Student Evaluation Instruments: Online Versus Paper

Brian Balyeat, Xavier University
Julie Cagle, Xavier University

ABSTRACT

Results from student evaluations of teaching are regularly used in the merit review process that determines faculty raises, as well as for the rank and tenure process. This paper examines the implications of the switch from paper based student evaluations of teaching to electronic delivery based student evaluations of teaching. Data is analyzed based upon gender of the instructor, rank of the instructor, type of course, and the past performance of the instructor. Results show that the student response rate decreased significantly for the overall dataset and each subset. Additionally, the teaching performance scores dropped significantly for the overall dataset. Performance under electronic evaluations dropped more significantly for full professors and instructors than assistant professors and associates. Differences are also noted based on type of course and relative teaching performance. These results suggest caution for drawing inferences about teaching quality over time periods in which there is a switch in the media (paper vs. electronic) used for student evaluations of teaching.

INTRODUCTION

The purpose of this study is to examine the difference in response rates, instructor ratings, and course ratings that resulted from a switch from paper course evaluations to online course evaluations in the William's College of Business (WCB) at Xavier University. Course evaluations are understandably a focal point for faculty. These evaluations are used in merit processes and rank and tenure processes as one measure to evaluate the quality of teaching. Clayson (2009) reports that almost all business schools (99.3%) use student evaluations of teaching (SET) and deans generally place higher importance on them than peer evaluations of teaching. Therefore, it is no surprise that switching from paper evaluations to online evaluations would be a concern to faculty. In the case of Xavier University, the switch was also made with a short timetable and minimal faculty input, adding the concern of lack of shared governance to the process. Switching to online evaluations is typically done as a cost savings measure and to free up instructional time, but there may also be disadvantages to online evaluations.

LITERATURE REVIEW

Response rates are a particular concern noted in prior research. Guder and Malliaris (2013) indicate that most research on student response rates collected online are far lower than those collected on paper and cite Nulty's (2008) finding of a drop on average of about 23% compared to in-class response rates. Guder and Malliaris' (2010) review of prior literature on response rates indicates a range from a low of 31% to a high of 89%. Their review of literature also indicates mixed results regarding feedback, with more favorable ratings, less favorable ratings, and no change in ratings being reported in studies comparing online to paper evaluations. Carini et al. (2003) find college students respond more favorably on all eight scales contained in

a survey on student engagement using the web-based survey versus the paper surveys. This was true for both male and female students and younger and older students.

Guder and Malliaris (2010) examine the roles of faculty status (full-time vs. part-time), course type (core vs. advanced classes), and class size (large vs. small classes). While they note changes in instructor evaluations with the switch from paper to online evaluations, the changes were similar to the changes they noted between semesters with paper evaluations. The notable difference in the semester when online evaluations were used was between core and advanced classes. Advanced classes have a higher response rate and rank courses and instructors slightly higher, while core classes have a lower response rate and rank courses and instructors slightly lower. Similar results were reported for full-time versus part-time instructor with higher response rates and ratings for the full-time. Larger classes had lower response rates, higher instructor ratings, and lower course ratings versus smaller classes.

Gender remains an issue in reviews of performance. Basow (1995) in examining student evaluations at a private liberal arts college finds male professor ratings were unaffected by student gender, while female professor ratings were highest among female students and lowest among male students. Female faculty ratings also varied by divisional affiliation. In a subsequent student, Basow (1998) concludes the gender effect is small at about 3% of variance, but that significant interaction effects between gender and other context variables may cumulatively disadvantage female faculty. Basow and Silberg (1987) report male students gave female professors poorer ratings than male professors on six teaching evaluation measures, while female students rated female faculty lower on three measures.

In a more recent student by Benjamin Schmidt, a Northeastern University history professor, using 14 million reviews on Rate My Professor website, the results indicated people think more highly of men than women in professional settings (Miller (2015)). For example, they focus on a woman's personality or appearance, while they focus on a man's intelligence. Mr. Schmidt offers a caution that evaluations must be viewed keeping in mind cultural conditioning. Perhaps the most startling recent evidence on gender and student evaluations is out of North Carolina State University (MacNell, Driscoll, and Hunt (2015)). Two professors taught online classes, removing the possibility of gendered behavior, and switched their identities. While neither the male nor female professor received significantly higher ratings, the male instructor had lower ratings when then students thought he was a female instructor and the female instructor had higher ratings when the students thought she was a male instructor. This is convincing evidence of gender bias in student evaluations of teaching.

Driscoll and Cadden (2010) offer caution about using student evaluations to make comparative evaluations of teaching quality. Their results indicate statistical differences between departments that suggest a global standard (school or college) would be unfair. They also find differences in course type (required by core, required by major, or elective) and students' anticipated grades, which would further argue against a global standard.

DATA AND METHODOLOGY

Given that most prior research suggests a decline in response rates when switching from paper to online evaluations, we hypothesize a decline in the response rate with the switch to online evaluations in the WCB. Prior research indicates that core courses have lower instructor

and course evaluations than those required for the major or electives. Student evaluations of teaching may differ between full and part-time faculty or by gender of the instructor. Cross-sectional analysis will be used to examine whether differences in instructor and course ratings with the switch to online evaluations differed by gender of the instructor, rank of the instructor (instructor, assistant professor, associate professor, or full professor), or type of course (business core courses, required courses for specific majors, or elective courses). Analysis is also conducted to determine if the drop in scores is related to the magnitude of the score (i.e. did poorer performing professors with paper evaluations drop more under the electronic system or did better performing professors exhibit a greater drop).

Spring 2014 is used as a baseline as that is the last semester paper evaluations were used in the WCB. Evaluations for fall 2014 are included as that is the first semester of online evaluations for Xavier University, including the WCB. Descriptive statistics for the number of sections in the data set are provided below in Table 1.

Table 1 – Descriptive Statistics

	Number of Sections (Spring 2014)	Number of Sections (Fall 2014)	Total
Number of Sections	165	149	314
Sections Taught by			
Instructors	32	30	62
Assistant Profs.	31	29	60
Associate Profs.	61	52	113
Full Professors	41	38	79
Females	73	66	139
Males	92	83	175
UG Core classes	68	75	143
UG Major-Required	38	31	69
UG Major-Elective	17	9	26
UG Business Elective[#]	2	1	3
MBA Core	29	21	50
MBA Elective	10	9	19
Executive MBA[#]	2	2	4

[#]Groups do not contain enough observation for any further analysis.

The data consists of student evaluations for 314 sections of classes taught that are divided between 165 Spring sections and 149 Fall sections. Any sections taught by faculty who did not teach in both semesters (e.g. new hires, retiring faculty, sabbatical, etc.) are excluded from the above counts and any subsequent analysis. The data also excludes sections taught by adjunct faculty. Current WCB policy excludes sections taught as overloads from the merit review process; therefore, those sections are also excluded from the analysis. The data is also broken down by professor rank (instructor, assistant, associate, and full professor), gender, undergraduate class type (business core, required for the major, and elective in the major), and MBA class type (core, elective, or executive).

RESULTS

At the university level, the response rate declined from 83% with paper evaluations to 70% with online evaluations. Though there is no standard question regarding instruction quality, some version of overall instructor rating existed for all departments. The overall rating for instructor for the university went from 4.46 (on a five point scale) with a standard deviation of 0.240 to 4.30 (on a five point scale) with a standard deviation of 0.267. The overall university results are suggestive of a drop in response rate and instructor evaluation with the switch from paper to online evaluations (Herbert, 2015).

We delve into these issues with more robust analysis specifically for the WCB. In the WCB, the evaluation instrument consists of 35 questions grouped into 8 categories (presentation ability, organization/clarity, grading/assignments, intellectual/scholarly, student interaction, student motivation, instructor rating, and course rating). Unconditional results are provided in Table 2 for the instructor rating, course rating, and for an overall rating (the average score across the eight categories) for both paper evaluation (Spring 2014) and online evaluations (Fall 2014). Additionally, the response rate for both semesters is provided.

Going from paper evaluations to online evaluations resulted in a response rate decline from 84.2% in the spring to 71.5% in the fall. This 12.8% drop in the response rate is significant at the 1% level. The result is not surprising given that paper evaluations were usually done towards the end of the semester during class time. This usually resulted in a high response rate. However, the online evaluations were generally done by the students outside of class. Filling out the online evaluation was not part of the student's grade nor was filling out the evaluation tied to students receiving their grades. Thus, it is not surprising that the response rate dropped significantly with the introduction of an online evaluation system.

Table 2 – Unconditional Results

	Paper Evaluations (Spring 2014)	Number of Sections (Spring 2014)	Online Evaluations (Fall 2014)	Number of Sections (Fall 2014)	Difference in Mean
Unconditional Results		165		149	
Instructor Rating	4.32		4.18		0.14***
Course Rating	4.22		4.14		0.08*
Overall Rating	4.42		4.29		0.12***
Response Rate	84.2%		71.5%		12.8%***

*Significant at the 10% level
**Significant at the 5% level
***Significant at the 1% level

Within the WCB, looking at professors who taught in both semesters, the average score on the evaluations dropped from a 4.42 in the spring (paper evaluations) to a 4.29 in the fall (first semester for the electronic evaluations). A difference in means test showed that the 0.12 difference was significant at the 1% level. Results for the instructor rating and course rating questions showed similar decreases. These results are very consistent with the results for the university as a whole.

Additional tests were performed to determine if the rank of the professor correlated to the general decrease in performance observed from the unconditional results in Table 2. Professors were broken down into four groups: instructors (non-tenure track), assistant professors, associate professors, and full professors. At Xavier, instructors are hired primarily to teach and there is a very high teaching hurdle required to become a full professor.

Table 3 – Professor Rank

	Paper Evaluations (Spring 2014)	Number of Sections (Spring 2014)	Online Evaluations (Fall 2014)	Number of Sections (Fall 2014)	Difference in Mean
Panel A: Instructors		32		30	
Instructor Rating	4.47		4.28		0.18**
Course Rating	4.33		4.17		0.16*
Overall Rating	4.55		4.37		0.18***
Response Rate	86.2%		78.9%		7.3%**
Panel B: Assistant Prof.		31		29	
Instructor Rating	4.20		4.16		0.04
Course Rating	4.03		3.93		0.10
Overall Rating	4.36		4.28		0.08
Response Rate	84.0%		74.2%		9.8%***
Panel C: Associate Prof.		61		52	
Instructor Rating	4.26		4.21		0.05
Course Rating	4.18		4.23		-0.05
Overall Rating	4.35		4.32		0.03
Response Rate	82.7%		70.7%		12.0%***
Panel D: Full Prof.		41		38	
Instructor Rating	4.38		4.08		0.30***
Course Rating	4.34		4.16		0.18**
Overall Rating	4.46		4.21		0.25***
Response Rate	85.1%		64.4%		20.7%***

*Significant at the 10% level
**Significant at the 5% level
***Significant at the 1% level

For all four professor ranks, the response rate decreased significantly with the move to electronic evaluations. The decrease ranged from 7.3% for instructors to an alarming 20.7% decrease for full professors. This compares to the overall unconditional decrease of 12.8% denoted in the last row of Table 2.

While the evaluations decreased for instructor rating, course rating, and overall rating for each category (except for the course rating for associate professors), these decreases were only significant for instructors and full professors. It is interesting to note that instructors and full

professors outperformed assistant and associate professors with the paper evaluations given in the spring 2014 semester. However, the results are more mixed in the fall 2014 semester with electronic evaluations. Moving to electronic evaluation had a more dramatic effect on the better performing instructors and full professors than on assistant or associate professors.

The effect of professor gender is analyzed in Table 4. The purpose of this analysis was not to see if female professors out or underperformed their male counter parts. Rather this analysis shows the effect of switching to electronic evaluations for both genders separately.

Table 4 – Gender

	Paper Evaluations (Spring 2014)	Number of Sections (Spring 2014)	Online Evaluations (Fall 2014)	Number of Sections (Fall 2014)	Difference in Mean
Panel A: Female		73		66	
Instructor Rating	4.35		4.20		0.15**
Course Rating	4.27		4.15		0.12*
Overall Rating	4.47		4.32		0.15*
Response Rate	86.2%		77.0%		9.3%***
Panel B: Male		92		83	
Instructor Rating	4.29		4.16		0.13**
Course Rating	4.18		4.13		0.04
Overall Rating	4.38		4.27		0.10**
Response Rate	82.7%		67.1%		15.6%***

*Significant at the 10% level
**Significant at the 5% level
***Significant at the 1% level

As with professor rank, the response rate for both female and male professors decreased significantly with the move to electronic evaluations. Results for female professors decreased 9.3% and decreased 15.6% for male professors. Both of these results were significant at the 1% level.

The results also dropped for all three rating categories (instructor, course, and overall) for both females and males. These decreases were statistically significant in all cases except for male course ratings. The drops in the evaluations were larger for female professors than the corresponding drop for their male counterparts. Additionally, although the results are not statistically significant, female professors outperformed their male counterparts in all three performance measures in the spring 2014 semester. Like in the professor rank analysis in Table 3, the drops in the evaluations are the largest for the group with the highest spring 2014 paper evaluations.

It is possible that the results are related to the type of class that is being taught. Table 5 looks only at the undergraduate courses in the sample. Undergraduate classes are divided into three categories: business core, courses required for a major, and electives. In the case that a course is required for one major and can also be counted as an elective in another major, the

course was coded as a required course for a major. As can be seem from Table 5, the majority of the undergraduate classes fall into the business core classification.

For all three undergraduate class types the decrease in the response rate is significant at the 1% level. The response rate for undergraduate core classes only decreased by 7.8% while the response rate for undergraduate electives decreased by 22.0%. This result is surprising in that one might expect the response rate to be related to student interest and one would expect students to be more interested in electives in their major than business core courses.

Table 5 – Undergraduate Class Type

	Paper Evaluations (Spring 2014)	Number of Sections (Spring 2014)	Online Evaluations (Fall 2014)	Number of Sections (Fall 2014)	Difference in Mean
Panel A: UG Business Core		68		75	
Instructor Rating	4.26		4.18		0.08
Course Rating	4.11		4.10		0.01
Overall Rating	4.34		4.27		0.07
Response Rate	80.6%		72.8%		7.8%***
Panel B: UG Major - Required		38		31	
Instructor Rating	4.47		4.27		0.20**
Course Rating	4.46		4.33		0.14
Overall Rating	4.56		4.40		0.16**
Response Rate	88.5%		76.5%		12.0%***
Panel C: UG Major - Elective		17		9	
Instructor Rating	4.38		3.90		0.48**
Course Rating	4.26		3.94		0.32*
Overall Rating	4.47		4.06		0.40**
Response Rate	83.0%		61.0%		22.0%***

*Significant at the 10% level
**Significant at the 5% level
***Significant at the 1% level

While the performance decreased in all three categories for business core classes, none of these results were statistically significant. However, the drop in the performance is statistically significant at the 5% level for the instructor rating and overall rating for courses that are required for the major. Undergraduate elective course show the greatest decrease across all three categories and these results are all statistically significant. The results conditional on undergraduate class type are different than the previous results conditional on instructor rank and gender in that the highest rated group (courses required for the major) did not show the largest decrease in performance.

Table 6 shows the results for the MBA classes. The course ratings decrease for MBA core classes and actually increase for MBA elective classes. However, none of these results are

statistically significant. For both groups of MBA classes, the participation rate decreases by a little over 20% and this decrease is statistically significant.

For many of the previous tables, the largest decrease in the evaluations occurred for the group with the highest spring evaluations. Professor ratings for the three evaluation criteria are averaged over the Spring 2014 semester. The results are weighted by the number of responses in each class. The Spring 2014 results are then sorted into terciles (top third, middle third, and bottom third). The analysis in Table 7 shows the performance of each professors in each tercile (as measured by their Spring 2014 performance) against the courses those professors taught in the Fall 2014 semester. Counts in each tercile are not identical because the methodology sorted professors into terciles and each professor did not necessarily teach the same number of courses. Additionally, due to changing service reductions to teaching loads, many professors taught a different number of classes in the spring 2014 semester than they did in the Fall 2014 semester.

Table 6 – MBA Class Type

	Paper Evaluations (Spring 2014)	Number of Sections (Spring 2014)	Online Evaluations (Fall 2014)	Number of Sections (Fall 2014)	Difference in Mean
Panel A: MBA Core		29		21	
Instructor Rating	4.25		4.11		0.14
Course Rating	4.10		4.03		0.07
Overall Rating	4.39		4.26		0.13
Response Rate	87.1%		67.0%		20.1%***
Panel B: MBA Elective		10		9	
Instructor Rating	4.11		4.31		-0.20
Course Rating	4.12		4.33		-0.21
Overall Rating	4.32		4.42		-0.10
Response Rate	83.6%		61.8%		21.8%***

*Significant at the 10% level
**Significant at the 5% level
***Significant at the 1% level

Like the other conditional analysis, the response rate decreased significantly in all categories. However, the effect on student course evaluations differs significantly based upon Spring 2014 performance. Panel A in the table shows the results for the top performers in 2014. The top performers showed a significant decrease across all three performance metrics. This decrease is significant at the 1% level. With the exception of the course rating questions, professors in the middle third show a significant decrease in performance with online evaluations. This decrease in performance was significant at the 1% level and the magnitude of the decrease is similar to the results for the top third.

However, the results for the bottom third of performers in Panel C, is very different from the other two terciles. For the bottom performers in the Spring 2014 semester going to electronic evaluations actually increased their performance in all three categories. However, this increase was not statistically significant.

The results in Table 7 are consistent with the patterns in the data for professor rank and gender. The greatest decrease in performance occurred in the group with the highest evaluations. Thus, the best performers in the college were hurt the most by the switch to electronic evaluations.

Table 7 – Conditional Performance

	Paper Evaluations (Spring 2014)	Number of Sections (Spring 2014)	Online Evaluations (Fall 2014)	Number of Sections (Fall 2014)	Difference in Mean
Panel A: Top Third					
Instructor Rating	4.68	55	4.42	51	0.25***
Course Rating	4.62	55	4.38	50	0.24***
Overall Rating	4.71	56	4.50	52	0.21***
Response Rate	91.7%	53	74.6%	50	17.1%***
Panel B: Middle Third					
Instructor Rating	4.38	60	4.12	54	0.26***
Course Rating	4.29	59	4.25	54	0.04
Overall Rating	4.42	59	4.24	53	0.18***
Response Rate	83.9%	65	72.7%	55	11.2%***
Panel C: Bottom Third					
Instructor Rating	3.84	50	3.97	44	-0.13
Course Rating	3.71	51	3.75	45	-0.05
Overall Rating	4.09	50	4.11	44	-0.03
Response Rate	76.4%	47	66.3%	44	10.1%***

*Significant at the 10% level
**Significant at the 5% level
***Significant at the 1% level

CONCLUSION

In the fall of 2014, the WCB at Xavier University switched from paper course evaluations to an electronic system. This switch provided a unique opportunity to examine the effect of moving to electronic evaluations on the participation rate and student ratings. Results are presented unconditionally and based upon professor type (instructor, assistant, associate, and full), gender, class type (undergraduate core, undergraduate required for the major, undergraduate elective, MBA core, and MBA elective), and past performance.

The results show that the response rate significantly decreases both unconditionally and for every subset of the data. Additionally, instructors and full professors show a significant and much greater decrease in evaluation performance with electronic evaluations than do assistant or associate professors. Both female and male professors are negatively impacted by electronic evaluations as are professors who teach courses in undergraduate majors, especially electives in those programs. Lastly, both conditional on prior performance and for many subsets of the data the decrease in performance is the greatest for higher performing professors. These results offer

a caution for using student evaluations to evaluate faculty teaching performance in time periods in which the media of the student evaluations (paper vs. electronic) changes.

REFERENCES

Basow, S.A. Student evaluations of college professors: When gender matters. Journal of Educational Psychology, 1995, 87(4), 656-665.

Basow, S.A. Student evaluations: Gender bias and teaching styles. In L.H. Collins, J.C. Chrisler, and K. Quina (Eds.), Career Strategies for Women in Academe: Arming Athena, 1998, Sage, Thousand Oaks, CA.

Basow, S.A. and N.T. Silberg. Student Evaluations of College Professors: Are Female and Male Professors Rated Differently? Journal of Educational Psychology, 1987, 79(3), 308-314.

Carini, R.M., J.C. Hayek, G.D. Kuh, and J.A. Ouimet. College Student Responses to Web and Paper Surveys: Does mode matter? Research in Higher Education, 2003, 44(1), 1-19.

Clayson, Dennis. Student Evaluations of Teaching: Are They Related to What Students Learn? Journal of Marketing Education. April 2009. 31(1), pp. 16-30.

Driscoll, Jennifer and David Cadden. Student Evaluation Instruments: The Interactive Impact of Course Requirement, Student Level, Department and Anticipated Grade. *American Journal of Business Education*. May 2010, 3(5), pp. 21-29.

Guder, Faruk and Mary Malliaris. Online and Paper Course Evaluations. *American Journal of Business Education*. February 2010, 3(2), pp. 131-137.

Guder, Faruk and Mary Malliaris. Online Course Evaluations Response Rates. American Journal of Business Education. May/June 2013, 6(3), 333-337.

Herbert, Steve. Course Evaluation Results. Email to faculty. April 10, 2015.

MacNell, Lillian, Adam Driscoll, and Andrea Hunt. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. Innovation in Higher Education. 2015, 40, pp. 291-303.

Miller, Carol Cain. Is the Professor Bossy or Brilliant? Much Depends on Gender. February 6, 2015. http://www.nytimes.com/2015/02/07/upshot/is-the-professor-bossy-or-brilliant-much-depends-on-gender.html?_r=0&abt=0002&abg=0

Nulty, D. The Adequacy of Response Rates to Online and Paper Surveys: What can be Done? Assessment and Evaluation in Higher Education, June 2008, 33(3), 301-314.