A Comparison of Three Modes of Student Ratings of Teacher Performance

Larry S. Rice and Eric V. Van Duzer

Humboldt State University

Abstract

One of the primary means of evaluating teaching effectiveness in higher education is through the use of end-of-term student ratings of teacher performance. A number of universities have begun experimenting with replacing the traditional paper and pencil rating forms with on-line forms. Because student ratings often play a key role in personnel decisions, a study was conducted to evaluate the comparability of two methods of on-line rating with traditional paper-and-pencil ratings. A total of 53 students enrolled in one section of a course taught by a single instructor were randomly assigned to one of three conditions. One group completed traditional paper forms. The second group completed on-line forms as a class in a computer lab. The third group completed on-line forms independently. The primary difference in outcomes suggested by the data was not between paper-and-pencil and online groups, but rather between the groups that completed the ratings in whole-class *vs*. independent contexts.

A Comparison of Three Modes of Student Ratings of Teacher Performance

Larry S. Rice and Eric V. Van Duzer

Humboldt State University

End of term student ratings of teacher performance are widely used in evaluating the teaching effectiveness of college faculty participating in the retention, tenure, and promotion process. Over the past several years, a number of universities have begun experimenting with using on-line rating forms to replace traditional paper-and-pencil forms. On-line administration reduces the costs of collecting, transcribing, and reporting data, which makes the proposition attractive to institutions of higher education, especially in challenging budgetary times.

The comparability of student ratings data between traditional paper and pencil forms and on-line ratings has not been established. Therefore, it is possible that the use of on-line forms for data collection will lead to retention, tenure, and promotion decisions being made on the basis of an incorrect assumption that the two forms produce comparable outcomes. Two previous studies (Rice & Van Duzer, 2002 & 2003) produced conflicting results. Those studies were limited by the minimal variation in the student ratings. The present study, involving 53 students from a single course, was conducted with a general education course that enrolled students ranging from freshman to seniors. Research suggests that this course should experience greater variation in student ratings (Civian and Brennan, 1996) and was therefore selected to overcome past limitations in examining the question of comparability between modes of student's completing course ratings.

Student Ratings of Teacher Performance

Student ratings of teacher performance (SRTP) are among the most widely used measures of teaching effectiveness (Centra, 1993). Summaries of these ratings are often highly consequential in that they provide the basis for teaching improvement and personnel decisions. Because of the clear relationship between quality teaching and optimal student achievement (Darling-Hammond, 1997), it is desirable that faculty teaching effectiveness be carefully monitored. Procedures for measuring faculty teaching effectiveness vary by university; however, student ratings of teacher performance are typically considered in the process, and they are generally important elements of tenure and promotion decisions (Haskell, 1997; Marsh, 1987).

Faculty concerns regarding the use of student-completed rating forms as assessments of teaching quality have been well documented (Cashin, 1995; Haskell 1997; Mark 1982; Marsh 1987). Instructors might agree on the need for student input into the evaluation of their teaching, however, many express concerns about reliability and validity of these measures. Mixed results from tests of reliability and validity of student ratings of teacher performance underscore that concern (Haskell, 1997; Simmons, 1996). According to

Feldman's (1989) review of research, student ratings of teacher performance have been shown to relate only moderately to other indices of teaching effectiveness correlating, on average, 0.29 with instructor's ratings, 0.39 with administrator ratings, 0.55 with colleague ratings, and 0.40 with student achievement. However, even though faculty may mistrust student evaluations, most agree that they are useful, and that they are here to stay (Greenwald, 1997; Callahan, 1992).

Studies suggest that a number of factors other than the quality of teaching may be related to SRTP. For example, course grades have been correlated with student ratings of teacher performance, and higher grades were related to higher ratings (Greenwald & Gillmore, 1997a; Tang, 1999; Stumpf and Freedman, 1979). In other studies, SRTP was correlated with student workloads, (Greenwald and Gillmore, 1997), course rigor (Overbaugh, 1998), and congruence between student and instructor cognitive styles (Carroll, 1995). Interestingly, use of active teaching techniques such as group assignments and discussion sections, which students might be expected to find more interesting or entertaining than a straight lecture format, was not related to more positive ratings of the teacher's performance (Leeds, Stull, and Westbrook, 1998). Research on class size, while mixed, tends to support a weak relationship between higher SRTP and smaller class size (Fernandez, Mateo, and Muniz, 1998; Marsh and Roche, 1997).

In one of the most comprehensive efforts to date, Civian and Brennan (1996) used hierarchical linear modeling to investigate predictors of Harvard University student ratings of teacher performance. Examining results across more than 1000 undergraduate courses, their data suggest a number of robust predictors of positive student ratings, including: course subjects; higher level of difficulty; high proportion of majors; course taught by Assistant or Associate Professor; and being a freshman in a course with few freshmen. Predictors of less positive student ratings included: math/science courses; high proportion of students taking the course as a requirement; and being in a math/science course and finding it more difficult than other students.

It is clear from the literature that a variety of variables other than teacher effectiveness are related to student ratings of teacher performance. The use of on-line forms of student course evaluations introduces two additional variables: technology mediated communication, and a situational variable where students complete the evaluations independently at home at their convenience rather than as part of a group of students completing the forms at the same time and place. Research in a variety of settings suggests the use of technology may affect both the content and quantity of communication (), and that situational variables can influence various types of scores (). Given that previous research indicates that a range of non-teaching related variables affect course ratings, the authors posed the following questions as the basis of this study.

> Do on-line ratings differ significantly from paper and pencil ratings in either the overall numerical scores or in the length of comments provided by the students?

Do scores completed independently at various times differ significantly from those completed in a whole-class setting in either the overall numerical scores or in the length of comments provided by the students?

Given the seriousness of RTP decisions based in part on student ratings, it is incumbent upon institutions of higher education to assure comparability of scores obtained using paper-and-pencil and on-line data collection where traditional standards of interpretation are employed.

Method

The subjects for this study consisted of 60 students, drawn from a Sociology course, taught in the fall semester of 2004. The students who attended the final course meeting were randomly divided into three groups using a random number table. Of the 60 students, a total of 53 completed the evaluation form. The three conditions consisted of a web-based-at-home rating form group (n=16), a web-based-computer-laboratory rating form group (n= 22), and a group which completed the standard paper–and-pencil course rating form in class (n= 15).

At the conclusion of the class session, color-coded instruction sheets were distributed to the students. The paper & pencil group remained in the classroom where instruction had just been completed. The at-home group was given instructions and dismissed. The laboratory group was led *en masse* to a computer laboratory in the same building, which contained a sufficient number of computers to allow all members of the group to complete the form simultaneously. None of the groups were aware of the other groups' conditions.

All three groups were read a standard script that explained the purpose and procedures for completing course evaluations. Both the in-laboratory and at-home computer groups were given additional instructions on accessing the web-based evaluation forms. All groups were told that they must either complete the evaluation or indicate on the paper or web-based form that they were deliberately choosing not to participate. Students were informed of the requirement to complete the evaluation (or choose not to participate) in order to avoid a delay in posting their final grade for the course.

The on-line process for these two experimental groups used a web-based format that the students were familiar with from registration and other campus business. The on-line evaluation form presented the seventeen course evaluation questions in the same order with the same response format and as the paper-and-pencil forms. However, the on-line forms presented the questions one at a time, with students using a mouse click to move to the next question. Written comments were solicited in the same manner on both the web and paper instruments, although there was no limit on the space for comments on the computer form. The paper form had practical limits, but none of the comments in any format were, in fact, long enough to exceed the available space on the paper form.

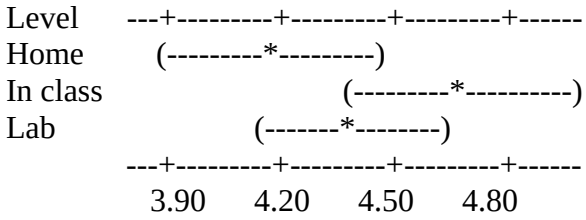
Data were collected, descriptive statistics calculated, and the One-way ANOVA procedure was used to calculate the probability that the groups' mean ratings differed significantly based on the response mode**.** Two variables were analyzed, mean course ratings, which were calculated by averaging the ratings for all seventeen questions on the evaluation form, and length of written comments, calculated by totaling the number of letters in the comment section of the form.

Results

Students were informed that posting of their final grades might be delayed until they completed the course evaluations or indicate their choice not to participate by marking that option on the web-based forms. Despite this, a total of seven at-home students failed to complete the evaluations. As a result, 100% of the classroom group, 100% of the lab group, and 69% of the at-home group completed the forms or indicated that they chose not to participate. Visual inspection of the data suggested that the three conditions differed. A One-way analysis of variance showed a difference that was suggestive but non-significant (*F = 2.73, p = 0.075*) between course ratings for the on-line at home (*M = 4.10, SD = 0.758, n = 16*), on-line in lab (*M = 4.45, SD = 0.574, n = 22*), and the paper-and pencil-group (*M = 4.60, SD = 0.486, n = 15*).

The mean scores for questions on the form for all groups were quite high.

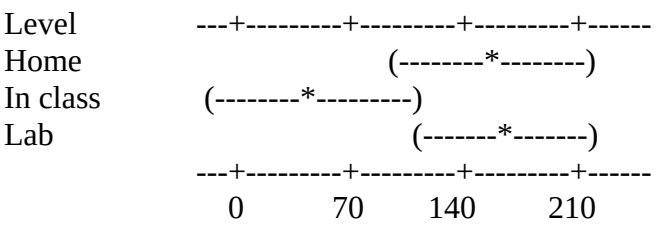
Chart 1. Individual 95% Confidence Intervals for Mean Based on Pooled StDev



There was also no significant difference when the two on-line groups were combined and compared using a t-test to the paper & pencil group who completed traditional paper-and-pencil ratings in class. The most pronounced difference was found when the two whole-class groups, (the on-line group that was escorted *en masse* to the computer lab, and the control group) were combined into a single "whole-class" group, and compared to those that completed on-line evaluations independently. A t-test between the whole-class condition and the independent on-line condition was again suggestive, but also failed to reach the level of statistical significance (*T = -1.96, p = .064*).

In addressing the question as to whether the length of comments varied by the three conditions a One-way ANOVA was conducted. The ANOVA results showed a significant difference (*F = 4.75, p = 0.014*) between the three conditions. For on-line at home (*M = 155, SD = 180.9, n = 13*), on-line in lab (*M = 167.8, SD = 94.8, n = 19*), and the paper-and pencil-group (*M = 44.8, SD = 38.8, n = 13*). Paper and pencil comments were significantly shorter than either of the two on-line groups.

Chart 2. Individual 95% Confidence Intervals for Mean Based on Pooled StDev



In sum, results of the present study indicated that there were non-significant, but highly suggestive, differences in mean course ratings but significant differences in length of comments between the on-line and paper-and-pencil groups.

Discussion

Results in the present study indicate that on-line ratings may be comparable to traditional paper and pencil ratings particularly when both are completed in similar settings. It does appear, however, that ratings may differ when students complete the on-line forms independently at home as opposed to whole-group conditions. This means that the comparability of on-line ratings and traditional ratings may, in fact, be dependent on setting. However, the low response rate (60%) of the at-home group makes definitive conclusions difficult at this time. This result, despite efforts to ensure full participation, reflects a commonly reported problem with on-line ratings (e.g, Johnson, 2002). The current study provides fresh evidence that the differences between on-line and paper & pencil ratings are not related to the use of technology and its effect on scores. Rather, the evidence suggests that the situational variables, and possibly the issue of who responds and who doesn't, are more likely to affect comparability of scores.

Results of the present study seem to suggest that it may in fact be feasible to use cost effective, on-line procedures without significantly affecting interpretations of ratings for faculty. The distinction that the on-line ratings should be completed in a group setting is important, however. The present results suggest that simply putting the ratings forms on-line and allowing students to complete them at their leisure, may not yield comparable results.

Where the use of technology does result in significant differences is in the length of written comments. Both on-line conditions resulted in significantly longer comments than when paper & pencil were used to complete the ratings. In the current study the content of the comments was not analyzed. Future research should investigate whether the net positive or negative content of longer comments reflects similar patterns in paper and pencil ratings to ensure comparability for interpretation.

References

Callahan, J. P. (1992). Faculty attitude towards student evaluations. *College Student Journal, 26,* 98-109.

Cashin, W.E. (1995). *IDEA technical report no. 32: Student ratings of teaching: The research revisited.* Manhattan, KS: Center for Faculty Evaluation and Development, Kansas State University.

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

Civian, J. T. and Brennan, R. T. (1996) *Student and course factors predicting satisfaction in undergraduate courses at Harvard University.* Paper Presented at the American Education Research Association Conference (New York)

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching.* New York: National Commission on Teaching and America's Future.

Feldman, K.A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external observers. *Research In Higher Education, 30,* 137-194.

Fernandez, J. J., Mateo, J., and Muniz, M. A. (1998). Is there a relationship between class size and student ratings of teacher quality? *Educational and Psychological Measurement, 58,* 596-604.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182-86.

Greenwald, A. G., and Gillmore. G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209-17.

Greenwald, A. G., and Gillmore. G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743-51.

Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives 5(6).*

Johnson, T. (2002). *Online Student Ratings: Will Students Respond?* (ERIC Document Reproduction Service No ED465794).

Leeds, M., J. Stull, and W. A. Westbrook. (1998). Do changes in classroom techniques matter? Teaching strategies and their effects on teaching evaluations. *Journal of Education for Business, 74*, 75-8.

Mark, S. F. (1982). Faculty evaluation in community college. *Community Junior College Research Quarterly, 6(2),* 167-78.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. [Monograph summary]. *International Journal of Educational Research, 11*, 253-387.

Marsh, H. W., and L. A. Roche. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187-97.

Overbaugh, R. C. (1998). The effect of course rigor on preservice teachers' course and instructor evaluation. *Computers in the Schools, 14*, 13-23.

Rice, L. S. & Van Duzer, E. (2004) Student evaluation of teaching. Paper presented at the American Educational Research Association Conference (San Diego).

Rice, L. S. & Van Duzer, E. (2003) Online instructor evaluation: report on research in progress. Paper presented at the Hawaii International Conference on Education (Waikiki).

Simmons, T. L. (1996). Student evaluation of teachers: Professional practice or punitive policy? *JALT Testing & Evaluation N-SIG Newsletter, 1(1),* 12-6.

Stumpf, S. A., and R. D. Freedman. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71*, 293-302.

Tang, S. (1999). Student evaluation of teachers: Effects of grading at college level. *Journal of Research and Development in Education, 32*, 83-8.