Interpretation Guide

For

Student Opinion of
Teaching Effectiveness (SOTE) Results

SAN JOSÉ STATE
UNIVERSITY

Prepared by:

Office of Institutional Research
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 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.[1] 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

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 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 SOTE. 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

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." Tpiach3 0 TDiots arr 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

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 inArreola9.385 0 TD up85 divison rcun0

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