SAN JOSÉ STATE UNIVERSITY
OIRblog
Institutional Research Newsletter
www.oir.sjsu.edu/OIRblog

In this issue:

The Newsletter for the Office of Institutional Research at San José State University Spring 2011

Student Evaluations of Teaching: A Comprehensive Study of Online versus Paper Modes

During the past decade the most important change in teaching evaluations has been the switch from paper evaluations (PE) to online evaluations (OE) (Lovric, 2006). Colleges and universities may hesitate to implement this change due to "comparability of results" (Donovan, Mader, and Shinsky, 2006), specifically, there was concern that OE maybe more negative than PE. There are many studies comparing student responses between PE and OE (Avery, Bryant, Mathios, Kang, and Bell, 2006; Cao, Clark, Schrimer, and Nelson, 2007; Chang, 2003; Donovan et al., 2006; Hardy, 2003; Heath, Lawyer, and Rasmussen, 2007; Laubsch, 2006; Spooner, Jordan, Algozzine, and Spooner, 1999). However, these evaluations were limited in the number of teachers evaluated, the number of students who participated in the evaluations, and the length and breadth of the courses the teachers taught. This study's goal was to be the most comprehensive in evaluating student responses between PE and OE to date so that other colleges and universities can apply the results to their circumstances.

I. Pros and Cons of Online Evaluations versus Paper Evaluations

The majority of studies on teaching evaluation modes have found no significant difference in the total quantitative evaluation scores between OE and PE (Donovan et al., 2006; Hardy, 2003; Heath, et al., 2007; Laubsch, 2006; Spooner, et al., 1999). Some studies have found that an overarching question (overall effectiveness) is answered more favorably by OE students, with the rest of the questions showing no significant difference (Liu, 2006). Other studies have found no significant difference in the total mean quantitative score, but show differences when comparing individual questions (Avery et al., 2006; Cao, et al., 2007). Not all studies have found that OE is either positive or neutral. Chang (2003) found that PE produced higher scores for individual questions and total scores.

One of the most popular OE, *RateMyProfessor.com*, was found to have comments similar to traditional teaching evaluations (Silva, Silva, Quinn, Draper, Cover, and Munoff, 2008). *RateMyProfessor.com* responses were highly correlated to PE responses at a well-known college (Brown, Baille, and Fraser, 2009). *RateMyProfessor.com* does not conduct a controlled survey (Silva et al., 2008) yet it produces similar results. Among OE advantages are more efficiency and accessibility (Krajewski & Pike, 2005), reduction in time and cost for managing and recording the results (Lovric, 2006; Anderson et al., 2005), ease of use (Avery et al., 2006; Ha, Mars, and Jones, 1998; Layne, DeCristoforo, and McGinty, 1999), and better flexibility for student answers to open-ended questions (Gamliel & Davidovitz, 2005). Lastly, if paper surveys are stored, that storage space can be used for other purposes.

There are drawbacks to using OE, including low response rates (Cao et al., 2007; Dommeyer, Baum, and Hanna, 2002; Heines & Martin, 2005; Krajewski & Pike, 2005; Lovric, 2006) and student concern with anonymity (Laubsch, 2006, Heines & Martin, 2005; Dommeyer et al., 2002). College faculty were also concerned with security of results (Heines & Martin, 2005). Faculty who are not proficient with computers feel that they are at a disadvantage (Anderson et al., 2005). Finally, faculty perceive traditional mode (PE) to be more accurate (Donovan et al., 2006).

II. Online and Paper Evaluation Procedures

The study was conducted at a comprehensive public, urban university in the West, enrolling nearly 32,000 students with approximately 20,000 of them full-time undergraduates. The university consists of seven colleges. The administration of the same university-wide teaching evaluation form is required in all colleges at the end of each semester. The in-class assessment was coordinated by an assigned proctor. A checklist that fully and clearly describes the student's procedure to be followed was given in advance.

The evaluation instrument consisted of four sections. The first section (13 questions) focused on the instructor's teaching effectiveness, with the 13th question asking about an overall effectiveness. The items ranged from a low of 1 (strongly disagree) to a high of 5 (very strongly agree), the majority of the ratings are in the 3-5 range. The next section asked about student's expected overall grade in the course, their year in college, and questions about undue influence from other students and instructors. The third section allowed faculty to

include their own questions. If no additional queries were provided, this section was left blank. The final section had three open-ended questions regarding the strengths of the instructor's teaching, weaknesses and/or areas in need of improvement, and other comments. The two delivery modes had identical formats.

To examine whether there were significant differences in student responses between paper and online evaluations, this pilot project was conducted in the spring semester of 2009. The request to participate in the evaluation was issued in the Student Administrative System.

The compared OE vs. PE evaluations were from the same courses and taught by the same faculty during the last five semesters. If multiple PE counterparts were identified, the most recent evaluation was selected. If a course with multiple sections was taught by the same faculty in a selected semester, all of those sections were used. All OE that did not have PE counterparts, or had only two or fewer responses (per section) were excluded from the analysis. Any "Not Applicable/No Opportunity to Observe" responses were also excluded from both the descriptive analysis and the comparison of means (t-test and ANOVA).

The impacts of the migration from paper to online student evaluation of teaching effectiveness are gaining greater attention, and are now becoming clearer in higher education. However, there has been limited empirical research on this topic (Liu, 2006; Lovric, 2006; Cao, et al., 2007; Heath et al., 2007). This study provides empirical data to determine whether there were differences in student responses between online and paper course evaluations. A total of 291 course sections (52% for OE and 48% for PE) were used in this analysis. The student responses consisted of 4,654 students (32% OE and 68% PE) from undergraduate lower and upper divisions, as well as graduate courses. Response rates for PE were much higher than for OE surveys, in fact more than double. Overall the response rate for PE was 73% compared to 31% for OE.

III. Online Evaluation Pilot Study Findings

Is there a difference in student's ratings (average scores) between online and paper evaluations?

Because an independent variable consisted of two values (online and paper evaluations), with between-group design, and the dependent variable (evaluation scores) is normal or scale data, the independent sample t-test was selected. This technique to examine the mean differences of two independent groups is consistent with previous studies (Avery et al., 2006; Cao et al., 2007; Chang, 2003; Donovan et al., 2006; Liu, 2006; Spooner et al., 1999).

In this study, the survey instrument asked 13 multiple choice questions that the first 12 questions assessed particular aspects of teaching and the last question examined the overall effectiveness. The data in Table 1 indicates that OE students rated their faculty more favorably in four questions (Responsive to Questions and Comments; Facilitated Learning; Approachable for Assistance; and Responsive to Diversity). The remaining nine questions, including an Overall Rating, revealed no significant difference between these two delivery modes.

Table 1.

Mean Comparisons of Evaluation Questions

Question[1]	Online Evaluation	Paper Evaluation	t-test
Relevance of Course Content	4.40	4.40	0.253
Enhanced Learning	4.20	4.16	1.284
Emphasized Points	4.29	4.25	1.529
Responsive to Questions	4.43	4.35	2.611*
Facilitated Learning	4.29	4.23	2.041*
Approachable for Assistance	4.40	4.32	2.855**
Responsive to Diversity	4.44	4.38	2.177*
Interest in Teaching	4.51	4.49	0.778
Intellectual Challenge	4.14	4.12	0.733
Fair Grading	4.23	4.20	0.778
Analysis of Ideas	4.19	4.13	1.925
Meaningful Feedback	4.15	4.10	1.318
Overall Rating	4.38	4.37	0.210

* $p < 0.05$; ** $p < 0.005$

1 = refer to Appendix for full question description

Are there differences in student's ratings (evaluation scores) varying on survey delivery modes and/or course levels?

The analysis of variance (ANOVA) was used to examine whether the mean of the evaluation scores were equivalent between delivery mode and course level. In this study the ratings on the thirteen questions were analyzed using a 2 X 3 [(survey delivery mode) X (course level)] ANOVA. The main effect of delivery mode, shown in Table 2, indicated that only one question (Approachable for Assistance) was rated significantly different between PE and OE students. The main effect of course level showed a significant difference in evaluation scores for three questions (Relevance of Course Content; Intellectual Challenge; and Overall Rating).

Table 2.

Analysis of Variance:

Survey Delivery Mode and Course Level - F values

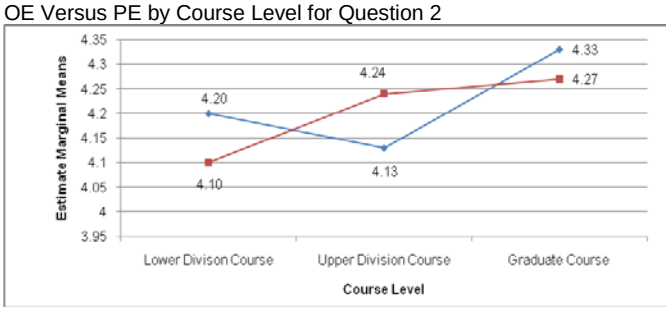
Question[1]	Main Effect: Survey Delivery Mode	Main Effect: College	Interactive Effect
Relevance of Course Content	0.09	3.15*	1.42
Enhanced Learning	0.06	2.98	4.97*
Emphasized Points	0.91	2.78	0.16
Responsive to Questions	2.33	0.56	1.41
Facilitated Learning	0.46	0.48	3.22*
Approachable for Assistance	6.55*	2.41	2.52
Responsive to Diversity	2.68	0.56	2.62
Interest in Teaching	0.04	1.37	1.18
Intellectual Challenge	0.09	4.64*	2.57
Fair Grading	0.43	0.40	7.07**
Analysis of Ideas	1.29	0.56	3.32*
Meaningful Feedback	1.19	0.35	1.36
Overall Rating	0.04	5.20*	0.167

* $p < 0.05$; ** $p < 0.005$

1 = refer to Appendix for full question description

The interactive effect was statistically significant on four questions (Enhanced Learning; Facilitated Learning; Fair Grading; and Analysis of Ideas). This means that the "effect" of survey delivery mode on evaluation scores for these questions depends on which course level is being considered. For example, student's responses on teaching effectiveness regarding "Used assignments to enhance learning" in Figure 1, the results suggest that evaluation scores from OE students were relatively lower for lower division courses and higher for upper division courses. The ratings for graduate courses were inconclusive.

Figure 1. Estimated Marginal Means of OE Versus PE by Course Level for Question 2



Are there differences in student's ratings (evaluation scores) varying on survey delivery modes and/or colleges?

A 2 X 7 [(survey delivery mode) X (college)] ANOVA was used to answer these questions. Similar to responses for Questions 2 and 3, data in Table 3 suggests evaluation scores were significantly different for PE and OE students on only one question (Approachable for Assistance). In regards to the college main effect, students in each college rated their instructors differently for all thirteen questions. Only one question (Provide meaningful feedback) shows significant interactive effect on evaluation scores between survey delivery mode and college.

Table 3.

Analysis of Variance:

Survey Delivery Mode vs. College - F values

Question[1]	Main Effect: Survey Delivery Mode	Main Effect: College	Interactive Effect
Relevance of Course Content	0.08	4.77***	0.25
Enhanced Learning	0.87	3.69**	0.41
Emphasized Points	0.51	7.15***	1.38
Responsive to Questions	1.45	3.44**	1.19
Facilitated Learning	0.36	5.88***	1.59
Approachable for Assistance	4.86*	3.66**	1.05
Responsive to Diversity	1.75	7.40***	0.63
Interest in Teaching	0.34	6.46***	1.43
Intellectual Challenge	1.78	2.42*	1.02
Fair Grading	0.17	9.70***	0.32
Analysis of Ideas	1.61	4.93***	1.14
Meaningful Feedback	0.29	5.99***	2.40*
Overall Rating	0.40	6.87***	1.02

* $p < 0.05$; ** $p < 0.005$; *** $p < 0.0005$

1 = refer to Appendix for full question description

Are there differences in student's ratings (evaluation scores) varying on survey delivery modes and/or subject areas?

Thirty-three subject areas were represented in this study. A 2 X 33 [(survey delivery modes) X (subject areas)] ANOVA was selected to quantify whether the evaluation scores were influenced by differences in delivery mode and subject area. Table 4 shows that there was no significant difference in the main effect of survey delivery for any of the thirteen questions. However, the main effect for subject areas displayed the opposite. There was a significant difference in student's ratings varying on subject areas.

The interaction between survey delivery modes and course subject is statistically significant for all thirteen questions. Interestingly, the "effect" of survey delivery modes on evaluation scores depends on which subject area is being considered.

Table 4.

Analysis of Variance:

Survey Delivery Modes vs. Subject areas - F values

Question[1]	Main Effect: Survey Delivery Mode	Main Effect: College	Interactive Effect
Relevance of Course Content	0.57	5.93***	2.23***
Enhanced Learning	1.02	5.48***	1.90**
Emphasized Points	1.59	7.40***	2.18***
Responsive to Questions	1.26	9.97***	2.48***
Facilitated Learning	0.15	10.77***	2.78***
Approachable for Assistance	1.65	10.80***	1.80**
Responsive to Diversity	1.38	7.15***	1.98**
Interest in Teaching	0.00	7.25***	2.61***
Intellectual Challenge	0.30	6.14***	1.77*
Fair Grading	0.05	9.38***	2.01**
Analysis of Ideas	0.36	8.66***	1.92**
Meaningful Feedback	0.20	11.39***	2.51***
Overall Rating	0.02	10.02***	3.38***

* $p < 0.05$; ** $p < 0.005$; *** $p < 0.0005$

1 = refer to Appendix for full question description

Are there differences in student's ratings (evaluation scores) varying on Grade and Classification?

Two background questions were asked to determine the differences in academic achievement and student classification. The first question asked what grade that student expected to receive in the course at the end of the semester. Only five-letter grades (A, B, C, and D or F) were considered in this comparison (see Figure 2). Table 5 indicates a significant difference in expected grades between PE and OE students (t = 6.050, $p <$ .0005). In other words, OE students expected to receive a higher grade. Another question asked for their current undergraduate classification: freshman, sophomore, junior, or senior (See figure 3). Table 6 suggests a significant difference in self-reported classification (t = -3.871, $p <$ .0005). Thus, PE students tend to be in a higher classification than OE students in this study. Table 6 suggests a significant difference in self-reported classification (t = -3.871, $p <$ .0005). Thus, PE students tend to be in a higher classification than OE students in this study.

Table 5.

Current estimate of overall grade in class

	Online Evaluation		Paper Evaluation	
Grade	No.	%	No.	%
A (4)	416	38%	824	33%
B (3)	498	46%	1,170	47%
C (2)	160	15%	472	19%
Either D or F (1)	16	1%	47	2%
Total	1,090		2,513	
Mean	3.20		3.10	
t-test		6.050*		

* $p < 0.0005$

Figure 2. Current estimate of overall grade in class

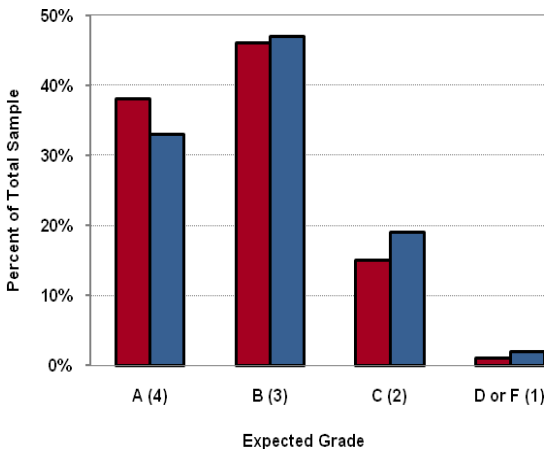


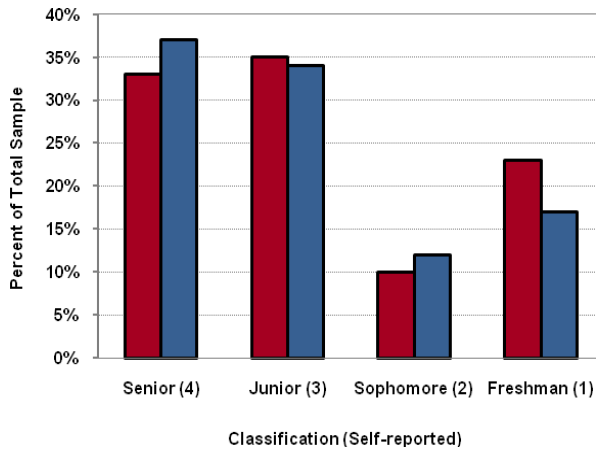
Table 6.

Self identification of college year

	Online Evaluation		Paper Evaluation	
Classification	No.	%	No.	%
Senior (4)	357	33%	959	37%
Junior (3)	379	35%	866	34%
Sophomore (2)	110	10%	309	12%
Freshman (1)	250	23%	425	17%
Total	1,096		2,559	
Mean	2.77		2.92	
t-test		-3.871*		

* $p < 0.0005$

Figure 3. Self identification of grade (classification) level



The study results suggest for a majority of the questions no significant difference between the two delivery modes (OE vs. PE) of teaching evaluation. This supports findings in previous studies (Donovan et al., 2006; Hardy, 2003; Heath et al., 2007; Laubsch, 2006; Spooner et al., 1999). When there was a difference, it was found that OE produced higher evaluation results than PE.

When delivery mode and course level were compared, the delivery mode main effect suggests no significant difference in student's responses between online and paper evaluations for 12 out of the 13 questions. The main effect of course levels show significant difference in only three questions in student's responses among lower and upper divisions, and graduate.

However, conclusions differed when the comparison between delivery modes and colleges. As in Layne, et al. (1999) study a significant difference exists in the main effect of colleges. This means that courses in each college are rated differently by the students. This difference may occur due to student expectations, teaching methodology due to specialized subject matter, and the subject rigor and scholarship differences in each college (Clayson, 2009). Lastly differences in ratings may occur because of the reciprocity between student evaluation and the student's expected grade (Clayson, 2007).

When addressing subject areas, a significant difference was found in the main effect of subject areas. A significant difference appeared in the interactive effect between the mode of delivery and the subject areas for all the questions. Surprisingly, this interactive effect for the overall teaching effectiveness indicates that some subject areas receiving low ratings using PE had higher ratings when using the OE. Therefore, faculty in these subject areas could benefit from the switch to online mode.

Finally, there was a significant difference in the expected grade. The OE students expected to receive a higher grade. McGregor (2007) suggested that students with better grades tend to test higher in the quality of initiative. Students who took the OE survey administered outside the classroom had to initiate the process. The PE, on the other hand, was given to all students in the classroom and no initiative is needed. Therefore, the initiative to participate in online evaluation outside the classroom environment mirrored the motivation that students had toward their course work and academic performance.

IV. Evaluation Mode Proposal

Any change in the teaching evaluation process produces faculty's anxiety and tension. The most notable change during the last decade has been the migration to online evaluations. This study concluded that, for a majority of survey questions, no significant difference in student's evaluations, whether online and paper. However, an important difference was that OE evaluators rated their instructors more highly than PE evaluators. The migration to online mode seemed to produce higher ratings for some survey items and no changes in other items.

For all of the multiple choice questions, a significant difference existed in student's responses when varied from college to college. This may be due to differences between the colleges and should be taken into account when comparing faculty from one college to another.

This study also found that there was a significant difference between subject areas and the interactive effect between subject areas and mode of delivery. The interactive effect indicated that for subject areas with low ratings, the switch to online mode could produce an improvement in their ratings.

Finally, OE students expected a higher grade than PE students. This difference may be explained by the higher achieving students having more initiative overall, whether it be course work or completing faculty evaluations. Such an effort by OE students could lead to a positive correlation between grades received and responses in online evaluation.

In conclusion, this study demonstrates that student ratings of paper and online evaluations are compatible and can be used interchangeably. With the rising costs and sustainability concerns with paper evaluations, it is an appropriate time to migrate to online faculty evaluations. We recommend for further study the use of student ratings for instructional improvement and how student ratings are being used and the extent of misinterpretations and misuses.

References

Anderson, H. M., Cain, J., & Bird, E. (2005). Online student course evaluations: review of literature and a pilot study. *American Journal of Pharmaceutical Education, 69*, 34-43.

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*, 21-37.

Brown, M. J., Baille, M., & Fraser, S. (2009). Rating ratemyprofessors.com: A comparison of online and official student evaluations of teaching. *College Teaching, 57*, 89-92.

Cao, Y., Clark, A., Schrimer, J., & Nelson, M. (2007). *Online and paper course evaluations: Are the response rates and results different?* Paper presented at the Association of Institutional Research Annual Forum, San Francisco, CA.

Chang, T. S. (2003, April). *The results of student ratings: The comparison between paper and online surveys.* Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Clayson , D. E. (2007). Conceptual and Statistical Problems of using between-class data in educational research. *Journal of Marketing Education, 29*, 34-38.

Clayson , D. E. (2009). Student evaluations of teaching: Are they related to what students learn? *Journal of Marketing Education, 31*, 16-30.

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002). College student's attitude towards methods of collecting teaching evaluations: In-class versus on-line. *Journal of Education for Business, 78*, 11-15.

Donovan, J., Mader, C. E., & Shinsky, J. (2006). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning, 5*, 283-296.

Gamliel, E. & Davidovitz, L. (2005). Online versus traditional teaching evaluation: Mode can matter. *Assessment & Evaluation in Higher Education, 30*, 581-592.

Ha, T. S., Mars, J., & Jones, J. (1998). *A web-based system for teaching evaluation*. Paper presented at National Curriculum for Initial Teacher Training 1998 Conference, Hong Kong.

Hardy, N. (2003). Online ratings: fact and fiction. *New Directions for Teaching and Learning, 96*, 31-41.

Heath, N. M., Lawyer, S. R., & Rasmussen, E, B. (2007). *Teaching Psychology, 34*, 259-261.

Heines, J. M. & Martin, D. M. (2005). *Development and deployment of a web-based course evaluation system: Trying to satisfy the faculty, the students, the administration, and the union.* Paper presented at Proceedings of the First International Conference on Web Information Systems & Technologies (WebIST), Miami, FL.

Krajewski, S., & Pike, D. (2005). *Student evaluation of courses: Kicking and screaming into the 21st century.* Paper presented at the Conference on Innovations in the Scholarship of Teaching and Learning at Liberal Arts Colleges, Northfield, MN.

Laubsch, P. (2006). Online and in-person evaluations: A literature review and exploratory comparison. *Journal of Online Learning and Teaching, 2*, 62-73.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*, 221-232.

Liu, Y, (2006). A comparison study of online versus traditional student evaluation of instruction. *International Journal of Instructional Technology & Distance Learning, 3*. Retrieved April 23, 2009, from http://www.itdl.org/journal/april_06/index.htm.

Lovric, M. (2006). *Traditional and web-based course evaluations: Comparison of their response rates and efficiency.* Paper presented at 1st Balkan Summer School on Survey Methodology, Opatija, Croatia.

McGregor, A. (2007). Academic success, clinical failure: Struggling practices of a failing student. *Journal of Nursing Education, 45*, 504-511.

Silva, K. M., Silva, F. J., Quinn, J. N., Draper, J. N., Cover, K. R., & Munoff, A.A. (2008). *Teaching Psychology, 35*, 71-80.

Spooner, F., Jordan, L., Algozzine, R., & Spooner, M. (1999) Student rating of instruction in distance learning and on-campus classes. *The Journal of Educational Research, 92*, 132-

Appendix

Question	Abbreviation used in Tables	Full Question
1.	Relevance of Course Content	Demonstrated relevance of the course content
2.	Enhanced Learning	Used assignments that enhanced learning
3.	Emphasized Points	Summarized/emphasized important points
4.	Responsive to Questions	Was responsive to questions and comments from students
5.	Facilitated Learning	Established an atmosphere that facilitated learning
6.	Approachable for Assistance	Was approachable for assistance
7.	Responsive to Diversity	Was responsive to diversity of students in this class
8.	Interest in Teaching	Showed strong interest in teaching this class
9.	Intellectual Challenge	Used intellectual challenge teaching methods
10.	Fair Grading	Used fair grading methods
11.	Analysis of Ideas	Helped students analyze complex/abstract ideas
12.	Meaningful Feedback	Provided meaningful feedback about student work
13.	Overall Rating	Overall, this instructor's teaching was