Interpretation Guide

For

Student Opinion of Teaching Effectiveness (SOTE) Results

Prepared by:

Office of Institutional Research
and
Student Evaluation Review Board

May 2011
Table of Contents

Introduction ......................................................................................................................... 3
The SOTE Form .................................................................................................................. 3
Interpretation of the SOTE Ratings .................................................................................... 5
Factors Affecting SOTE Ratings ....................................................................................... 7
  Overview of Reliability and Validity .............................................................................. 7
  Expected and Actual Grades ......................................................................................... 8
  Class Size ..................................................................................................................... 9
Student Level ................................................................................................................... 9
Course Choice .................................................................................................................. 9
College Level Comparisons ............................................................................................ 10
Online vs. Paper Administration ....................................................................................... 10
Other Factors .................................................................................................................. 10
References ....................................................................................................................... 12
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 Policy F04-1).

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 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.

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.

---

1 The SOLATE form was designed in 1998. Norms for the form have not been updated since then. The 1998 interpretation guide, Interpretation Guide for SJSU SOTE and SOLATE Surveys, should be used for the SOLATE form. It can be found at: [http://www.sjsu.edu/senate/s98-4.htm](http://www.sjsu.edu/senate/s98-4.htm).
Table 1
SOTE Rating Questions and Correlations

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Questions</th>
<th>Correlation With Q13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>1. Demonstrated relevance of the course content</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>2. Used assignments that enhanced learning</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>4. Was responsive to questions and comments from students</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>5. Established an atmosphere that facilitated learning</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>8. Showed strong interest in teaching this class</td>
<td>0.71</td>
</tr>
<tr>
<td>Learning Environment</td>
<td>3. Summarized/emphasized important points</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>9. Used intellectually challenging teaching methods</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>11. Helped students analyze complex/abstract ideas</td>
<td>0.78</td>
</tr>
<tr>
<td>Helping Students Think</td>
<td>6. Was approachable for assistance</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>7. Was responsive to the diversity of students in class</td>
<td>0.69</td>
</tr>
<tr>
<td>Responsiveness to Students</td>
<td>10. Used fair grading methods</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>12. Provided meaningful feedback about student work</td>
<td>0.74</td>
</tr>
<tr>
<td>Grading/Feedback</td>
<td>13. Overall, this instructor’s teaching was</td>
<td></td>
</tr>
<tr>
<td>Overall Effectiveness</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. 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

---

2 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.
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
instructor. Pages 3-4 allow students to provide written evaluations of the instructor’s
strengths and/or weaknesses, and to provide 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 are 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.) The usefulness and validity of the ratings will be degraded if ratings within the norm area are interpreted as anything other than typical. 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 consistently below (or above) the norm then the item should be noted as important. If an item mean is inconsistently 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.

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.3 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.

---

3 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 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 of instructors, have been found to have little or no 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**

<table>
<thead>
<tr>
<th>Background characteristics</th>
<th>Summary of findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior subject interest</td>
<td>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.</td>
</tr>
<tr>
<td>Expected grade-actual grade</td>
<td>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.</td>
</tr>
<tr>
<td>Reason for taking a course</td>
<td>Elective courses and those with a higher percentage of students taking the course for general interest tend to be rated higher.</td>
</tr>
<tr>
<td>Workload -difficulty</td>
<td>Harder, more difficult courses requiring more effort and time are rated somewhat more favorably.</td>
</tr>
<tr>
<td>Class size</td>
<td>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.</td>
</tr>
<tr>
<td>Level of course or year in school</td>
<td>Graduate-level courses are rated somewhat more favorably; weak, inconsistent findings suggest upper division courses are rated higher than lower division courses.</td>
</tr>
<tr>
<td>Instructor's rank</td>
<td>Mixed findings but little or no effect.</td>
</tr>
<tr>
<td>Sex of instructor or student</td>
<td>Mixed findings but little or no effect.</td>
</tr>
<tr>
<td>Academic discipline</td>
<td>Weak tendency for higher ratings in humanities and lower ratings in sciences, but too few studies to be clear.</td>
</tr>
<tr>
<td>Purpose of ratings</td>
<td>Somewhat higher ratings if ratings are known to be used for tenure-promotion decisions.</td>
</tr>
<tr>
<td>Administrative conditions</td>
<td>Somewhat higher if ratings are not anonymous and the instructor is present when ratings are being considered.</td>
</tr>
<tr>
<td>Students' personality</td>
<td>Mixed findings but apparently little effect, particularly because different &quot;personality types&quot; may appear in somewhat similar numbers in different classes.</td>
</tr>
<tr>
<td>Online vs. paper ratings</td>
<td>Mixed findings but little or no effect.</td>
</tr>
</tbody>
</table>

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


