Transition from Paper to Online Course Evaluation: Preliminary Trends in Student Response Rate and Overall Professor Evaluation

Christopher R. Dennison, Cagri Ayranci, Pierre Mertiny and Jason P. Carey
Department of Mechanical Engineering, University of Alberta, Edmonton AB, Canada
Corresponding Authors: cdenniso@ualberta.ca and jpcarey@ualberta.ca

Abstract – Student evaluation of teaching (SET) has been used as a metric to arguably evaluate instructor effectiveness and quality of instruction since the 1920s. SET is used in decisions regarding annual evaluation (of faculty) and is one of the most researched topics in evaluation of instructor effectiveness. Central research questions associated with SET include whether SET is an appropriate measure of effectiveness and whether it leads to improved teaching and quality of graduates.

In the fall 2013, the Faculty of Engineering at the University of Alberta for the first time administered SET online. The transition from paper-based and in class SET to online and out of class SET 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 would suggest protocol-related bias. Our preliminary results show lower response rates for online SET and effectiveness scores that were outside one standard deviation of the previous 5-year mean. These findings show the importance of continuing to monitor web-based SET results and point to directions of further research. At this time, this will not be possible, as the Faculty of Engineering has discontinued online SET testing after a single term.

Keywords: USRI, universal student rating of instruction, SET, student evaluation of teaching, engineering, education, measurement, effectiveness

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.

Our overall objective was to identify changes in response rate and instructor effectiveness from preliminary data and structure research questions for future work. Focusing on two Department of Mechanical Engineering core design courses, typically deemed academically demanding with high survey response rates, involving the same three instructors, we compare response rate and 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). Concerns about response bias are discussed in terms of continued USRI testing value and potential effect on junior faculty.

2. MATERIALS AND METHODS

2.1. Study Design

This preliminary study focused on two core design courses in the Mechanical Engineering Department identified here as DC1 and DC2 in the results. These courses were selected because they have stable curricula and have been taught by the same core of professors for the past five years, are academically challenging, interactive, and involve large term projects. Historically, the response rate to the SET questionnaire for these courses has been high.

SET Data collected was the response rate (% of total class that responded to questionnaire) and question 221 of the questionnaire: “overall I find the instructor excellent”. Q221 can be ranked:

- 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, involving the same three instructors, was collected and compared with results of the Fall 2013 term web-based SET results.

Ethics approval was received from the University of Alberta Ethics board (Study ID Pro00045934) for this study; participant anonymity was preserved through coding the data a priori.

2.2. Participants and Data

Three instructors participated in this study. Each participant is at a different stage of his or her academic career; the three represent all professorial ranks. Each participant has taught at least one of the courses no less than 3 times. Because of the three program streams in our department, (traditional, coop I and coop II) class sizes range from 70 to 130.

2.3. Data Analysis

All data presented is randomly assigned a unique identifier to prevent participant recognition. The

exception is the data for the single web-based results. Although, participant information is not directly provided, this allows for greater possibility of identification through university of Alberta website searches; participants were aware of this limitation in agreeing to participate.

Term- and aggregate-based comparisons of data is presented. Descriptive statistics (standard deviations and averages) are used to discuss the findings.

3. RESULTS

Table 1 provides raw scores for response rates and Median Q221 responses for DC1 (top) and DC2 (bottom) for paper and web-based results.

Table 1: Unique identifier for course offering, response rate, and median score on question 221 for Design Course 1 (DC1) and Design Course 2 (DC2). Grayed background denotes data for online protocol.

DC1			
Delivery Method	Unique identifier	Response rate (%)	Median Q221
Paper based	1	58.9	4.5
	2	-	-
	3	-	-
	4	58.6	4.4
	5	-	-
	6	25.7	4.5
	7	100.0	4.4
	8	48.0	4.4
	9	26.0	4.6
	10	41.0	4.4
Web-based	11	44.0	4.7

DC2			
Delivery Method	Unique identifier	Response rate (%)	Median Q221
Paper based	1	62.0	4.6
	2	-	-
	3	64.4	4.4
	4	80.0	4.6
	5	62.6	4.2
	6	69.0	4.7
	7	62.5	4.5
	8	-	-
	9	-	-
	10	83.6	4.6
Web-based	11	40.2	4.3

3.1 Response rate

Figure 1 provides the response rate for DC1 and DC2 as well as the average (solid lines) and banded by one-standard deviation lines (dotted lines, e.g. DC1+S.D.) for the 5 years of paper based results (1 to 10) versus the web-based results (11). Note that the result for the DC2 web-based test falls well outside one standard deviation of the mean of the paper-based tests. More specifically, it falls outside two standard deviations of the mean. Conversely, for DC1, the web-based result is near the mean result.

Figure 1: Response rate (%) for DC1 and DC2 over the preceding five years. Vertical gray bar denotes data for online protocol.

3.2 USRI Scores

Figure 2 provides the Q221 scores for DC1 and DC2 as well as the average (solid lines) and banded by one-standard deviation lines (dotted lines, e.g. DC1+S.D.) for the 5 years of paper-based results (1 to 10) versus the web-based results (11). For both DC1 and DC2, web-based results fall outside one standard deviation from the mean for the respective courses. DC1 result appears more of an abnormality, and is outside two standard deviations of the mean of paper-based results. DC2 paper-based results have a much larger standard deviation, largely due to a consistent gap between two instructors, where one averages 4.60 ± 0.07 and the other 4.30 ± 0.14 .

4. DISCUSSION

Most universities ask students to complete course evaluations (SET) as part of university wide faculty and teaching review processes [2]. 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 teaching evaluation? There are many on-going 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.

Figure 2: Median Q221 score for DC1 and DC2 over the preceding five years. Vertical gray bar denotes data for online protocol.

4.1 Response rate and Q221 score

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][6]. 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], [6]–[11].

Considering response rate, 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]. Dommeyer and colleagues (College of Business) have conducted several studies in the context of response rate [7], [8] and their findings indicate lower response rates for online (in some cases 50% lower) but also highlight that incentivizing students and using multiple reminders can bring online rates up to be comparable with paper-based rates [9]. Thorpe [11] investigated response rates and non-response bias from a large dataset spanning three courses and found that overall, online response rates were lower than those

associated with the paper protocol. Thorpe's results also suggest that female students and students with relatively high grade point averages (GPA) were more likely to fill out the online form. This latter finding could be viewed as problematic because feedback from students with poorer GPAs, it could be argued, may be different than that from students with high GPAs if one presumes that the structure of the course has some bearing on student success. Given that Universities strive to increase the success of students through instruction quality and course design, this non-response bias of low-GPA students could be viewed as worrisome.

While our dataset is limited, our observations are consistent with that of the above research: we noted that the first online protocol response rates were lower than the mean rate for the previous five years of paper based SET. While we cannot assert statistical significance of this observation, we can note that our result is in agreement with previous findings and recommend that if online evaluation persists we should monitor SET results to answer the question: *Is response rate consistently lower with online SET and are there non-response biases that could make some students feedback go un-reported and as a consequence impact success of students?* To answer these questions, a sound understanding of the common motivators of students to complete SET in the engineering context is needed. Select research suggests that motivating students to help improve course design and instructor effectiveness is key [12].

Considering evaluation of instructor effectiveness, an overall theme from the literature is that transitions from online to paper based protocols do not necessarily manifest in significant changes to USRI scores [2], [13]. Layne and colleagues [13] found that there was no significant change in rating distribution across protocol when considering a student sample of approximately 2,500 students. Kasiar and colleagues [10] 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 (roughly 1/3 survey questions indicated a significant difference in Likert score). A recent study by Dommeyer [9] which considered business students of 16 volunteer instructors suggests that the overall rating of instructor effectiveness is not significantly influenced by online or paper protocol and further that even when incentivized, there is no significant difference. While these previous investigations seem to convey a unified theme, the authors typically warn against extrapolating their results to other schools and instead recommend testing their findings at schools contemplating transition to online evaluation.

Due to limitations in our data, it is difficult to compare our results to that of previous work. We observe that for both design courses, the Q221 score for the online protocol falls outside one standard deviation calculated on

the previous five years of data but we also note there is significant scatter in our data. Considering DC2, the Q221 score for the online protocol falls within the range on Q221 score for the past five years (Figure 2). In contrast, the DC1 Q221 score (online) falls outside the range shown for the past five years (Figure 2). This difference may be due, in part, to the fact that DC2 is offered by two different instructors and as a result with continued surveillance we propose to study the following 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?*

The value of answering these questions is that any changes in Q221 score for a given instructor could be considered in light of the change in protocol and against typical variability in Q221 score and these considerations could give both instructors and administrators the information required to determine if changes in Q221 should be considered significant or, in other words, cause for concern. Other factors that could be included in research questions similar to the above would include the influence on Q221 of course difficulty and workload which is a common concern for instructors [14][15]. Select research suggests that exam score correlates with instructor ratings [16] which further suggests that a considered approach when judging Q221 scores should be implemented, which would necessitate surveillance of factors not currently included on judgments of Q221 score.

4.2 Considerations for Faculty and Administration

This paper focuses on two facets, administration and possible interpretation of results of a small subset of SET testing. Our focus on two design courses, as well as three instructors, between 5 years of paper-based data and one term of web-based data cannot provide a thorough interpretation of department, faculty or university wide results; nor do the authors attempt to make such grand commentary.

Design courses are very time consuming for students, but also very much so for instructors since one of the core beliefs of our design program is that *design cannot be taught, it must be mentored*. Design courses therefore require regular one-on-small group activities, meetings, discussions and mentorship. These cannot represent, currently, other fundamental courses in our program. We further focused on these two courses because, contrary to capstone design courses, students are also provided new analytical approaches and tools during the term. Students and instructors must thus balance issues of design with applied science concepts. This was important since that double content is often more challenging and can lead to

a different (lower) SET scoring. As part of our pilot study, we focused on two outcome measures of SET evaluations, response rate and Q221.

Particular concerns of using SET for evaluation or for constructive feedback is response rate. DC1 varied from 25 to 100% paper-based response rate, with a web-based response of 44%. DC2 varied from 62 to 84% response rate for paper-based and 40% for web-based. 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. DC1 is an 8 am starting class, while DC2 is a 9:30 or 10am starting class depending on the term. Some of the discrepancy between response rates during paper-based evaluations could result from class schedule. However, one advantage of paper-based responses was that students that were in class filled out the SET evaluation. The intent of web-based SET was to provide flexibility to students; however, there was no incentive. The latter could explain why DC2 response fell for the web-based response. It begs the questions: *what incentive is there for a student to fill in the evaluation? What subset of the class would respond?* Possibilities range from absolutely loving the instructor and the course content to reviling the same set.

The Q221 results discussed in this work provide a limited comparison set; interpretation of what they can truly provide is limited, but do offer an opportunity to start being vigilant of SET scores. We should first note that among instructors, the average results for the two courses are 4.46 and 4.51 for DC1 and DC2, respectively, for paper-based evaluations. The three instructors are known as engaging and approachable and have won a number of teaching awards. This could explain the high Q221 scores. The two web-based results fall outside of the one standard deviation of the means for their respective course. DC1 web-based result falls well outside of two standard deviations. This could indicate that the 44% that responded were biased positively (which would agree with previous research as described above). The data spread for DC2 is much larger and the below-average web-based response is within the spread. This spread is largely due to the typical score of one instructor versus the other. It is noteworthy that University Policies state that that small variations in Q221 should not be considered significant; however, that allowable variance is not defined.

Do our findings demonstrate a true measure of student assessment of instructor effectiveness or not? That question cannot be answered and thus it is imperative to be careful in solely using such information for evaluation purposes. University of Alberta General Faculty Council policies state that “Evaluation of teaching shall be multifaceted”; further, the collective

agreement with Faculty members states that teaching evaluation should be multi-faceted:

“13.06 The standards for evaluation of teaching performance shall be broadly based, including course content, course design and performance in the classroom. Such evaluation may take into account information such as statistical summaries of responses to student questionnaires, comprehensive reviews of student commentary; reviews by peers, reviews by administrative officials and reviews of teaching dossiers and other materials provided by the staff member.” (UA, Faculty Agreement, 2006, p17.)

SET evaluations are only one tool at the disposal of instructors to present a case to faculty evaluation committees. UofA Centre for Teaching and Learning, offers peer teaching evaluations for the entire university. However, few of the other tools are at the disposal of instructors at department levels. The Department of Mechanical Engineering started in 2013 a teaching peer mentorship program to provide willing instructors with feedback on their teaching performance and course content and encourage peers to interact with the students to get early feedback that instructors can apply to the current class and not solely to the next. Such approaches were shown to be highly successful in the literature [17,18]. This peer-based assessment is meant to provide formative feedback as well as a second facet for evaluation purposes if the mentored instructor wishes to do so. The guidelines of the program indicate that this should be a two-year evaluation process allowing instructors to improve their skills, this does not happen overnight.

Teaching evaluation is a key consideration for junior faculty members vying for tenure or promotion. With due respect to these very important processes, it is of critical importance that teaching evaluation be assessed fairly. It is also vital that any SET evaluation have these progress-through-the-ranks reviews as a primary objective and allow for formative feedback. Our results could not conclusively determine if the web-based method introduced bias. It must be noted that a recent report by the University of Alberta’s Renaissance Committee [19] suggested that if interested parties could not put forward a SET with clear objectives and desired outcomes, that a moratorium of SET be enacted, the authors hope it does not come to that.

The authors cannot provide more in depth analysis with such a small set of data. We believe, as many others do, that all teaching evaluation must be multifaceted. We believe it vital that such research be undertaken with a larger set of data when or if it becomes available. It is noteworthy that the Faculty of Engineering, after one web-based SET evaluation (fall 2013), returned to paper based the next term (winter 2014). We believe that the

web-based approach has value, and it must be further explored.

5. CONCLUSIONS

In our pilot study, we found based on limited data that there were some possible differences in the response rate and primary evaluation question of SET teaching evaluations for two core design courses, with three different instructors, when comparing 5 years of paper-based versus 1 web-based testing. These results must be viewed as preliminary and not statistically significant, but we believe that they should further encourage the University of Alberta to pursue further the questions raised (above) regarding SET and Q221 score and, further to enforce multifaceted faculty teaching evaluation and explore new ways of engaging students in the evaluation process. 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 a multi-faceted evaluation techniques that we have outlined.

Acknowledgements

The authors acknowledge the help of Mr. Fuhrer for data collection as well as all participants. This work was unfunded.

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.*, vol. 69, no. 1, pp. 34–43, Feb. 2005.
- [3] H. K. Wachtel, "Student evaluation of college teaching effectiveness: A brief review," *Assess. Eval. High. Educ.*, vol. 23, no. 2, pp. 191–212, 1998.
- [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] C. Ballantyne, "Online Evaluations of Teaching: An Examination of Current Practice and Considerations for the Future," *New Dir. Teach. Learn.*, vol. 2003, no. 96, pp. 103–112, 2003.
- [7] 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.*, vol. 27, no. 5, pp. 455–462, 2002.
- [8] 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.*, vol. 78, no. 1, pp. 11–15, 2002.
- [9] 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.*, vol. 29, no. 5, pp. 611–623, 2004.
- [10] 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.*, vol. 66, pp. 268–270, 2002.
- [11] S. W. Thorpe, "Online Student Evaluation of Instruction: An Investigation of Non-Response Bias. AIR 2002 Forum Paper," Jun. 2002.
- [12] J. J. Giesey, Y. Chen, and L. B. Hoshower, "Motivation of Engineering Students to Participate in Teaching Evaluations," *J. Eng. Educ.*, vol. 93, no. 4, pp. 303–312, Oct. 2004.
- [13] B. H. Layne, J. R. Decristoforo, and D. McGinty, "Electronic versus traditional student rating of instruction," *Res. High. Educ.*, vol. 40, no. 2, pp. 221–232, Apr. 1999.
- [14] K. C. Dee, "Student Perceptions of High Course Workloads are Not Associated with Poor Student Evaluations of Instructor Performance," *J. Eng. Educ.*, vol. 96, no. 1, pp. 69–78, Jan. 2007.
- [15] 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.*, vol. 102, no. 2, pp. 289–318, Apr. 2013.
- [16] J. A. Centra, "Student Ratings of Instruction and Their Relationship to Student Learning," *Am. Educ. Res. J.*, vol. 14, no. 1, pp. 17–24, 1977.
- [17] S. Dennison, "Peer mentoring: Untapped potential," *Journal of Nursing Education*, 49 (6), pp. 340–342, 2010.
- [18] P. Boyle, B. Boice, "Systematic mentoring for new faculty teachers and graduate teaching assistants," *Innovative Higher Education*, 22 (3), pp. 157–179, 1998.
- [19] University of ALberta Renaissance Committee report, hwww.renaissance.ualberta.ca/~media/renaissance/RenComReportComplete20131209Footnote.pdf, 2013