2012

Comparing Ratings: In-class (paper) vs. Out of Class (online) Student Evaluations

Ronald R. Mau
Embry-Riddle Aeronautical University, maur1@erau.edu

Rose Opengart
Embry-Riddle Aeronautical University - Worldwide, opengarr@erau.edu

Higher Education Studies; Vol. 2, No. 3; 2012
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

Comparing Ratings: In-class (paper) vs. Out of Class (online) Student Evaluations

Ronald R. Mau[1] & Rose A. Opengart[1]

[1] Department of Business Administration, Embry Riddle Aeronautical University—Worldwide, Daytona Beach, Florida, USA

Correspondence: Ronald R. Mau, Department of Business Administration, Embry Riddle Aeronautical University, Florida, USA. E-mail: Ronald.Mau@erau.edu

Received: July 3, 2012 Accepted: July 17, 2012 Online Published: August 22, 2012

doi:10.5539/hes.v2n3p55 URL: http://dx.doi.org/10.5539/hes.v2n3p55

Abstract

Student evaluations of teaching (SET) are used by institutions of higher learning in the tenure and promotion process and in awarding merit pay increases. The trend at some institutions has been towards using an online student assessment instrument (SAI) in lieu of the traditional paper –based, in-class assessment. This study examines the difference in student evaluations in two contexts; online and paper-based, in a finance course taught to non-finance majors. The evidence strongly indicates faculty receives higher evaluations using a paper-based instrument administered during class than with an online assessment instrument which students complete on their own time.

Keywords: student evaluations, teaching evaluations, ratings

1. Introduction

Every semester a majority of institutions conduct a student assessment of teaching, typically by employing a survey or rating form completed by students currently enrolled in a class for the purposes of evaluating a faculty member's teaching effectiveness. The use of these instruments was introduced to U.S. institutions of higher learning in the 1920s (Doyle, 1983). Not surprisingly, the 1920s also brought the first research studies on student ratings (Remmers & Brandenburg, 1927). These student ratings are often the primary measure of teaching effectiveness and are used by many institutions for promotion, tenure and merit pay increase decisions. The use of these instruments in human resources-related decisions makes the issues related to student evaluation of faculty teaching effectiveness an important concern for faculty and administrators.

Traditionally, evaluations have been conducted using pencil and paper. Students would rate professors on a scale, filling in scanned bubble sheets and answering a list of questions. These paper and pencil surveys were completed during a class period near the end of the course. Recently, institutions have migrated toward the use of online evaluations, citing lower costs in administering the evaluations and compiling results. The online evaluations are not typically completed by students during a class period; students access and complete the web-based evaluation on their own during a set time period near the end of the term.

The literature related to student evaluations has investigated a wide range of topics including purpose and goals of evaluation (Arreola, 2007; Theall, Abrami, & Mets, 2001), validity, and reliability (Abrami, d'Appolonia & Cohen, 1990; Centra, 1993; Marsh, 2007). Some authors (Feldman, 1998; Theall & Franklin, 1990) have argued validity is more related to process and day-to-day practice than to quality and soundness of instruments and analysis. Still others suggest that evaluation issues result from a lack of agreement on a definition of teaching excellence and that "homemade evaluations" are developed by people lacking knowledge of psychometrics (Theall et. al., 2005).

Given the increase in online learning and use of online instructor evaluations, as well as their potential importance to a faculty member's career, this research is critical to determine whether there are statistically significant differences in evaluations completed online, on the student's own time, as compared to paper evaluations completed within the classroom setting.

The purpose of this research is to compare instructor evaluation scores from online evaluations completed at the

student's convenience and in-class paper methods of administering the evaluations. By limiting the results to the same class and professor, a number of bias issues are eliminated, allowing the primary question of online versus in-class paper evaluations to be addressed. The class used was an introductory finance class taught to non-finance majors at a regional comprehensive institution in the southeast region of the United States.

2. Literature Review

Numerous studies investigating student evaluations of teaching have found biasing factors potentially affecting an instructor's evaluation. These factors are not necessarily related to teaching effectiveness and may introduce potential biases into the evaluations (Basow & Silberg, 1987; Cohen, 1981; Franklin & Theall, 1992; Kulik. 2001). Examples include time of day the class is taught (Koushki & Kuhn, 1982), class size (Greenwald & Gilmore, 1997), level of the class (Marsh 2007), student interest, student GPA, and gender of the faculty (Sidanius & Crane, 1989).

Faculty characteristics such as charisma and enthusiasm and the use of an entertaining style have been shown to relate positively to teacher evaluations, particularly on presentation skills items (Abrami et.al., 1982; Chan, 2004; Lees & Barnard, 1999). Another example of biases found in teacher evaluations having nothing to do with teacher evaluations is the number of rows in a classroom. Safer et. al. (2005) studied college algebra classes at California State University and found that as the number of rows increased in the classroom the average student rating declined. The authors also provided further evidence of the positive link between expected grades and teacher evaluation rating.

Despite the continued uncertainty and long list of biases unrelated to teaching effectiveness, the use of student evaluations has grown. Seldin (1993) found use of evaluations increased from 29 to 86 percent of institutions. Emery, Kramer and Tian (2003) discussed how evaluations from students have increased in importance, particularly at institutions emphasizing teaching. Wines and Lau (2006) discuss the legal issues and unfairness of how the use of teacher evaluations has evolved to its current use in making personnel decisions. Use has grown and become more important at teaching institutions yet legal scholars continue to express concerns with the use of student evaluations.

Evidence indicates there are many ways to assess faculty performance, using both quantitative and qualitative methodologies (Theall et. al., 2005). With approximately 40 years of research (Theall et. al., 2005) and as much as 90 years of use (Doyle, 1983), debates continue, arguing for and against the validity and/or reliability of ratings and evaluations as accurate measures of teaching performance. A potential cause of this is lack of agreement regarding the definition of teaching excellence and uncertainty as to which measure(s) to use as the criterion of teaching effectiveness. These are critical issues in ensuring evaluations of an instructor's effectiveness are based on student learning rather than their opinions/ratings (Kulik, 2001; Theall et. al., 2005). However, when ratings instruments are properly constructed, administered, and analyzed, they are useful tools (Marsh, 1983; Theall et. al., 2005).

Research suggests that in order to improve the practice of evaluating faculty it is important to employ strategic evaluations with multiple data sources, rigor, and consideration of contextual factors (Arreola, 2007; Centra, 1993; Feldman, 1998; Scriven, 1994; Seldin; 1991; Theall et al., 2005; Theall and Centra., 2001; Theall & Franklin 1990). One important contextual factor to consider is location/modality: in-class paper vs. out of class online evaluations.

Modality appears to affect response rates, as research has indicated online evaluations may produce lower response rates than classroom evaluations. This could be attributed to the often voluntary nature of out of class, online evaluations. Layne, DeCristofor, and McGinty (1999) found a response rate of 60.6 percent for in-class paper respondents and 47.8 percent for respondents in the online group. Their results indicated seniors and students with lower GPAs are the least likely to respond to online evaluations. Other studies have also found lower response rates for online (outside of class) evaluation instruments, with response rates 20 to 40 percent higher using classroom paper methods (Ewell, 2000; Rosenberg et. al., 2001).

Avery et.al. (2006) offer a number of possible explanations for the lower response rates from online evaluations. Potential explanations include the captive audience in the classroom and social pressure to complete the evaluation in the classroom which is not present in the online version, the impact of the environment on students (distractions when out of class), multiple evaluations to fill out at the end of the semester on their own time and the perception of students that responses are not anonymous.

Johnson (2002) studied the effects of student characteristics on mean evaluation scores. The author's students

completed evaluations before and after receiving their grade. The results found students who expected higher grades were more likely to provide a favorable review. Students who expected an A minus were 20 to 30 percent more likely to provide a favorable review then those expecting a B. The expected B students were 20 to 30 percent more likely to provide a favorable rating than those expecting a C and so on. In addition, after receiving grades, students who did not receive the grade expected actually lowered their responses and those who received higher than expected grades increased their ratings.

Student evaluations are a form of survey, and as such, the survey literature could be examined for parallels. One stream of research related to the participation in surveys is the investigation of customer satisfaction/dissatisfaction. Best and Andreasen (1977) found consumer dissatisfaction varied among product categories, with more consumer discontent from appliances such as washers and dryers than from cameras and stereo equipment.

An analogy can be drawn to student evaluations, as numerous studies have found a relationship between the nature of the discipline and student ratings. For example, Feldman (1998) found instructors in highly quantitative courses received lower ratings than those in social science courses. Similar results have been found by Centra (1993), Marsh and Dunkin (1992), Nuemann (2000) and others. Anderson (1998) concluded that the category of customers most likely to complain and express negative statements are those customers that are the most dissatisfied. This finding, in addition to the relationship of dissatisfaction to category or discipline, could be linked to differences in classes and student interest in completing the out-of-class online evaluations. The content of the course may matter and the most dissatisfied students may disproportionately represent those students who respond and complete online evaluations.

Research comparing online vs. classroom evaluations is conflicting in its conclusions. Layne et al. (1999) found differences by academic areas but no effect on mean scores between paper and electronic evaluations. Ewell (2000) found higher scores for paper-and-pencil methods and Avery et al. (2006) found online evaluations did not adversely affect evaluation scores. Hardy (2003) found online evaluations were 0.25 points lower (on a six-point scale) than were paper-based evaluations. Linse's 2010 study at Penn State found that while response rates were lower using online evaluations, average online scores were similar to the average scores using paper evaluations.

The prior results predominately indicate no significant differences between online and paper evaluations. Since the results are mixed with the limited research that has been conducted, additional data is needed to address the question of whether or not there is a significant difference in evaluation results when comparing classroom-based paper evaluations with online evaluations. This further research is necessary in order to address the issues related to the chosen mode of obtaining student evaluations, and the effects of that mode on the evaluation means.

3. Study Objectives

The objective of this study is to compare the results of online out-of-class teaching evaluations to in-class paper evaluations. This study used an introductory finance class for non-finance majors which is a required course for graduation. As it is a finance class, the class does contain an emphasis on quantitative concepts. The class was taught in the same classroom each term, by the same instructor. By this construction a majority of the effects of biases found in prior research (i.e. instructor characteristics, room characteristics) are controlled and the primary questions can be evaluated.

The two primary research questions are:

1) Do online outside-of-class evaluation systems requiring students to complete evaluations on their own time result in lower response rates relative to paper evaluations completed within class?
2) Do these online evaluations lead to significantly different mean course evaluation scores relative to paper evaluations?

4. Methods

The data were collected for a total of 8 semesters, starting in the fall of 2007 when the online system was first used on campus, through the spring 2011 semester. The fall of 2007 assessments used a 5 point Likert scale while all other semesters used a 4 point Likert scale. Because of the different scales, the fall 2007 data is not used in this study or reported. However, the results were similar to those reported here.

Students were asked to complete online evaluations between weeks 12 through 14 of the 15-week semester. The online evaluation was the official evaluation used by the university. Results were made available after final

grades were submitted. Students were informed of the online rating system in class via university wide e-mails and banners and posters prominently displayed across campus. In some terms, student's names were entered into campus-wide drawings for prizes after completing an evaluation for a course.

The typical presentations of the evaluation scores in the tenure and promotion process consist of an overall mean score for responses to all of the questions. A simple table consisting of courses taught by semester and the average for each course is reported. Reports provided to faculty and department heads provide results for individual questions as well.

Paper evaluations were completed by students in class during the last week of class (week 15) and consisted of the same questions/instrument as the online evaluation. Paper evaluations were given after the online evaluations were available for students to complete and after the online evaluations had closed. Students were never informed paper evaluations would be given in addition to the online evaluations. This process was used to reduce or eliminate lower response rates during the online evaluation window by students, in the event this knowledge might reduce on-line evaluation participation.

The evaluations were given to the same introductory finance course class taught to non-finance majors by the same instructor in each of the semesters studied. By using this process, various issues related to class characteristics such as class level/difficulty, time of day of the class, male or female instructor have been controlled. These characteristics have been found in some cases to affect evaluations although some characteristics have not been shown to possess a consistent effect (Basow & Silberg, 1987; Franklin & Theall, 1992; Koushki & Kuhn, 1982; Kulik, 2001).

5. Results

Regarding response rates, prior literature has indicated response rates are lower for the web-based evaluations. A testable hypothesis is:

H (1): There is no difference in the response rates between online out of class evaluations and in class paper evaluations.

The predominate result in previous studies as discussed in the literature review is there is no difference in the results of the evaluations between the on-line out of class evaluation and the traditional paper based in class evaluation.

To test our second objective our hypothesis is:

H (2): There is no difference in the mean ratings between online out of class evaluations and in class paper evaluations.

Table 1 provides the questions asked in the student evaluation. Students responded by selecting 1: strongly disagree, 2: disagree. 3: agree or 4: strongly agree. In addition to the questions, each group of four questions is designed to provide information on five factors of teaching as identified by the bold headings in Table 1. The instrument is the standard form used by the university during the course of the research. The authors had no control over the questions, some of which are not recommended in the literature. For example, asking about instructor preparation (Item 1) is discouraged in the literature.

Table 1. Student Assessment Instrument

No.	Category and Question
	Organization and Clarity
1	My instructor is well prepared for class meetings.
2	My instructor explains the subject manner clearly.
3	My instructor clearly communicates course goals and objectives.
4	My instructor answers questions appropriately.
	Enthusiasm and Intellectual Stimulation
5	My instructor is enthusiastic about teaching the course.
6	My instructor presents the subject in an interesting manner.
7	My instructor stimulates my thinking.
8	My instructor motivates me to do my best work.
	Rapport and Respect
9	My instructor helps students sufficiently with course-related materials.
10	My instructor is regularly available for consultation.
11	My instructor is impartial in dealing with students.
12	My instructor respects opinions different from his or her own.
	Feedback and Accessibility
13	Assessment methods accurately assess what I have learned in this course.
14	Grades are assigned fairly.
15	The basis for assigning grades is clearly explained.
16	The instructor provides feedback on my progress in the course on a regular basis.
	Student Perceptions of Learning
17	My instructor advances my knowledge of course content.
18	My instructor promotes my understanding of important conceptual themes.
19	My instructor enhances my capacity to communicate effectively about the course subject matter.
20	My instructor encourages me to value new viewpoints related to the course.

Table 2 presents the response averages for the semesters used in this study. Averages were calculated for each course/semester from the responses. Totals represent the averages of all responses for a given evaluation method. The term, year, class size, delivery mechanism, number of responses and overall mean rating are indicated. In addition t-tests and Wilcoxon test results (Table 2) were used and are reported to test for differences between the online mean and the paper mean. The aggregated results (totals) are also provided in Table 2. In each semester and for the totals the mean score is higher for paper evaluations than for online evaluations. These results do not support the hypothesis (H1) of no difference being observed between online and paper evaluations and do not provide additional support of previously performed studies. The results reported in Table 2 indicate instructors receive higher evaluations using in class paper method of evaluation. Prior research has predominantly found there is no difference between the two modes of delivery of student evaluations.

Table 2. Response Rate Averages and Test Results

Term	Yr	Class Size	Delivery	Responses	Response Rate (pct)	Overall Mean Score	t-value (p-value)	Wilcoxon Statistic (p-value)
Spr	2008	37	online	11	30	3.00	7.08	6.74
Spr	2008	37	paper	29	78	3.40	(0.00)	(0.00)
Fall	2008	32	online	7	22	3.24	6.47	7.33
Fall	2008	32	paper	26	81	3.60	(0.00)	(0.00)
Spr	2009	40	online	13	33	3.52	0.32	-0.37
Spr	2009	40	paper	35	88	3.54	(0.75)	(0.71)
Fall	2009	33	online	13	39	3.46	2.97	3.02
Fall	2009	33	paper	30	91	3.58	(0.00)	(0.00)
Spr	2010	42	online	18	43	3.32	6.36	5.31
Spr	2010	42	paper	38	90	3.60	(0.00)	(0.00)
Fall	2010	33	online	23	70	3.55	3.05	3.18
Fall	2010	33	paper	33	100	3.64	(0.00)	(0.00)
Spr	2011	19	online	14	74	3.72	4.39	4.46
Spr	2011	19	paper	18	95	3.86	(0.00)	(0.00)
Totals		236	online	99	42	3.43	9.33	8.89
		236	paper	209	89	3.59	(0.00)	(0.00)

Table 2 also shows 6 of the 8 online response rates were below 43 percent and half of these are below 34 percent. These response rates are below generally accepted criteria for decision-making (Theall & Franklin, 1991) and lead to further questions regarding the use of online evaluations when evaluating faculty. The other key result is the low response rates of the online evaluation relative to the paper evaluations. Online response rates ranged from 22 percent to 74 percent with a total response rate over the seven semesters of 42 percent while paper evaluations ranged from 78 to 100 percent with a total response rate over the seven semesters of 89 percent. The only semester found to not have statistically significant difference (test results not reported) in the response rate was Spring of 2011. Cumulative results indicated a statistically significant difference in response rates. The results do not support hypothesis H0 but do support H1 that hypothesized there would be differences in response rates.

Table 2 indicates for this course, the online evaluations are significantly lower than paper evaluation in six of seven semesters. (Note 1) For this particular department and the criteria in the departments tenure and promotion documents in the spring of 2008, fall 2008 and spring 2010 the online results indicate the faculty member only met expectations (average scores between 3.00 and 3.39). Paper scores are all above 3.4 which exceed the department's expectations for this component of teaching evaluations. This result could potentially play a critical role in tenure, promotion and merit raise consideration.

The savings of administering online evaluations potentially results in higher costs to individual faculty in terms of tenure and promotion decisions. This is potentially more of an issue if institutions and departments have not adjusted scales to define exceeding, meeting and not meeting expectations. Another observation from the results in Table 2 is despite higher response rates for online student assessments in the fall of 2010 and spring of 2011 (70 and 74 percent respectively), paper assessments still resulted in statistically significant higher overall mean assessment results.

Table 3 provides the results for individual questions. The results indicate the ratings are higher for each of the 20 questions when the evaluations were completed on paper. The higher ratings were statistically significant for 13 of the 20 questions. The largest difference between classroom paper and online evaluations was 0.29 (question 1).

Table 3. Means, t-Test and Wilcoxon Statistic Results by Question

No.	Category and Question	Paper Mean	On line Mean	t-value (p-value)	Wilcoxon Statistic (p-value)
	Organization and Clarity				
1	My instructor is well prepared for class meetings.	3.87	3.58	5.44(0.00)	5.34 (0.00)
2	My instructor explains the subject manner clearly.	3.67	3.42	3.51 (0.00)	3.38 (0.00)
3	My instructor clearly communicates course goals and objectives.	3.68	3.51	2.54 (0.01)	2.56 (0.01)
4	My instructor answers questions appropriately.	3.62	3.45	2.31 (0.02)	1.99 (0.05)
	Enthusiasm and Intellectual Stimulation				
5	My instructor is enthusiastic about teaching the course.	3.67	3.49	2.58 (0.01)	2.62 (0.01)
6	My instructor presents the subject in an interesting manner.	3.41	3.27	1.54 (0.12)	1.14 (0.25)
7	My instructor stimulates my thinking.	3.52	3.34	2.17 (0.03)	1.82 (0.07)
8	My instructor motivates me to do my best work.	3.47	3.34	1.53 (0.13)	1.44 (0.15)
	Rapport and Respect				
9	My instructor helps students sufficiently with course-related materials.	3.57	3.46	1.50 (0.13)	1.40 (0.16)
10	My instructor is regularly available for consultation.	3.57	3.44	1.76 (0.08)	1.43 (0.15)
11	My instructor is impartial in dealing with students.	3.57	3.34	2.78 (0.01)	2.76 (0.01)
12	My instructor respects opinions different from his or her own.	3.58	3.42	2.23 (0.03)	2.15 (0.03)
	Feedback and Accessibility				
13	Assessment methods accurately assess what I have learned in this course.	3.52	3.35	2.02 (0.04)	1.89 (0.06)
14	Grades assigned fairly.	3.65	3.46	2.57 (0.01)	2.25 (0.02)
15	The basis for assigning grades is clearly explained.	3.60	3.47	1.69 (0.09)	2.05 (0.04)
16	The instructor provides feedback on my progress in the course on a regular basis.	3.53	3.41	1.43 (0.15)	1.47 (0.14)
	Student Perceptions of Learning				
17	My instructor advances my knowledge of course content.	3.62	3.46	2.26 (0.02)	2.12 (0.03)
18	My instructor promotes my understanding of important conceptual themes.	3.57	3.47	1.37 (0.17)	1.11 (0.27)
19	My instructor enhances my capacity to communicate effectively about the course subject matter.	3.54	3.46	1.01 (0.31)	0.77 (0.44)
20	My instructor encourages me to value new viewpoints related to the course.	3.55	3.46	1.20 (0.23)	1.27 (0.21)

The five factors (organization and clarity, enthusiasm and intellectual stimulation, rapport and respect, feedback and accessibility and students perceptions of learning) are also analyzed by combining the results of questions

from each of the factors as identified in Table 1.

Table 4 summarizes the results for the factors and similar to results in Tables 2 and 3, the paper evaluations produce significantly higher ratings then online evaluations in each of the five categories.

Table 4. Test Results by SAI Factors

Factor	Paper Mean	On line Mean	t-value (p-value)	Wilcoxon Statistic (p-value)
Organization and Clarity	3.71	3.49	6.53	6.36
			(0.00)	(0.00)
Enthusiasm and Intellectual Stimulation	3.52	3.36	3.78	3.42
			(0.00)	(0.00)
Rapport and Respect	3.57	3.42	4.18	3.89
			(0.00)	(0.00)
Feedback and Accessibility	3.57	3.43	3.82	3.82
			(0.00)	(0.00)
Student Perceptions of Learning	3.57	3.47	2.92	2.62
			(0.00)	(0.01)

Additional analysis was performed to examine the proportions of each of the responses. Clason and Dormody (1993) discussed some of the issues related to the analysis of Likert scale data. The underlying issue is whether the Likert scale is an ordinal scale or an interval scale and the resulting inferential errors that can result from improper analysis if the underlying variables are continuous.

To address these issues, this study recognizes the discrete nature of the observation and also summarizes and performs the analysis on the counts (percentages) observed in the data. The proportions for each term and factor are graphed in Figures 1 through 12. The figures show the largest difference in responses is fewer “4s” are received in the online ratings. This is observed for each term except the Spring 2009 term. This result of fewer 4’s is observed in each of the five categories the assessment is designed to address. This result is similar to results found in the marketing literature that has found unsatisfied customers are more likely to respond or complain about product or service issues.

Table 5 provides the results of chi-squared tests and provides evidence the distributions between paper and online evaluations are different for each semester and confirms the differences observed in Figures 1 through 12.

The individual question response percentages were aggregated by category as shown in Table 6 and Figures 7 through 12. Chi-squared test results are reported in Table 6 and provide additional evidence the paper responses are not the same as the online responses. The percentage responses result in fewer 3’s (agree) and more 4’s (strongly agree) when using paper evaluations.

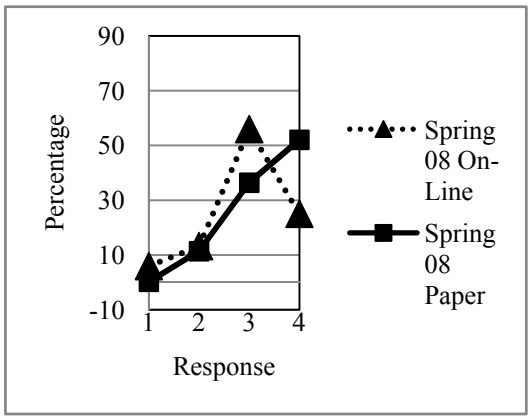

Figure 1. Spring 2008

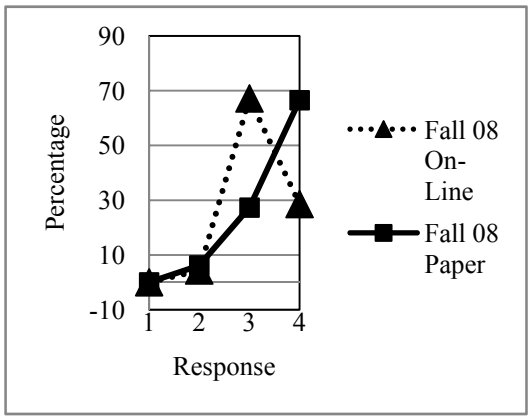

Figure 2. Fall 2008

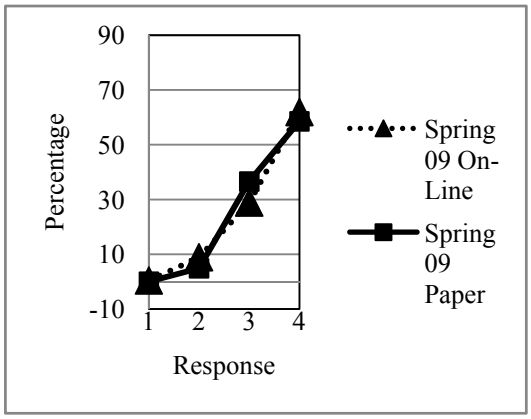

Figure 3. Spring 2009

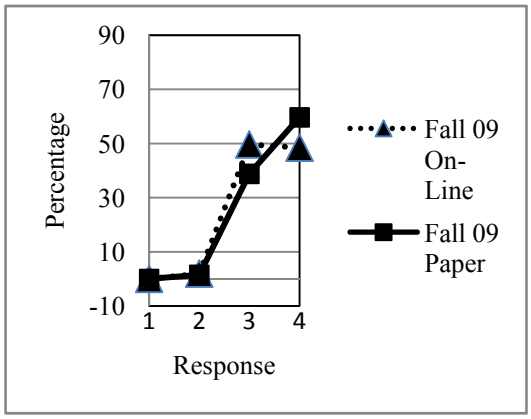

Figure 4. Fall 2009

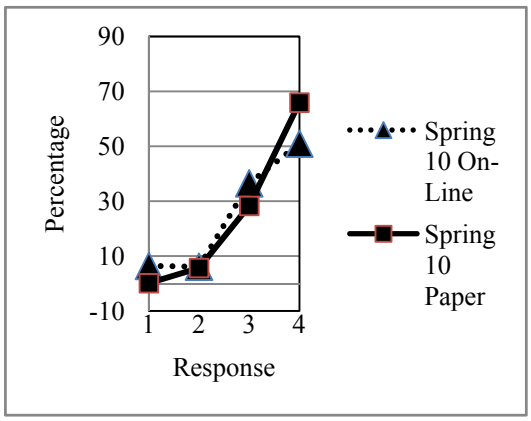

Figure 5. Spring 2010

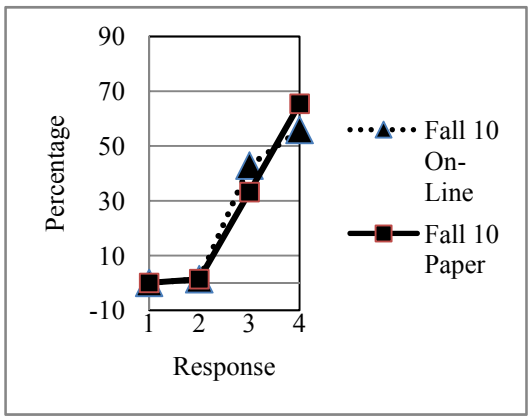

Figure 6. Fall 2010

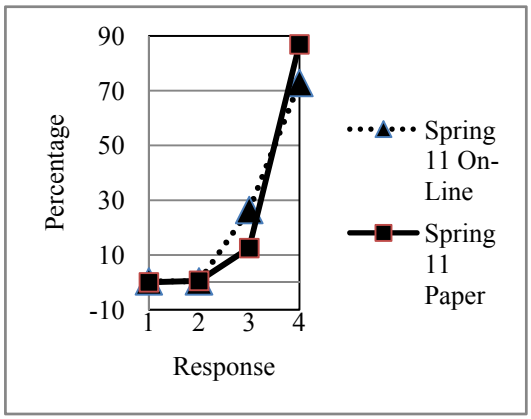

Figure 7. Spring 2011

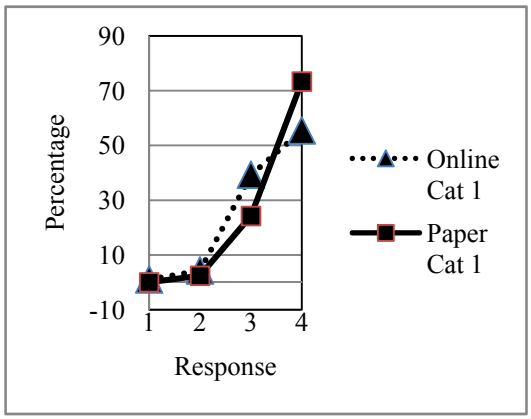

Figure 8. Category I

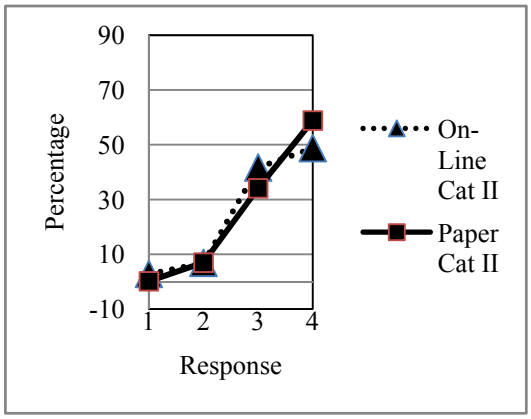

Figure 9. Category II

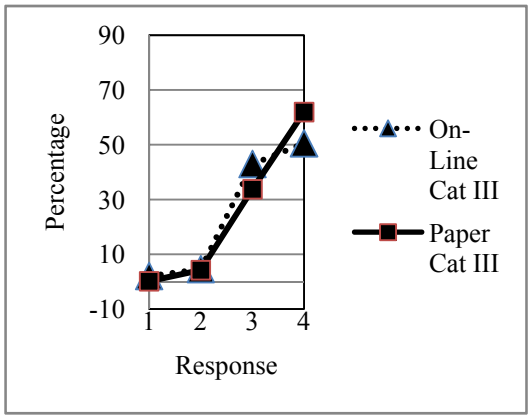

Figure 10. Category III

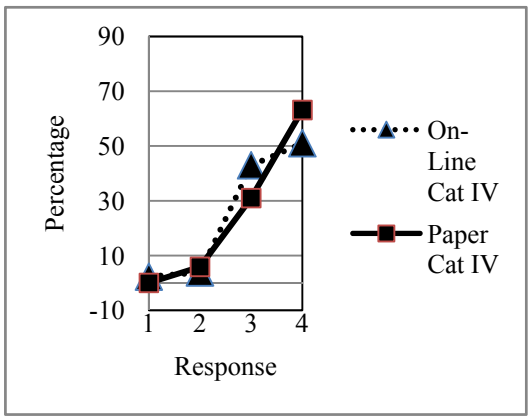

Figure 11. Category IV

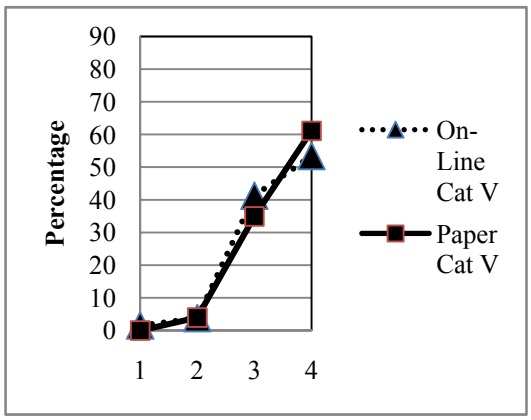

Figure 12. Category V

6. Limitations

One limitation of this research involves the samples size(s) used in some of the analysis. Some of the response rates on an individual course level (particularly online response rates) resulted in limited data which could create issues with analysis. The authors understand this limitation but also believe this is an additional indication of potential issues with the use of online evaluations for evaluating faculty when response rates are low.

The authors recognize previously enrolled students could have informed current students paper evaluations would also be given for the class. This could have resulted in a lower on-line response rate. Anecdotally the authors did not have any reports of this occurring, but understand it is a potential limitation of this research.

Table 5. On-line vs. Paper Evaluation by Term

Term	Year	Delivery	Pct. 1	Pct. 2	Pct. 3	Pct. 4	Pearson Chi-Squared (p-value)	Likelihood Ratio Chi-squared (p-value)
Spring	2008	On-line	5.9	13.2	55.9	25.0	93.04	93.40
Spring	2008	Paper	0.2	11.4	36.4	52.1	(0.00)	(0.00)
Fall	2008	On-line	0.0	4.3	67.1	28.6	113.73	97.91
Fall	2008	Paper	0.0	6.2	27.3	66.5	(0.00)	(0.00)
Spring	2009	On-line	0.4	8.9	28.9	61.9	11.18	12.59
Spring	2009	Paper	0.0	4.9	36.6	58.6	(0.01)	(0.00)
Fall	2009	On-line	0.0	1.9	49.6	48.5	12.95	13.00
Fall	2009	Paper	0.0	1.5	38.8	59.7	(0.00)	(0.00)
Spring	2010	On-line	6.4	6.1	36.5	51.0	45.54	67.71
Spring	2010	Paper	0.1	5.7	28.3	65.9	(0.00)	(0.00)
Fall	2010	On-line	0.0	1.3	42.8	55.9	17.36	17.79
Fall	2010	Paper	0.0	1.4	33.2	65.5	(0.00)	(0.00)
Spring	2011	On-line	0.4	0.4	26.4	72.9	30.01	35.38
Spring	2011	Paper	0.0	0.6	12.5	86.9	(0.00)	(0.00)

Table 6. Aggregated Results by Question Category

	Delivery	Pct 1	Pct 2	Pct 3	Pct 4	Pearson Chi-Squared (p-values)	Likelihood Ratio Chi-squared (p-values)
Cat I	On-line	1.01	4.30	39.24	55.44	13.66	15.26
Cat I	Paper	0.00	2.39	24.25	73.36	(0.00)	(0.00)
Cat II	On-line	2.78	6.82	41.67	48.74	6.56	7.55
Cat II	Paper	0.12	2.39	24.25	73.36	(0.09)	(0.06)
Cat III	On-line	2.02	4.55	42.93	50.51	6.46	6.46
Cat III	Paper	0.00	5.86	30.98	63.16	(0.09)	(0.09)
Cat IV	On-line	2.27	3.79	42.93	51.01	9.17	11.20
Cat IV	Paper	0.00	5.86	30.98	63.16	(0.03)	(0.01)
Cat V	On-line	1.52	3.80	41.27	53.42	4.09	6.08
Cat V	Paper	0.00	3.95	34.93	61.12	(0.25)	(0.10)

7. Implications and Conclusions

The use of online evaluations has gained usage due to the ease and reported cost savings of an electronic/online assessment instrument. The use of online assessments will likely continue to grow and become more widespread. Faculty and administrators should be aware of issues related to how the scores are analyzed (means vs. counts), and based on this study, the lower assessment scores received by faculty. Guidelines and standards may need to be adjusted for online evaluations as scores are significantly lower than paper responses.

In addition, there may be a similar effect present for students as found with consumer dissatisfaction and variation of product categories (Best & Andreasen, 1977). Thus, courses which are considered to be unimportant or unrelated to the student's major may represent courses for which students are more dissatisfied. Anderson (1998) also found the most dissatisfied customers are the ones most likely to complain. This study cannot address those issues specifically but if one considers the type of course as the product in which the customer (student) is evaluating in combination with the finding that the most dissatisfied customers are one of the groups most likely to discuss their issues, this produces a possible explanation for the lower scores observed in this study. One method of potentially considering this issue would have been the effect of expected grades. However, the standard instrument provided by the university did not address grade expectations. The course evaluated had rather specific criteria. The online evaluations require students to use their own time to complete the evaluations and lack the social pressure of completing the evaluation in a course setting. Students who strongly dislike the class would be in one of the categories noted by Andersen (1998) and Best and Andreasen (1977).

This study has demonstrated evaluation scores are lower using an electronic/online assessment than paper and pencil assessment for a finance course taught to non-finance majors. Scores were lower each semester for most of the factors the assessment is designed to evaluate. Evaluations were lower using online evaluations by as much as 10 percent of the scale relative to paper evaluations. Proportions of scores for the 4-point Likert scale were also lower using the online evaluations. These results were consistent even in semesters when response rates for the online evaluations were relatively high (greater than or equal to 70 percent). In each term studied, the response rate was always higher using the classroom/paper-based evaluations. The study is limited due to the data being from one class and one instructor and may not be generalizable.

The evidence presented in this study indicated online evaluations resulted in lower teaching evaluation scores. This result could be impacted by the low response rates to the online evaluation. The lower ratings could be the result of the low response rates and if so this would still lead to questioning the use of the online evaluation if response rates are low. If the reduction in scores is uniform throughout a department, the effect on merit raises could be insignificant. However, if there are standards used to evaluate faculty, and those standards define acceptable teaching scores, and if they have not been adjusted to account for situational variables such as lower scores from online evaluations, then tenure, promotion and merit pay processes could be adversely affected.

The research regarding the use of online course evaluations has not incorporated specifics related to course characteristics. This study employed a very specific sample and course characteristics and raises questions regarding the potential effect of the course characteristics. Future research should carefully consider the characteristics of the course and the potential impact course characteristics have student teacher evaluations.

References

Abrami, P. C., d'Appollonia, S., & Cohen, P. A. (1990). The validity of student ratings on instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219-231. http://dx.doi.org/10.1037/0022-0663.82.2.219

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research*, *52*(3), 446-464. http://dx.doi.org/10.2307/1170425

Anderson, E. W. (1998). Customer satisfaction and word-of-mouth. *Journal of Service Research*, *1*(1), 5-17. http://dx.doi.org/10.1177/109467059800100102

Arreola, R. A. (2007). *Developing a Comprehensive Faculty Evaluation System: A Guide to Designing, Building, and Operating Large-Scale Faculty Evaluation Systems* (3rd ed.). Bolton, Mass: Anker Publishing Company.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations?. *The Journal of Economic Education*, *37*(1), 21-37. http://dx.doi.org/10.3200/JECE.37.1.21-37

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors

rated differently?. *Journal of Educational Psychology*, *79*(3), 308-314. http://dx.doi.org/10.1037/0022-0663.79.3.308

Best, A., & Andreasen, A. R. (1977). Consumer re-sponse to unsatisfactory purchases: A survey of perceiving defects, voicing complaints, and obtaining redress. *Law & Society*, *11*, 701-742.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

Chan, D. W. (2004). Perceived emotional intelligence and self-efficacy among Chinese secondary school teachers in Hong Kong. *Personality and Individual Differences*, *36*(8), 1781–1795. http://dx.doi.org/10.1016/j.paid.2003.07.007

Clason, D. L., & Dormody, T. J. (1993). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, *35*(4), 31-35. http://dx.doi.org/10.5032/jae.1994.04031

Doyle, K. O., Jr. (1983). *Evaluating Teaching*. Lexington, MA: Lexington Books.

Emery, C. R. (1995). *Student evaluations of faculty performance*. Unpublished manuscript, Clemson University, Clemson, SC.

Emery, C. R., Kramer, T. R., & Tian, R. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*(1), 37-46. http://dx.doi.org/10.1108/09684880310462074

Ewell, B. (2000). *The United States Air Force Academy (AFA) desired to convert its mid- and end-of course evaluations from paper-pencil to computer administered methodology*. Retrieved April 18, 2012 from http://home.att.net/~bobewell/oleval.htm

Feldman, K. A. (1998). Reflections on the effective study of college teaching and student ratings: One continuing quest and two unresolved issues. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research*. New York: Agathon Press. http://dx.doi.org/10.1007/978-94-011-3971-7_2

Franklin, J., & Theall, M. (1992). *Disciplinary differences*, *instructional goals and activities*, *measures of student performance*, *and student ratings of instruction*. Paper presented at the Seventy-Third Annual Meeting of the American Educational Research Association, San Francisco.

Greenwald, A. G., & Gilmore, G. M. (1997). Grading Leniency is a Removable Containment of Student Ratings, *American Psychologist*, *52*(11), 1209-1217. http://dx.doi.org/10.1037/0003-066X.52.11.1209

Hardy, N. (2003). Online ratings: Fact or fiction. In D. L. Sorenson, & T. D. Johnson (Eds.), *New Directions for Teaching and Learning: Online Student Ratings of Instruction*, *96*. San Francisco: Jossey-Bass.

Johnson, V. E. (2002, April 14). An A is an A is an A and that's the problem. *New York Times Education Today*, p. 14.

Koushki, P. A., & Kuhn, H. A. (1982). How reliable are student evaluations of teachers?. *Engineering Education*, *72*, 362-67.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* (pp. 9-25). San Francisco: Jossey-Bass.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional students rating of instruction. *Research in Higher Education*, *40*(2), 221-232. http://dx.doi.org/10.1023/A:1018738731032

Linse, A. R. (2010). *Online Student Ratings of Teaching Effectiveness: Analysis of Data from Select Semesters (2009-2010)*. Report prepared for the Committee on Faculty Affairs of the University Faculty Senate, the Pennsylvania State University. Retrieved April 18, 2012 from http://www.srte.psu.edu/pdf/Online_vs_Paper_Fall2010.pdf

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology*, *75*, 150-166. http://dx.doi.org/10.1037/0022-0663.75.1.150

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*. New York: Springer. http://dx.doi.org/10.1007/1-4020-5742-3_9

Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook on Theory and Research* (pp. 143-234). New York: Agathon Press.

Neumann, R. (2000). Communicating student evaluation of teaching results: Rating interpretation guides (RIGs). *Assessment & Evaluation in Higher Education*, *25*(2), 121-134. http://dx.doi.org/10.1080/02602930050031289

Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, *13*, 519-527.

Rosenberg, M. E., Watson, K., Paul, J., Miller, W., Harris, I., & Valdivia, T. D. (2001). Development and implementation of a Web-based evaluation system for an internal medicine residency program. *Academic Medicine*, *76*, 92–95. http://dx.doi.org/10.1097/00001888-200101000-00024

Safer, A., Farmer, L. S. J., Segalla, A., & Elhoubi, A. F. (2005). Does the distance from the teacher influence student evaluations?. *Educational Research Quarterly*, *28* (3), 27-34.

Scriven, M. (1994). Using student ratings in teacher evaluation. *Evaluation Perspectives*, *4*(1), 1-6.

Seldin, P. (1991). The teaching portfolio: A practical guide to improved performance and promotion/tenure decisions. *New Directions in Teaching and Learning*. Bolton, MA: Anker Publications.

Seldin, P. (1993). The use and abuse of student ratings of instruction. *The Chronicle of Higher Education*, A-40.

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, *19*(2), 174-197. http://dx.doi.org/10.1111/j.1559-1816.1989.tb00051.x

Theall, M., & Centra, J. A. (2001). Assessing the scholarship of teaching: Valid decisions from valid evidence. *New Directions for Teaching and Learning*, *86*, 31-43. http://dx.doi.org/10.1002/tl.14

Theall, M., & Franklin, J. (1990). Student ratings of instruction: Issues for improving practice. *New Directions for Teaching and Learning*, *43*. San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. L. (1991). Using student ratings for teaching improvement. In M. Theall & J. Franklin (Eds.), R. J. Menges (Series Ed.), *Effective Practices for Improving Teaching* (*New Directions for Teaching and Learning*, *48*). San Francisco: Jossey Bass.

Theall, M., Abrami, P. A., & Mets, L. (Eds.). (2001). The student ratings debate: Are they valid? How can we best use them?. *New Directions for Institutional Research*, *109*. San Francisco: Jossey Bass.

Theall, M., Abrami, P. C., Arreola, R., Franklin, J., Nuhfer, E., & Scriven, M. (2005). Valid Faculty Evaluation Data: Are There Any?. Paper presented at AERA Annual Meetings Program Interactive Panel Presentation, American Educational Research Association Symposium, Montreal.

Wines, W. A., & Lau, T. J. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decision and its implications for academic freedom. *William & Mary Journal of Women and Law*, *13*(1), 167-202.

Note

Note 1. Results were significantly different for the fall of 2007 as well. However, the student assessment instrument that term used a five point scale, while all the results reported are from semesters when a four point scale was used.