Interpretation Guide

For

Student Opinion of Teaching Effectiveness (SOTE) Results

Prepared by:

Office of Institutional Effectiveness and Analytics

and

Student Evaluation Review Board

May 2011

Table of Contents

Introduction

The Student Evaluation Review Board is an Operating Committee of the Academic Senate comprised of faculty representing each campus college. It is responsible for overseeing student evaluations of teaching, including developing and revising the SOTES (Student Opinion of Teaching Effectiveness) and SOLATE (Student Opinion of Laboratory and Activity Teaching Effectiveness) forms, and authoring and updating the interpretation guides for SOTE and SOLATE (see Senate Resolution F02-2).

In addition, SERB is charged with updating the department, college, and university norms (or averages) that are contained on the SOTE reporting forms (Senate Policy Recommendation S08-6). The norms compare an instructor's ratings with the average ratings of colleagues and, therefore, make it possible to form a better judgment about an instructor's teaching effectiveness. New SOTE norms were calculated in Fall 2003 and again in Fall 2008. In the latter semester faculty members were urged to evaluate all their sections so that the resulting norms would not be biased by a small or unrepresentative sample. As a consequence, 3,639 sections were evaluated, comprising 91% of all sections. The new norms were calculated based on the 76,086 SOTE forms completed by students.

Because of the calculation of new SOTE norms, SERB is issuing this updated interpretation guide. The information presented here provides a description of the SOTE form, explanations for the statistics included in the SOTE report, and factors that influence SOTE ratings.

The SOTE Form

Following several years of development by SERB, the current SOTE rating form was adopted for implementation beginning in the Fall 2003 semester. The rating form contains four numbered pages. Page 1 contains thirteen standardized rating items that assess students' perceptions on teaching effectiveness and the learning experience. The first 12 items are answerable with a five-point Likert scale:

- (5) Very Strongly Agree
- (4) Strongly Agree
- (3) Agree
- (2) Disagree
- (1) Strongly Disagree

There also is a sixth option, (NA), Not Applicable/No Opportunity to Observe. These items address different aspects of teaching, as summarized in the following table.

Table 1
SOTE Rating Questions and Correlations

Aspects	Questions	Correlation With Q13

Relevance	1. Demonstrated relevance of the course content	0.74
Learning Environment	2. Used assignments that enhanced learning	0.75
	4. Was responsive to questions and comments from students	0.73
	5. Established an atmosphere that facilitated learning	0.78
	8. Showed strong interest in teaching this class	0.71
Helping Students Think	3. Summarized/emphasized important points	0.77
	9. Used intellectually challenging teaching methods	0.73
	11. Helped students analyze complex/abstract ideas	0.78
Responsiveness to Students	6. Was approachable for assistance	0.69
	7. Was responsive to the diversity of students in class	0.69
Grading/ Feedback	10. Used fair grading methods	0.71
	12. Provided meaningful feedback about student work	0.74
Overall Effectiveness	13. Overall, this instructors teaching was	

The last item is a summary measure of teaching effectiveness and is also answerable with a five-point Likert Scale:

- (5) Very Effective
- (4) Effective
- (3) Somewhat Effective
- (2) Ineffective
- (1) Very Ineffective

Question 13 is strongly correlated with all of the other items and therefore is a good index of overall effectiveness.[1] **Nonetheless, evaluations of teaching effectiveness should be based on all 13 items and not solely on ratings for item 13.**

Each of the 13 items is presented in a separate box of its own rather than in a matrix of questions. This layout was designed to maximize the likelihood that each item would be read and considered on its own, and to reduce the likelihood that students would simply endorse the same rating for each item by marking the same number in a matrix.

Page 2 asks students about their expected grade in class and their class level. It also asks whether or not their ratings were unduly influenced by other students or the

[1] The Pearson product moment correlation measures the strength of linear dependence between two variables, and varies between -1 and 1. A value of -1 means two variables are perfectly inversely related. A value of 0 implies that there is no linear relationship, while a value of +1 means variables increase or decrease in perfect lock-step. As a rule of thumb, correlations between .00 and .29, even when statistically significant, are not practically useful. Correlations between .30 and .49 are practically useful. Correlations between .50 and .70 are very useful but are not common when studying complex phenomena. The correlations presented in the Table 1 were calculated from the Fall 2008 SOTE data and are all highly statistically significant.

instructor. Pages 3-4 allow students to provide written evaluations of the instructor's strengths, weaknesses, and other helpful comments. The written comments are returned to the instructor only after course grades have been released.

Interpretation of the SOTE Ratings

The official SOTE reporting forms consist of two pages. The first page provides the instructor's means, standard deviations, and medians for the 13 rating items. To aid in interpretation, it also provides the norm data (means, standard deviations, and medians) for the instructor's college, and the university as a whole.

- The **mean** is the arithmetic average of student responses. Means are reported to the first decimal place.

- The **standard deviation** is a measure of agreement among respondents. It indicates the variability among the responses. That is, how much, on the average, student responses vary from the mean. Standard deviations for most items are very close to 1.0. A large standard deviation (greater than approximately 1.3) indicates that students frequently do not agree about what rating should be assigned (i.e. students use three or more descriptors for a single item). A small standard deviation (less than approximately .7) indicates that students generally agree about what rating should be assigned (i.e. students usually use only two adjacent descriptors for a given item). We do not expect to often see 100% agreement among students – an excellent teacher for one student may be only average for another student given differential preparation or experiences of the two students.

- **Means and standard deviations should be interpreted with caution** when 10 or fewer students complete the ratings. Both statistics are highly influenced by even one or two aberrant scores if the number of ratings is fewer than about 10. Thus classes and/or items where fewer than 10 students have responded have been flagged with an asterisk and the following sentences will be printed directly below the rating items -***ITEM STATISTICS ARE BASED ON 10 OR FEWER STUDENTS. RESULTS SHOULD BE INTERPRETED WITH CAUTION***. Great caution should be used when interpreting means and standard deviations of such classes and/or items because the statistics may be unstable – check for consistency across classes and across rating occasions. In addition, when more than 30% of the students in a class leave an item blank or mark it "not applicable," that rating probably should not be interpreted.

- The **median** is the middle ranking. A median of 3.5 indicates that half the students gave ratings higher and half lower than 3.5. The median is helpful in cases where outliers might influence the mean and standard deviation; e.g. cases in which a few extremely high or extremely low ratings push the mean score in a direction that is not representative of the class as a whole. This is particularly

likely in smaller classes or classes with large numbers of blanks or "not applicable" ratings.

- **Norms:** As mentioned in the Introduction, data for new norms were gathered in the Fall 2008 administration of SOTEs. For departments, colleges, and the university as a whole, SOTE responses were aggregated to compute the means, medians, and standard deviations that serve as referent points for making comparisons. Without norms it is difficult to interpret an instructor's scores. Are the scores below, at, or above the scores of other instructors? Norms (university, college, and department) compare an instructor's ratings with the average ratings of colleagues and, therefore, make it possible to form a better judgment about an instructor's teaching effectiveness.

- **Comparisons between the class data and norm data** are best made using the graphic display on the second page of the report. Norm data for the college and university levels only are graphically displayed on page 2 of the printout. For each item the middle 60% of ratings received by instructors was determined for each college and the university as a whole. This range is displayed as a line of dashes. This line represents the usual range of ratings received by instructors for that item. The class mean is printed as an asterisk on the same line. Only if the class mean falls below the university or college norm (represented by an asterisk to the left of the dashes) or above the university or college norm (represented by an asterisk to the right of the dashes), can SOTE data can be used to identify exceptional teachers (those with rating means outside the norm average.) <u>The usefulness and validity of the ratings will be degraded if ratings within the norm area are interpreted as anything other than typical.</u> It should be noted that students tend to "agree" with the statements on the SOTE (giving scores of 3, 4, and 5) indicating a highly favorable evaluation of the typical SJSU instructor. SOTE interpretation should be done using trends across classes and semesters. If one item mean is <u>consistently</u> below (or above) the norm then the item should be noted as important. If an item mean is <u>inconsistently</u> above or below the norm, RTP committee members should request further information from the faculty member about the classes. It is especially important to note consistencies or inconsistencies in the same course preparation on different occasions. Thus it is possible to note steady improvement or decline.

Page 1 of the SOTE report also displays the frequencies of responses for the thirteen rating items, the percent of students who expect to receive As, Bs, Cs, etc., the percent of students by class level, and the average final GPA. These data also may assist with interpretation. As discussed below, student evaluations of teaching effectiveness may be affected by expected or actual grades and class level.

Finally, students' written comments provide additional information on teaching effectiveness. Subjective ratings of "officially" rated classes must be included in the dossier. **In interpreting these responses, members of RTP committees should take into account the majority of comments, rather than focusing on atypical responses.**

However, if comments are repeatedly observed for the same instructor across sections and time, then the RTP committees should consider further evaluations for that instructor.

Factors Affecting SOTE Ratings

Overview of Reliability and Validity

Student evaluations of teaching may be the most studied issue in higher education. Cashin's (1988) review of the literature studying the reliability and validity of evaluations reported that there were over 1,300 articles and books dealing with these two subjects. His updated review a few years later reported there were "now more than 1,500 references dealing with research on student evaluations of teaching" (Cashin, 1995). In the educational literature, reliability refers most often to consistency or interrater agreement between student ratings within a given class. Validity addresses the basic question: does the test measure what it is supposed to measure? For student ratings this translates into the extent to which student rating items measure some aspect of teaching effectiveness.

Researchers agree that reliability of students' ratings is generally good (D'Appollonia & Abrami, 1997; Centra, 1993; Kulik, 2001; Marsh, 1984). Marsh (1984, p. 717) concluded, "Given a sufficient number of students, the reliability of class-average student ratings compares favorably with the best objective tests." The ratings also are fairly stable. Studies have shown considerable agreement between retrospective ratings made by former students and those of currently enrolled students.

Although there is no agreed upon definition of "effective teaching" (Cashin, 1995; Kulik, 2001), researchers also conclude that student ratings are generally valid. In theory, effective teaching should be connected to greater student learning. The best evidence for this connection comes from student ratings in multi-section college courses. Instructors follow a common syllabus, use the same readings, and administer the same final examination. Correlations between average examination scores and average student ratings are usually positive. Researchers "have concluded therefore that students generally give high ratings to teachers from whom they learn the most, and they generally give low ratings to teachers from whom they learn the least" (Kulik, 2001, p. 12). Content analyses of students' written comments on evaluation forms also find strong positive correlations between the numerical ratings and the comments, indicating the numerical ratings and comments give nearly identical pictures of teaching effectiveness (Braskamp, Ory, and Pieper, 1981).

Despite the general acceptance of teaching evaluations as reliable and valid, researchers note that the ratings can be affected by a number of factors. Several factors were found to systematically influence SOTE ratings in the Fall 2008 data. Each is described below and references to similar findings from research on faculty evaluation conducted elsewhere are provided. These factors should be considered in any RTP

evaluation of SOTE data. It is the responsibility of the faculty member to assure that information about any of these factors is included in the dossiers along with the ratings.

Expected and Actual Grades

It is well established that students' evaluative ratings of instruction correlate positively with both expected and actual course grades (Stumpf & Freedman, 1979; Greenwald & Gillmore, 1997). Most researchers typically find a correlation of about .2 between grades and ratings and conclude that the possible effects of grades on ratings are small (Kulik, 2001). Greenwald & Gillmore (1997), however, concluded from their analyses that grading leniency exerts an important influence on ratings. The links between grades and ratings, however, do not necessarily invalidate ratings:

> The central principle of the teaching-effectiveness theory is that strong instructors teach courses in which students both (a) learn much (therefore, they earn and deserve high grades) and (b) give appropriately high ratings to the course and to the instructor. Thus, instructional quality is a third variable that explains the grades-ratings correlation in a way that raises no concern about grades having improper influences on ratings. (Greenwald & Gillmore, p. 1210)

As noted above, students are asked to report their expected grade at the time of the SOTE administration. Correlations between expected grades and ratings based on the Fall 2008 data correspond to those found in the literature, .26 (p=.000) for the summary evaluation of teaching effectiveness, and between .15 (p=.000) and .29 (p=.000) for the other 12 items. Although these positive correlations are statistically significant, they are very modest and perhaps not practically useful.

Nevertheless, frequencies for each possible grade are noted on the SOTE report, as is the actual average final GPA grade for the class. In general, expected grades should be distributed across the range of possible grades. When interpreting SOTE ratings RTP committees should note the distribution of expected grades. Classes in which the majority of students expect either low or high grades should be fairly rare (exceptions to this would be graduate and credential classes in which a grade lower than a "B" is often considered equivalent to a failing grade, and some classes in the Colleges of Science and Engineering in which grades are often lower than in other subjects). The expected average grades for a class should show some relationship to expected grades. In cases where there is a wide discrepancy (e.g. 80% of the class expects a grade of "A" while the actual average grade for the class is a 2.3) RTP committees should request further information from the instructor.

Class Size

Researchers find a relationship between class size and ratings, with small or moderate sized classes (<20) classes tending to produce higher ratings than larger (>20) classes (Mateo and Fernandez, 1996; Fernandez, Mateo, & Muniz, 1998). But the differences in ratings are usually found to be quite small. In addition, some researchers find curvilinear relationships where large classes also are rated favorably.

In the Fall 2008 data, the average ratings for overall effectiveness varied by class size: 1-10 students, 4.52; 11-30, 4.35; 31-50, 4.26; and 51 and above, 4.25. These differences in average ratings are statistically significant. But the correlations between class size and overall teaching effectiveness in the Fall 2008 data are weak, -.199 (p=.000) for total enrollments, and -.082 (p=.000) for the actual number of ratings.[2] Those interpreting SOTEs should consider average class sizes at the department, college and university levels when comparing a candidate's scores to the norms, as class size may influence SOTE scores.

Student Level

Faculty evaluation ratings can be influenced by student level. Ratings in graduate and credential classes tend to be higher than in undergraduate classes (Arreola, 2000; Marsh & Hocevar, 1991). However, the findings are weak and inconsistent regarding lower and upper division courses (Arreola, 2000; Aleamoni and Thomas, 1980; Stewart and Malpass, 1966). In the Fall 2008 data, average overall effectiveness ratings increase with level, 4.30 for lower division courses, 4.33 for upper division courses, and 4.36 for graduate courses. However, these differences are not statistically significant. And the correlation between level and average overall ratings is a very weak .038 and not statistically significant.

Course Choice

Students who take a class because of either an interest in the subject matter or because of the instructor's reputation tend to rate their instructors more favorably than students who take a course because it is required. Ratings given by students who are required to take a class are often lower than ratings by students for whom the class is an elective (Arreola, 2000). However, there is little support for these general findings in the Fall 2008 ratings. The average overall effectiveness rating for remedial courses was 4.38, 4.34 for GE courses, and 4.32 for other courses, presumably courses in the major and elective courses. But these differences in average ratings are not statistically significant.

[2] In Fall 2008, on average 64.3% of students in each class completed faculty evaluations.

College Level Comparisons

There are differences in the average ratings of overall teaching effectiveness between colleges in the Fall 2008 data:

- Applied Arts & Sciences, 4.38
- Business, 4.21
- Education, 4.38
- Engineering, 4.14
- Humanities and Arts, 4.38
- Social Sciences, 4.33
- Sciences, 4.21

These differences in average ratings are statistically significant. Not surprisingly, there are also differences in average ratings between departments within colleges as well. In light of this, it is important that RTP committees evaluating candidates from different departments and colleges (University level RTP) compare instructors to colleagues within their own departments and colleges (Arreola, 2000).

Online vs. Paper Administration

Several studies have found no significant difference in the total quantitative evaluation scores between online evaluations and paper evaluations (Donovan et al., 2006; Hardy, 2003; Heath, Lawyer, and Rasmussen, 2007; Laubsch, 2006; Spooner, Jordan, Algozzine, and Spooner, 1999). At SJSU, a study by Sujitparapitaya and Briggs (2010) indicated that there was no significant difference for a majority of the responses between online evaluations and paper evaluations (the overall the response rate for paper evaluations was 73% compared to 31% for online evaluations). Furthermore, in a presentation by Sorenson and Johnson (2006), there was no overall significant difference between online and paper ratings at Brigham Young University. Other studies have found that an overarching question is answered more favorably by online evaluation students, with the rest of the questions showing no significant difference (Liu, 2006). Various other studies have found no significant difference in the total mean quantitative score, but have differences when comparing individual questions (Avery et al., 2006; Cao, Clark, Schirmer, and Nelson, 2007). Not all studies have found that online evaluations are either positive or neutral. Chang (2003) found that paper evaluations produced higher scores for individual questions and total scores. Overall, there are mixed findings with little or no effect.

Other Factors

Table 2, reproduced from Marsh & Roche (1997, p. 1194), summarizes the factors discussed here, as well as other factors that have been discussed in the vast evaluation

literature as possible threats to validity. Some suspected factors, such as the gender or rank to instructors, have been found to have little or not effect. Others affect ratings. Interestingly, courses that are difficult or have heavy workloads tend to be rated higher than less challenging courses. Ratings tend to be somewhat higher if they are not anonymous or the instructor is present, which is why SOTES are supposed to be administered by student proctors with no interference from faculty members. They also tend to be higher if ratings are known to be used for tenure and promotion decisions.

Table 2
Overview of Relationships Found Between Students' Ratings and Background Characteristics

Background characteristics	Summary of findings
Prior subject interest	Classes with higher interest rate classes more favorably, although it is not always clear if interest existed before the start of the course or was generated by the course or the Instructor.
Expected grade-actual grade	Class-average grades are correlated with class-average students' evaluations of teaching, but the interpretation depends on whether higher grades represent grading leniency, superior learning, or preexisting differences.
Reason for taking a course	Elective courses and those with a higher percentage of students taking the course for general interest tend to be rated higher.
Workload -difficulty	Harder, more difficult courses requiring more effort and time are rated somewhat more favorably.
Class size	Mixed findings but most studies show smaller classes are rated somewhat more favorably, although some find curvilinear relationships where large classes also are rated favorably.
Level of course or year in school	Graduate-level courses are rated somewhat more favorably; weak, inconsistent findings suggest upper division courses are rated higher than lower division courses.
Instructor's rank	Mixed findings but little or no effect.
Sex of instructor or student	Mixed findings but little or no effect.
Academic discipline	Weak tendency for higher ratings in humanities and lower ratings in sciences, but too few studies to be clear.
Purpose of ratings	Somewhat higher ratings if ratings are known to be used for tenure-promotion decisions.
Administrative conditions	Somewhat higher if ratings are not anonymous and the instructor is present when ratings are being considered.
Students' personality	Mixed findings but apparently little effect, particularly because different "personality types" may appear in somewhat similar numbers in different classes.
Online vs. paper ratings	Mixed findings but little or no effect.

Note. Particularly for the more widely studied characteristics, some studies have found little or no relation or even results opposite to those reported here. The size, or even the direction, of relations may vary considerably, depending on the particular component of students' ratings that is being considered. Few studies have found any of these characteristics to be correlated more than .30 with class-average students' ratings, and most relations are much smaller.

References

Aleamoni, L. M., & Thomas, G. S. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9(1),* 67-84.

Arreola, R.A. (2000). *Developing a comprehensive faculty evaluation system*. Bolton, MA, Anker Publishing.

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.

Braskamp, L. A., Ory, J. C., and Pieper, D. M. "Student Written Comments: Dimensions of Instructional Quality." *Journal of Educational Psychology,* 1981, *73,* 65–70.

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.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research*. (IDEA Paper No. 20). Manhattan: Kansas State University, Center for Faculty Evaluation and Development. Available at: http://www.theideacenter.org/sites/default/files/Idea_Paper_20.pdf

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development. Available at: http://www.theideacenter.org/sites/default/files/Idea_Paper_32.pdf

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

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.

D'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52(11),* 1198-1208.

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

Fernandez, J., Mateo, M. A., & Muniz, J. (1998). Is There a Relationship between Class Size and Student Ratings of Teacher Quality? Educational and Psychological Measurement, 58(4), 596-604.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52(11),* 1182-1186.

Greenwald, A. G., & Gillmore, G. M. (1997a). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Education Psychology, 89(4),* 743-751.

Greenwald, A. G., & Gillmore, G. M. (1997b). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52(11),* 1209-1217.

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.

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 48(1),* 25-31.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & Mets, L.A. (Eds), *The student rating debate: Are they valid? How can we best use them*? (pp. 9-25). San Francisco: Jossey-Bass.

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

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.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707-754.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education,* 7, 9-18.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. The critical issues of validity, bias, and utility. *American Psychologist, 52(11),* 1218-1225.

Mateo, M.A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational and Psychological Measurement, 56(5),* 771-778.

McKeachie, W. J. (1997). Student Ratings. The validity of use. *American Psychologist, 52(11*), 1182-1186.

Nuhfer, E. B. (2003). Of what value are student evaluations? Pocatello: Idaho State University, Center of Teaching and learning. Available at: http://www.isu.edu/ctl/facultydev/extras/student-evals.html

Richardson, J. T. E. (2005). Instruments for obtaining feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30(4),* 387-415.

Sorenson, L. & Johnson, T. (2006). Online Student Ratings of Instruction. Paper presented at Town Hall Meeting April 12, 2006, San Jose, CA.

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,* 1332-140.

Stewart, C. T. & Malpass, L. F. (1966). Estimates of achievement and ratings of instructors. *Journal of Educational Research, 59,* 347-350.

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71(3),* 293-302.

Sujitparapitaya, S. & Briggs, J. (2010). Does a Delivery Method Matter? : A Comparison between Online and Paper Teaching Evaluations. Paper presented at the Student Evaluation Review Board, San Jose, CA.