

**Online Collection of  
Student Evaluations of Teaching**

**James A. Kulik**

**Office of Evaluations and Examinations**

**The University of Michigan**

**December 5, 2005**

More than a dozen universities now collect all their student ratings of teaching online, and many more schools collect some ratings online. But questions about online ratings still arise: Do enough students respond to online evaluation requests? Do students fill out online forms responsibly? Do online evaluation systems produce fair results? Before revising their evaluation systems, many school want to have better answers.

The two pioneers in the collection of online evaluations in this country were Northwestern University and Brigham Young University. Northwestern began collecting online evaluations in the spring semester of 2000, and Brigham Young University implemented online evaluations in the fall semester of 2002. Since then, Brigham Young has kept tabs on the spread of online teaching evaluations to other schools and colleges (Sorenson and Reiner, 2003). A Brigham Young website currently lists 16 universities that collect online evaluations campus-wide: Bates College, Brigham Young University, Carnegie Mellon University, Georgia Institute of Technology, Hong Kong University of Science and Technology, Laval University, Northwestern University, Polytechnic University of Brooklyn, Smith College of Social Work, Tel Aviv University, University of Idaho, University of North Texas Health Science Center, University of Virginia, Wellesley College, Whitman College, and Yale University. The website lists 43 schools that collect online evaluations in one or more departments but not campus-wide, and it also lists 25 other schools that collect online evaluations in one or more courses but not department- or campus-wide.

One of the great advantages of online systems is speed. Teachers receive the results of online evaluations immediately at the end of a course, whereas they usually have to wait three to eight weeks after a class ends for results from paper-based systems. Electronic data collection eliminates the clerical bottlenecks: the time-consuming distribution, mailing, sorting, and scanning of paper questionnaires. The elimination of such clerical jobs also makes online systems less costly to run. A careful analysis by Bothell and Henderson (2003) estimated the fiscal year 2002 cost of a paper-based system at BYU to be \$436,838 (or \$1.09 per paper student rating form) and the one-year cost of an online system to be \$186,617 (or \$0.47 per evaluation form).

Why then are so many schools still using paper evaluation forms? Inertia and the front-end development costs for online systems are certainly factors, but equally important are the validity concerns of educators. One issue is response rate in online evaluation systems. Response rates are usually high in paper-and-pencil systems, because the students who fill out paper forms are a captive audience. High response rates are usually harder to achieve in online systems, because students usually fill out online evaluations voluntarily on their own time. Many schools—including the University of Colorado, Duke University, Georgia Institute of Technology, the Air Force Academy, Kansas State University, Northwestern University, and the University of Idaho—have identified response rate as a significant problem with online ratings (Johnson, 2003).

Another concern is comparability of results from online and paper-based systems. If online and paper systems produce different results, that could be a problem for teachers who are accustomed to paper-based systems. It might be especially difficult to accept

results from online systems if they produce ratings that are systematically lower than those from paper-based systems. It would also be a problem if students write fewer or less complete comments with online evaluation systems.

E&E carried out two studies to compare the results of evaluations collected online with the results of in-class evaluations. The questions behind the studies were basic ones. Are students equally likely to complete on-line and in-class evaluations? Are evaluation results collected in the two ways equally favorable? Are the results equally reliable? Do the two methods produce the same number of student comments?

Researchers have studied these questions at other institutions, but it still seems important to ask them at the University of Michigan. The paper-based teaching evaluation system used at the University, Teaching Questionnaires (TQ), is one of the largest and most diverse teaching evaluation systems in higher education. Last year, teachers in nearly all of the University's 19 schools and colleges evaluated courses on TQ forms that were custom-tailored to their classes. In all, E&E printed a total of 431,853 TQ forms for 14,861 classes. It is important to know whether this large, diverse, and well-established system could be replaced by an online system.

## **Method**

Two studies were carried out in two different settings. The first study was conducted in introductory courses in the College of Engineering. The second was conducted in graduate courses in the economics, political science, and sociology departments of the College of Literature, Science, and the Arts.

### ***Study 1***

In the fall term of 2002, 18 instructors in the engineering program of the College of Engineering were offering two different sections of the same course. E&E randomly assigned one of each teacher's sections to an online condition and the other to an on-paper condition. In the online condition, students filled out their teaching evaluation forms online. In the on-paper condition, students filled out printed evaluation forms in class. E&E notified the students in the online classes by email that they were to fill out their teaching evaluation forms online during the last week of classes and then sent out email reminders to students as the week drew to a close. Students in the on-paper condition filled out standard paper evaluation forms, identical in content to the online forms, at class meetings during the last week of the term.

The main dependent measures in this study were response rate and rating favorability. Response rate was the percentage of students enrolled in a class who completed evaluation forms. Rating favorability was calculated from student responses to three key TQ questions. These questions were:

Question 1 – Overall, this was an excellent course.

Question 2 – Overall, this instructor was an excellent teacher.

Question 3 – I learned a good deal in this course.

Students responded to these questions on a 5-point scale, where 5 = Strongly Agree, 4 = Agree, 3 = Neutral, 2 = Disagree, and 1 = Disagree. The median response on each item indicated the rating favorability.

Additional measures used in this study were composite ratings for the teachers from semesters other than fall 2002. An examination of TQ archives showed that 16 of the 18 instructors had taught the same course at other times between fall 2000 and winter 2005. For each of the 16 teachers and each of the three questions, composite ratings were calculated. These composite ratings were considered to be highly reliable, since each of the composites was based on results from a number of different classes.

## *Study 2*

Data for the online condition came from 70 graduate-level courses offered during the winter term of 2005: 26 economics, 23 political science, and 21 sociology courses. E&E asked students in these 70 courses to fill out their teaching evaluations online. Data for the on-paper condition were archived data from 57 graduate courses offered during the winter term of 2004: 19 economics, 17 political science, and 21 sociology courses. Students in these courses filled out standard teaching evaluation forms during regular class meetings. Another source of on-paper data were a set of Xeroxed comments written on paper evaluations given out in graduate-level courses in the sociology department during fall 2004.

Four dependent measures were available for this study: response rate, rating favorability, comment rate, and comment length. The measures of response rate and rating favorability in this study were the same as the measures in Study 1. Comment rate was the percentage of enrolled students who wrote comments in response to questions on the evaluation form. Comment length was the number of words in a student's comment.

## **Results**

### *Response Rates*

In Study 1, response rates were essentially the same in online and on-paper conditions. In the 18 classes in which students submitted online evaluations, the average response rate was 74%. In the 18 matched classes in which students filled out evaluations during class meetings, average response rate was 75%. The difference in the two response rates is not statistically significant.

In Study 2, response rates were definitely lower in the online condition. In the economics departments, response rates were 80% in the on-paper condition and 52% in the online condition; in political science, response rates were 82% in the on-paper condition and 72% in the online condition; and in sociology, the rates were 80% in the on-paper condition and 65% in the online condition. Average response rate was 65% in the online condition and 80% in the on-paper condition. The overall difference is highly significant statistically, and the difference in each department is also statistically significant.

### ***Favorability of ratings***

In Study 1, ratings were higher in the on-paper than in the online condition (Table 1). For Question 1 (Excellent course) on-paper ratings were 0.10 points higher on the 5-point scale; for Question 2 (Excellent teacher) on-paper ratings were 0.17 points higher; and for Question 3 (Amount learned) on-paper ratings were 0.20 points higher. The difference on Question 1 was not statistically significant, but the differences on Questions 2 and 3 approached statistical significance.

Results were similar in Study 2 (Table 2). Average ratings were again higher in the on-paper condition than in the online condition. For Question 1, on-paper ratings were 0.17 points higher; for Question 2, on-paper ratings were 0.22 points higher; and for Question 3, on-paper ratings were 0.20 points higher. The difference on Question 1 was not statistically significant; the difference on Question 2 approached statistical significance; the difference on Question 3 was statistically significant.

Overall, online ratings were consistently less favorable than on-paper ratings. But the size of the difference in ratings in the two conditions was small.

### ***Reliability of ratings***

Correlations with composite ratings of online and on-paper ratings provide an indication of the reliability of the online and on-paper ratings. The composite ratings are the average ratings of Study 1 teachers in all semesters but fall 2002 between fall 2000 and winter 2005. Results show that the correlation between online ratings and composite ratings is as high as the correlation between on-paper and composite ratings (Table 3). For example, for Question 1 (Excellent course), the correlation with the composite rating was .74 for online ratings and .63 for on-paper ratings.

### ***Student comments***

The likelihood of students writing comments on their rating forms was the same for online and on-paper conditions. The evidence on this question comes from a comparison of comments made by students in graduate courses in the sociology department in fall 2004 (when evaluations were collected on paper) and in winter 2005 (when evaluations were collected online). Results show that 61% of the students in the online condition and 63% of students in the on-paper condition wrote comments. The difference is small, and it is not statistically significant.

The length of comments was also very similar for the two conditions. The average length of a written comment in the online condition was 52 words; the average length in the on-paper condition was 54 words. Again, the difference between conditions was small, and it was not statistically significant.

## **Related Work**

Researchers began studying the effectiveness of online evaluation systems while the systems were still undergoing development. This section of the paper reviews findings from pioneering systems at Northwestern University and Brigham Young University.

### ***Response rates***

Hardy (2003) calculated response rates for six classes at Northwestern University. The six classes were in a single department; they covered similar content; and each of the classes enrolled about 40 students. In four of the six classes, students evaluated teaching online; in the remaining two classes, students filled out paper evaluations. Response rates in the four classes using online evaluations were between 64% and 73%. Response rates in the two classes using paper forms were 100% and 83%.

Johnson (2003) reported on in-class and online response rates during development of the online system at Brigham Young University. The data came from 74 courses in which some students filled out online evaluations and others filled out paper forms in the 1999 academic year. For these 74 sections, the response rate was 50% for online ratings and 71% for paper ratings.

In Fall 2000, these researchers asked BYU teachers what they communicated to students about completing online forms. Seventeen (50%) of the 34 faculty members in the study responded to the inquiry. Their responses were categorized into four categories, which represented the degree to which the teachers used incentives and other forms of encouragement to get students to cooperate. Results showed that response rates were much higher when teachers made teaching evaluations a course assignment (whether or not the teachers gave points for completion of the assignment). Response rates were lower when teachers simply encouraged students to complete evaluations or when teachers did not mention teaching evaluations in their courses.

Overall, the results at Northwestern and Brigham Young are similar to the results at Michigan. Students at each of the schools were somewhat less likely to fill out online evaluation forms. The difference in online and in-class response rates at Michigan was about 10%. The difference between the two conditions appeared to be about 20% at both Northwestern and Brigham Young.

### ***Written comments***

Hardy (2003) also examined comments made by students on online and paper evaluation forms. The six classes in Hardy's study were in a single department; they covered similar content; and each class enrolled about 40 students. In four of the classes, students filled out online evaluations; and in the remaining two classes, they filled out paper evaluation forms.

Hardy found that students made more and longer comments on online forms. The percentage of students making comments on paper evaluations was 16% in one sections and 12% in the other section. The percentage making comments on online forms was between 37% and 69% in the four classes. In total, students in the online sections produced five times as much written commentary as students in the in-class sections did. Finally, the number of positive, negative, and mixed comments were similar in online and in-class conditions.

These findings are strikingly different from our results at Michigan. At Northwestern, more students wrote comments online and their comments were longer than comments written on paper evaluation forms. At Michigan, comments were similar with online and in-class formats.

### ***Favorability of ratings***

Hardy (2003) reported on favorability of ratings from online and in-class ratings collected from fall quarter 1999 through fall quarter 2000 at Northwestern University. Of the 5112 classes evaluated during the four quarters, 2457 classes used paper forms and 2655 classes used online forms. The online scores were 0.25 of a point lower than in-class scores on a six-point scale from 1 (low) to 6 (high).

Hardy also found similar results in a further analysis that was restricted to pairs of matched classes (same instructor and same course for each pair) that were evaluated through online and in-class systems. The results in the second analysis were precisely the same as those of the first analysis: online ratings were 0.25 points lower than in-class ratings.

These results are similar to results in both studies at Michigan. The difference in favorability of online and in-class ratings in the Michigan studies was about 0.15 points on a 5-point scale. At Northwestern the difference in favorability was about 0.25 points on a 6-point scale.

## Conclusion

Online rating systems produce results that are similar to results from in-class rating systems. However, the results of the two methods of data collection are not completely identical.

First, collecting data online can affect response rates. In Study 1, collection method did not affect response rate, but in Study 2, it did. The target population in the two studies was different, and this may account for the different results. Study 1 was carried out in introductory undergraduate courses; Study 2 was carried out in graduate-level courses. In addition, each student in Study 1 was asked to fill out only one course evaluation form online, whereas the graduate students in Study 2, who were enrolled in several graduate-level courses, were asked to fill out as many as six, seven, eight, or nine evaluation forms online.

Whatever produced the difference in results, the important point to note is that response rate was adequate, if not ideal, with online data collection. The overall response rate with online data collection was 74% in Study 1 and 65% in Study 2. In addition, it may be possible to increase the response rate with careful use of directions and reminders.

Second, ratings collected online are slightly less favorable than ratings collected in classes. It is unclear why rating favorability drops with online evaluation, and it is also unclear which ratings—online or paper ratings—give a fairer picture of a teacher and a course. An important point to note, however, is that the difference in ratings made under the two conditions is not large. Online ratings seem to average between 0.1 and 0.2 points lower than in-class ratings on a 5-point scale. In rating systems that include normative data on reports, it may be possible to ignore this small difference.

**Table 1**  
**Student ratings on core evaluation items, made under two conditions in Study 1**

	Online classes			On-paper classes			<i>t-ratio</i>	<i>Sig.</i>
	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>		
Q1. Excellent course	3.85	0.31	18	3.95	0.30	18	1.54	0.14
Q2. Excellent teacher	4.08	0.48	18	4.25	0.45	18	1.99	0.06
Q3. Amount learned	3.67	0.52	18	3.87	0.40	18	1.79	0.09

**Table 2**  
**Student ratings on core evaluation items, made under two conditions in Study 2**

	Online classes			On-paper classes			<i>F-ratio</i>	<i>Sig.</i>
	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>		
Q1. Excellent course	4.28	0.59	70	4.45	0.48	57	2.47	0.12
Q2. Excellent teacher	4.31	0.66	70	4.53	0.49	57	3.51	0.06
Q3. Amount learned	4.36	0.55	70	4.56	0.44	57	4.50	0.04

**Table 3**  
**Composite on online and in-class ratings with composite teaching ratings**  
**For 16 Study 1 teachers**

	Correlation	
	Online vs composite	In-class vs composite
Q1. Excellent course	.74	.63
Q2. Excellent teacher	.66	.62
Q3. Amount learned	.72	.66

## References

- Bothell, T. W., and Henderson, T. (2003) "Do Online Ratings of Instruction Make Sense?" In D. L. Sorenson and T. D. Johnson (Eds.), *New Directions for Teaching and Learning*, no. 96. San Francisco: Jossey-Bass. Pp. 69-80.
- Johnson, T. D. (2003) "Online Student Ratings: Will Students Respond?" In D. L. Sorenson and T. D. Johnson (Eds.), *New Directions for Teaching and Learning*, no. 96. San Francisco: Jossey-Bass. Pp. 49-60.
- Sorenson, D. L., and Reiner, C. (2003) "Charting the Uncharted Seas of Online Student Ratings of Instruction." In D. L. Sorenson and T. D. Johnson (Eds.), *New Directions for Teaching and Learning*, no. 96. San Francisco: Jossey-Bass. Pp. 1-24.