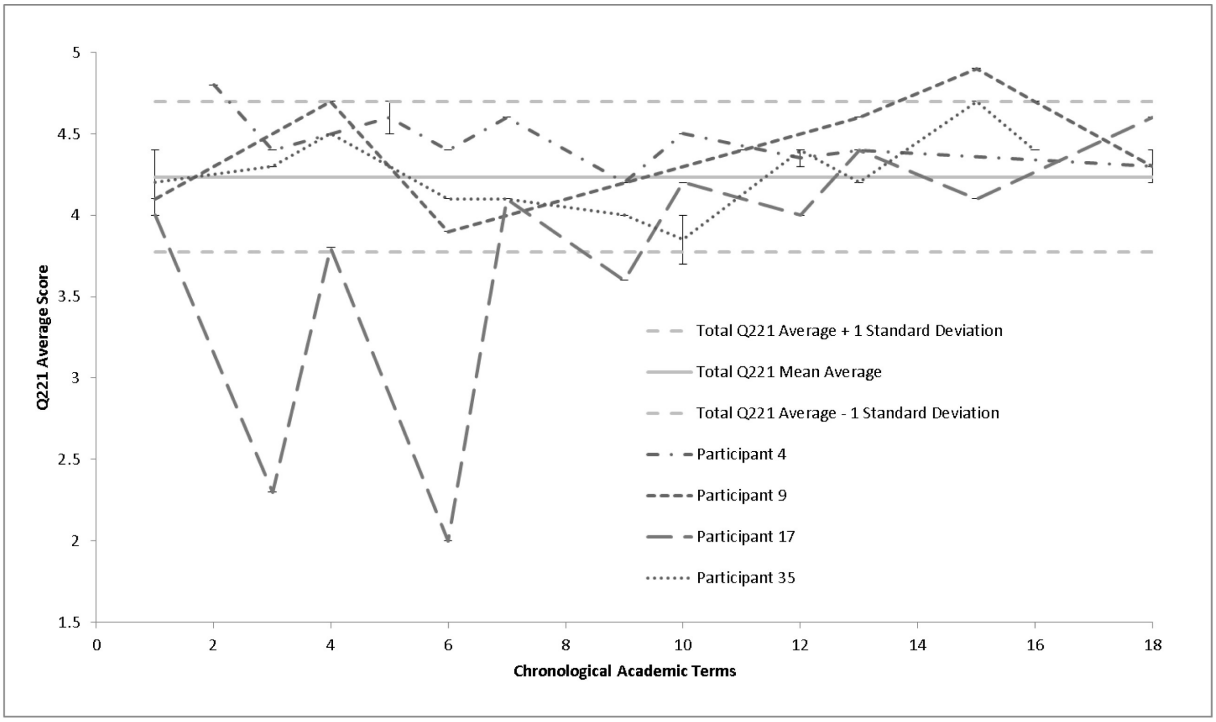


Fig. 12. Q221 score for four participants in chronological semesters. Average and standard deviations lines represent all participants. Participant lines with error bars represent the span of Q221 results as instructors were teaching more than one course in that semester.

Table 2. statistical comparisons using rank sum tests' significance set a $p < 0.05$

Basis of comparison (1) and (2) designate sample for sample size	Number of samples (n_1 vs n_2)	Significantly different
Response rate online (2) versus response rate paper (1)	17 vs 17	Yes
Q221 all online (2) versus Q221 all paper (1)	17 vs 17	No
Paper based response rate design courses (1) versus paper based response rate solid mechanics courses (2)	4 vs 4	No
Paper based Q221 design courses (1) versus paper based Q221 solid mechanics courses (2)	4 vs 4	No
Paper based response rate design (1) courses versus paper based response rate technical communications courses (2)	4 vs 3	No
Paper based Q221 design courses (1) versus paper based Q221 technical communications courses (2)	4 vs 3	No
Paper based response rate design courses (1) versus Paper based response rate thermo-fluid courses (2)	4 vs 4	No
Paper based Q221 design courses (1) versus paper based Q221 thermo-fluid courses (2)	4 vs 4	No
Response rate paper design courses (1) versus response rate paper all other lecture courses (2)	4 vs 13	No
Paper based Q221 design courses (1) versus Q221 paper all other lecture courses (2)	4 vs 13	No

3.5 Statistics

A number of comparisons of our findings were tested for statistical significant difference using a non-parametric rank sum test. Rank sum tests were used since histograms suggest data is not exactly normally distributed, we have relatively small sample sizes, and we sometimes have sample size mismatch. Significance was set at a $p < 0.05$. Table 2 lists the ten basis of comparison, the samples of each tests, and if a significant difference was found.

4. Discussion

Most universities ask students to complete course evaluations (SET) as part of university wide faculty and teaching review processes [2]; the University of Alberta is no different. Despite numerous critics and criticisms, SET is the dominant mechanism to evaluate teaching in North America. SET evaluations are contentious issues among faculty members, students and institutions. Their purpose is often ill defined. Are they to improve teaching? Are they a formal means by which students can provide constructive but anonymous feedback? Are they to provide data for faculty teaching evaluation? There are many ongoing debates on each of these questions, and we do not attempt to settle these debates using our preliminary data-set. Instead, we identify preliminary observations in response rate and Q221 score and structure future directions for further research at the University of Alberta.

Current research on SET evaluations mostly lead to one unified conclusion, if well designed, administered and interpreted, SET evaluations can be indicative of teaching quality [4]. This has been borne out by decades of research and is a common theme highlighted in SET reviews [2, 7]. Centered on concerns related to response rate and bias in results, significant research efforts have been directed into studies looking at SET in transitions from paper-based to online protocols [2, 7–12].

4.1 What the results indicate

In this paper we posed a number of questions of importance to those teaching Mechanical Engineering classes, and to a broader community. The data we gathered are in many cases unique and therefore potentially only immediately applicable to our department. The preliminary findings open the door for further long term examination of variables that could bias teaching evaluation and delivery, applicable to a broader university teaching community.

4.1.1 Is there a difference in USRI scores and response rates between paper- and web-based assessments? Is there a correlation between response rate and USRI score?

A critical question for all going through a transition from paper to web based assessments is *if there exists a difference in USRI scores and response rates between assessments?* Our findings show that there is a statistically significant difference between response rates between paper— and web—based assessments (Fig. 1). Confirming our findings, a recent review of paper and online SET, across many disciplines, indicates an overall trend for lower response rates for online protocols relative to paper [2]. Those authors conducted several studies in the context of response rate [8, 9] and their findings indicate up to 50% lower response rates for online assessments; our findings show the average response rate for all courses dropped approximately 25%. They also stress that incentivizing students and using multiple reminders can bring online rates up to be comparable with paper-based rates [10]. Thorpe [12] found that female students and students with relatively high grade point averages (GPA) were more likely to fill out the online form; it can be argued that this is not representative of a Mechanical Engineering Department student population. Universities strive to increase the success of students through instruction quality and course design, the non-response bias of low-GPA students is of concern. It is recommended that we continue to monitor online SET results to determine longer term trends and if there are non-response biases that could make some students feedback go un-reported and thus impact success of students. Select research suggests that motivating students to help improve course design and instructor effectiveness is key [13].

Considering evaluation of instructor effectiveness (Q221), the literature indicates that transitions from paper to web-based assessments do not necessarily lead to significant changes of USRI scores [2, 14]. Layne et al [14] found that there was no significant change in rating distribution between assessment methods. Kasiar et al [11] reported that in a single course (169 enrollment) the overall rating of the course based on Likert scales was largely the same between paper and online surveys. Dommeyer [10] in his study of business students of 16 volunteer instructors found online or paper assessments did not affect effectiveness scores, even when students where incentivized.

In our previous work we could not determine conclusively if there were any difference in Q221 scores between paper and web based assessments; our dataset was limited to design courses [15]. We were left with an important question: "*On a statistically significant data set considering several years of instruction and several instructors, are there significant differences in Q221 score on bases of protocol regardless of instructor and protocol where instructor is a study control?*" With our now larger dataset, we found that there was still no difference between Q221 results between paper and web based assessments (Table 2). It should still be noted that we only have one web-based term to compare to 5 years (17 terms) of paper data; further work is required to obtain a clearer answer. With more web-based data differences could emerge.

We also posed *if there is a correlation between response rate and USRI score?* It is important to note, for faculty member and evaluation committees, that from our department specific data there is no evidence to show that USRI scores are related to response rate (Fig. 11) for either paper or web based assessments. The correlations are insignificant, and thus any concern of bias could be dispelled if such trends can be confirmed with larger datasets.

4.1.2 Are there differences in USRI scores and response rates between different course types (labs, design, lecture based, miscellaneous)?

There is in our department, and likely in others, ongoing debate on the difference between course types. We found no literature in the area to verify our findings. Our results show no statistically difference between course types (Table 2 and Fig. 4). Response rates were on average greater for lecture courses than design and lab courses for both paper and web based assessments.

Figure 9 shows that lab courses have on average

the highest Q221 scores, followed by lecture then design courses, for both paper and web based courses. It should be noted that for Q221 scores, web results are all greater than paper based. Students appreciate hands-on work, and therefore good laboratory courses could have an inherent bias to be scored highly if the instructor is proficient, even if lab report writing is often disliked. Design courses are work intensive, which is a point of debate in the literature as to how it affects SET scores [16]. Design courses also involve open ended problems which students are not yet comfortable with. Lecture courses, which typically account for the majority of the curriculum, could be the baseline by which students compare all other courses.

4.1.3 Does the subject matter (solid mechanics, Thermofluids, etc) influence USRI scores and response rates?

Our current data set does not provide sufficient information to discriminate between subject matter. From Figure 5, we can clearly see that TC2, TF3 and MISC2 have much greater average response rates (~80%) than other courses, which all hover around an average of 60%. Only in the case of TF 3 (Fig. 10) does this translate to consistently greater USRI scores (both paper and web based). Again, the comparison group is small and a larger dataset may provide greater insight. However, assuming the instructor is selected to teach a course based on background and experience, the findings could indicate that instructors knowledgeable in the area, or that has taught the course a large number of times, is all that is needed fundamentally to deliver a course, and instructor does not influence response rate. Johnson [17] found a small negative correlation between experience and SET score for senior courses but a strong positive correlation between experience and USRI score in freshman engineering courses. Johnson also looked at course type, but only focused on core versus technical (program) electives. They found that technical electives had higher USRI scores than core courses; however, our sample does not include technical electives only core courses that all students must take. This question is very interesting and should be further investigated.

4.1.4 Are there differences in USRI scores and response rates between summer and traditional fall/ winter terms?

We cannot conclusively state if there are differences in USRI scores and response rates between summer and traditional fall/winter terms. Of our entire data set we only have three data points from summer terms (points 2, 5 and 11). This is a result of contract instructors during the summer terms; these individuals did not fall within our inclusion criteria. Both average response rate (Fig. 3) and USRI scores (Fig. 8) are higher than the average, and even outside the $\pm$ 1 standard deviation for points 2 and 5. Mechanical Engineering only teaches two courses, to smaller co-op stream classes, in the summer, thus we must be careful interpreting this data. However, these students typically have greater GPA than our traditional students which make up the bulk of fall and winter term attendees, which could support that better students typically have greater response rates and provide higher Q221 scores [17].

4.1.5 What is the variability in course USRI scores and response rates? Does this vary yearly?

When looking at a department, either from a 5,000 foot level or as an individual, it is important to assess the variability of SET assessments in the pool. We questioned *what was the overall and yearly variability in course USRI scores and response rates?* The standard deviation in response rates (Fig. 2) for all courses span 8 to 20% over the examined data. The standard deviation of this data spans 8-16%. The Q221 standard deviations for all courses (Fig. 7) range from 0.2 to 0.9 over the examined data. The standard deviation of the Q221 standard deviation spans 0.2–0.5. These findings are significant because they indicate, assuming a consistent material content per course as per accreditation requirements, that response rate and USRI score vary largely on either or both a yearly (students, term)—and instructor-basis. This would indicate that if an individual's variations varied by up to 0.5 from their average score it could not be considered a significant change as it may not depend on the instructor but on other factors. This can provide a means by which to judge what the University of Alberta Faculty member collective agreement [18] means by small changes in Q221 scores should not be consequential. This further makes us question if an instructor can be evaluated objectively solely by USRI score when so many other variables are at play. Such considerations are important at other institutions with faculty collective agreements which may identify such loose metrics.

This should be considered true on a yearly basis, especially when considering a one year drop in score; a one year drop can be a result of a number of variables, many outside of the instructor's control. Conversely, if an instructor has acknowledged working to improve their course and delivery, such changes should not be ignored but contextualized. Including new teaching techniques often take years to deliver correctly; these are risky endeavors that should be supported. If successful and lead to improved Q221 scores, instructors should be recognized and rewarded for their risk and self-improve-

ment. Conversely, taken in a context of no attempts to change course delivery, consistently low Q221 scores could be interpreted as ineffective teaching and a point of focus for formative attention and in time evaluation committees.

What we find in Figure 8 is that over the last five years the average Q221 score has showed no significant trends. The department has seen a significant change in demographics with a number of retirements and hires; the lack of change would be interesting to speculate on. On a positive note, it does show that our current instructors are meeting the teaching challenge of their generation. We should note that there is little room to significantly improve on scores that average 4.3 and thus drastic improvements cannot be expected. We now have a younger professoriate teaching to a new generation of students using methods they are used to seeing in their primary and secondary schools.

4.1.6 Is there a large difference in single participants and between participants in terms of USRI scores?

Figure 6, Table 1 and Figure 12 provide us some insight into these questions. The vast majority of responses of Q221 scores are 3.9 and above for all participants. From this, it is difficult to state there are large differences between participants in terms of excellence. Clearly some results, which are much lower, should be identified by instructors and supervisor to address possible concerns. Again this must be put in the context of a number of years and not simply from one point of data. It is of concern to note that Johnson [17] and Feldman [19] found that female instructors were significantly disadvantaged in terms of USRI score in lower level courses. Since our sample has only two females, this does not allow us to examine such data without breaking anonymity.

Single individual variability seen in Table 1 and Figure 12, show that the participant that has the greatest variability drops below the one standard deviation from the mean line in early years, but has shown dramatic improvement in Q221 scores in following years. We cannot speculate to the reason for this, be it formative or course substitution. Most of the data shown here falls within the $\pm$ 1 standard deviation of the department Q221 mean but do fluctuate; we cannot however ascertain what the root cause for this fluctuation is or what impact it could have on formative and evaluative processes. The average Q221 score for all instructors ranges from 3.12 to 4.80; the average is 4.23 ± 0.4, which is a common inflation from the expected 3.0 (i.e. signifying that the students on average agree that the instructor is excellent) found in SET evaluations [20]. Johnson [17] found in their study that Q221 average for mechanical engineering was 4.12 ± 0.48, showing our department fairs well. For each individual, the standard deviation of their Q221 score ranges from 0.09 to 0.79; these extremes cases indicate someone that is very consistently scored and one that has seen regular large jumps in scores or possibly significant progress/egress.

Particular concerns of using SET for evaluation or for constructive feedback is response rate. How high must it be to be representative of the class? Are low response rates indicative of apathy, satisfaction or even more simply the class starts at 8am? During the span of a term, student absenteeism increases, this is most obvious in 8 am classes.

There are no quantitative indications in the University of Alberta's Faculty collective agreement to what constitutes a good or poor Q221 score. Appendix A of the Faculty of Engineering Faculty Evaluation Standards and Process document states:

> "Promotion to the rank of Professor cannot be granted to individuals whose overall USRI Instructor Excellence median rating (USRI question 221) is less than 3.5 out of 5 in three or more of the preceding five years."

We could therefore infer that less that 3.5 is a poor score for Full Professors, when formative years are supposedly behind the instructor. However, with changing student demographics (Gen Y, Millennials) and new teaching approaches and technologies, the professorial teaching formative period should be continuous. Johnson found that assistant professors scores are greater than tenured colleagues and that older faculty receive lower USRI scores [17]; while others found just the opposite [21, 22]. Figure 6 shows the means $\pm$ 1 standard deviation range for paper and web based Q221 scores to be approximately 3.9–4.8 and 4.1–4.8, respectively. In this perspective, it would inappropriate to state that excellence is only if Q221 scores fall outside the upper standard deviation. In such cases excellent teaching would very seldom be rewarded. At what score does the question "was the instructor excellent?" truly indicate an instructor is excellent?

Again, Figure 12 shows the scores of a subset of four of our participants with all data averaged. Looking at these results, it is difficult to imagine how Q221 scores can be used other than punitively, or in rare occasions beneficially, since the average score is above the "agreed—the instructor was excellent", and the one standard deviation mark is only slightly below that threshold. In reality, this is not how the evaluation process works. What occurs is that all scores within the Faculty of Engineering are divided in quartiles per year of instruction. This separation between instruction year is important since there was found to be a positive correlation between year of instruction and set score (i.e. more

senior courses higher Q221 score). These quartiles are used to "bin" instructors, where those in the upper quartile are likely to receive recognition for their teaching, while those in the lowest quartile scrutinized for their teaching. Furthermore, at time of evaluation, a student weighted average Q221 score is considered and presented over a five year period; allowing for those who teach larger classes to be less affected by the negative correlation between class size and SET score [17]. Furthermore, a recently added measure in the faculty of engineering is the number of formal contact hour to account for those courses that include formal team meetings or other laboratory or seminar contact hours. If used during evaluation, these metrics do account for some measure of additional commitments required in these instances and level the playing field for all instructors.

5. Conclusions

In this paper we paved the way for a number of questions to be further explored. In our study, we found based on limited data that there were differences in the response rate but not for the primary evaluation question of SET teaching evaluations, when comparing 5 years of paper-based versus 1 web-based testing. Of the six questions we originally posed, only response rate between paper- and web-based was found to be significantly different. Further, we encourage faculty (especially junior) to elicit from students regular feedback in an effort to combat non-response biases while also addressing the need for multi-faceted evaluation techniques that we have outlined.

References

1. H. Kanuka, P. Marentette, J. Braga, K. Campbell, S. Harvey, R. Holte, J. Nychka, D. Precht, D. Read, C. Skappak and C. Varnhagen, *Evaluation of Teaching at the University of Alberta. Report of the Sub-committee of the Committee on the Learning Environment (CLE)*, University of Alberta, Jan. 2009.
2. H. M. Anderson, J. Cain and E. Bird, Online Student Course Evaluations: Review of Literature and a Pilot Study, *Am. J. Pharm. Educ.*, **69**(1), Feb. 2005, pp. 34–43.
3. H. K. Wachtel, Student evaluation of college teaching effectiveness: A brief review, *Assess. Eval. High. Educ.*, **23**(2), 1998, pp. 191–212.
4. T. Beran, T. Collin, E. Silva and R. Haukenfrers, *The Universal Student Ratings of Instruction Instrument at the University of Calgary: A review of a three year pilot project. A report by the USRI Review Committee*, University of Calgary, 2003.
5. R. A. Arreola, *Developing a comprehensive faculty evaluation system. A guide to designing, building, and operating a large-scale faculty evaluation*. San Francisco, California: Anker Publishing, 2007.
6. B. L. Yoder, *Engineering by the numbers: 2013 ASEE Survey of Engineering Colleges*, American Society of Engineering Education.
7. C. Ballantyne, Online Evaluations of Teaching: An Examination of Current Practice and Considerations for the Future, *New Dir. Teach. Learn.*, **2003**(96), 2003, pp. 103–112.
8. C. J. Dommeyer, P. Baum, K. S. Chapman and R. W. Hanna, Attitudes of Business Faculty Towards Two Methods of Collecting Teaching Evaluations: Paper vs. Online, *Assess. Eval. High. Educ.*, **27**(5), 2002, pp. 455–462.
9. C. J. Dommeyer, P. Baum and R. W. Hanna, College Students' Attitudes Toward Methods of Collecting Teaching Evaluations: In-Class Versus On-Line, *J. Educ. Bus.*, **78**(1), 2002, pp. 11–15.
10. C. J. Dommeyer, P. Baum, R. W. Hanna and K. S. Chapman, Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations, *Assess. Eval. High. Educ.*, **29**(5), 2004, pp. 611–623.
11. J. B. Kasiar, S. L. Schroeder, and S. G. Holstad, Comparison of Traditional and Web-based Course Evaluation Processes in a Required, Team-Taught Pharmacotherapy Course, *Am. J. Pharm. Educ.*, **66**, 2002, pp. 268–270.
12. S. W. Thorpe, Online Student Evaluation of Instruction: An Investigation of Non-Response Bias. AIR 2002 Forum Paper., Jun. 2002.
13. J. J. Giesey, Y. Chen and L. B. Hoshower, Motivation of Engineering Students to Participate in Teaching Evaluations, *J. Eng. Educ.*, **93**(4), Oct. 2004, pp. 303–312.
14. B. H. Layne, J. R. Decristoforo and D. Mcginty, Electronic versus traditional student ratings of instruction, *Res. High. Educ.*, **40**(2), Apr. 1999, pp. 221–232.
15. C. R. Dennison, C. Ayranci, P. Mertiny and J. P. Carey, Transition from paper to online course evaluation: preliminary trends in student response rate and overall professor evaluation, in *Proceedings of the 2014 Canadian Engineering Education Association (CEEA14) Conference*, Canmore AB, p. Paper 122 (pp. 1–6).
16. A. Greenwald and G. Gillmore, No pain, no gain? The importance of measuring course workload in student ratings of instruction, *J. Educ. Psychol.*, **89**(4), pp. 743–751.
17. M. D. Johnson, A. Narayanan and W. J. Sawaya, Effects of Course and Instructor Characteristics on Student Evaluation of Teaching across a College of Engineering, *J. Eng. Educ.*, **102**(2), Apr. 2013, pp. 289–318.
18. *University of Alberta Faculty Agreement: July 2006*, University of Alberta, July 2006.
19. K. A. Feldman, College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers, *Res. High. Educ.*, **34**(2), Apr. 1993, pp. 151–211.
20. W. E. Cashin, Student ratings of teaching: Recommendations for Use. Idea Paper No. 22, in *Center for Faculty Evaluation and Development, Kansas State University*, 1990.
21. M. A. McPherson and R. T. Jewell, Leveling the Playing Field: Should Student Evaluation Scores be Adjusted?, *Soc. Sci. Q.*, **88**(3), 2007, pp. 868–881.
22. M. A. McPherson, R. T. Jewell and M. Kim, *What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes*, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1649946, Jul. 2010.

Robert C. Butz, B.Sc., E.I.T., is an MSc student in Mechanical Engineering at the University of Alberta. He earned his BSc in Civil Engineering from the University of Alberta. He is a researcher in the biomedical instrumentation lab.

Jason P Carey, PhD, PEng, is a professor in Mechanical Engineering at the University of Alberta. He is the Associate Chair (undergraduate program) and Director of the Biomedical Engineering Undergraduate Option. He is the lead researcher in the Biomedical and Composite research laboratory.

Chris R. Dennison is an Assistant Professor of Mechanical Engineering, University of Alberta. He is principal investigator of the biomedical instrumentation lab that focuses on developing instrumentation for injury and basic biomechanics research.

R. Shawn Fuhrer, B.Sc., E.I.T., is an MSc student in Mechanical Engineering at the University of Alberta. He is a researcher in the Biomedical and Composite research laboratory. He earned his BSc in Mechanical Engineering from the University of Alberta.