Student Opinion of Teaching Effectiveness (SOTE)
Pilot Test Report

Jodie B. Ullman
SOTE Pilot Test Coordinator
CSUSB Department of Psychology

Submitted August 23, 2006

Acknowledgments

The following colleagues have contributed substantially to the completion and success of the SOTE pilot. Muriel Lopez-Wagner, Assistant Director of Institutional Research; Ramiro Diaz-Granados and Luisa Hawkins, Data Center Services; the SOTE Qualitative Coders: Jeff Hackel (Geography and Environmental Studies) Jake Zsu (Information and Decision Sciences), and K. Darcy Otto (Philosophy), Shawnee McMurran (Natural Sciences), and Rowena Santiago, Teaching Resource Center.

SOTE Overview

In AY 05-06 the SOTE (Student Opinion of Teaching Effectiveness) Task Force under the direction of Jeff Hackel, and with extensive consultation with CSUSB faculty, staff, and administrators, developed a teaching effectiveness survey based on current research. After development, which included several small pilot tests, the SOTE process was approved almost unanimously by the CSUSB Faculty Senate. The general goals of the SOTE are threefold: 1) provide a process whereby students can provide feedback on a faculty members' teaching, 2) provide instructors with better feedback to improve their teaching, and 3) provide evaluators with adequate data for the retention, promotion and tenure (RPT) awarding process.

The SOTE process contains three sections: 1) an evaluation section with five questions (three querying interest, attendance, and reason for taking the class and two questions probing teaching effectiveness); 2) an optional, supplemental form that allows individual faculty to select questions to investigate more deeply aspects of his/her teaching; and 3) an optional comment form to be used by faculty to describe any particular aspects of the class that may affect the SOTE evaluation. Those who designed the SOTE evaluation form sought primarily to: 1) encourage students to write more, and 2) reduce the number of evaluation questions, given that published, substantive research in this area has demonstrated that one dimension underlies most teaching effectiveness tools.

Goals of Pilot Study

Prior to SOTE implementation in AY 06-07, President Karnig requested that a pilot be done to ascertain that the SOTE process would produce student-based evaluative data on teaching that is comparable to the presently used SETE (Student Evaluation of Teaching Effectiveness). The Executive Committee of the Faculty Senate and the SOTE Implementation Committee selected Dr. Jodie Ullman, Associate professor of psychology, to direct the pilot. Dr. Ullman was asked to perform a full pilot study on the evaluative SOTE form; thus, the pilot did not assess the whole SOTE process but only that part of the process that produces data used directly in the RPT process. The pilot study assesses the SOTE evaluation form and the data produced from it. Specifically this pilot evaluates the extent to which the SOTE: 1) provides a process whereby students can provide feedback on a faculty members' teaching, 2) provides instructors with better feedback to improve their teaching, and (3) provide evaluators with adequate data for the retention, promotion and tenure (RPT) awarding process. Data collection for the pilot was done in Winter Quarter, 2006.

Pilot Test Procedure

SOTE Class selection

To ensure adequate statistical power for the analyses, a pilot sample of 200 classes was selected proportionally from each college's lower, upper, and graduate level courses (see Table 1). Due to RPT concerns, especially potential reactivity of the SETE and SOTE, the classes that were eligible for inclusion in the SOTE pilot test were classes that were not also evaluated with the SETE.

In Winter Quarter there were 2337 courses taught; of these, 1085 were evaluated with the SETE. Lecture and seminar type courses are the most prevalent type of the course offered on campus; therefore, due to time and budget constraints, the decision was made to pilot test the SOTE only in lecture and seminar type courses. After deletion of courses that were to be evaluated with the SETE, courses taught at the Palm Desert campus, and courses other than lecture and seminar (Distance Learning = 56, Palm Desert = 12, UNV = 6, Remedial = 1,Activity Class = 143, Lab Class = 96, Supervision = 453, Recitation = 3, Credit by Exam only class = 3), there were 479 classes that fit the criteria for selection. Institutional Research then drew a stratified random sample of 200 of these classes (6692 students). (See Table 1 for details of stratified random sample.) Table 1 presents the total number of lecture and seminar classes in the five Colleges in Winter Quarter 2006. The table is divided by division (lower, upper and graduate). The first line in each cell represents the actual number lecture and seminar classes offered in each of the five colleges divided by division. The second line gives the percent of courses offered by the College in that quarter within each division. For example BPA offered 28 lower division courses in Winter 2006 and this was 7% of the lower division course offered by the five colleges. (Note: due to rounding error percentages may not sum to 1.) The number following the percentage is the number of courses randomly drawn from the specific College and division to create a representative pilot sample of 200 courses.

Table 1. Winter Quarter Courses by College and Division.

Division	College					Grand Total
	BPA	EDU	HUM	NS	SBS	
Lower Division	28 (7% - 4)		205 (50% - 31)	111 (27%-17)	63 (15% - 9)	**407** **(30% - 61)**
Upper Division	124 (20% - 19)	56 (9% - 8)	158 (25% - 24)	135 (21% - 20)	155 (25% - 23)	**628** **(47% - 94)**
Graduate Division	41 (14% - 6)	137 (21% - 21)	29 (10% - 4)	37 (12% - 6)	53 (18% - 8)	**297** **(22% - 45)**
Grand Total						**1332** **(200)**

SOTE Form and SOTE Report Form

The SOTE evaluation form (see Appendix 1) contains five questions; three probe interest and motivation in the class: 1) "Rate your interest in the subject matter of this course before you

took the course," 2) "How many courses did you attend?," 3) "Why did you take this course?". And two questions that evaluate the course and the instructor: 4) "How would you rate the overall quality of instruction in this course?" and 5) "How would you rate the instructors' contribution to your learning in this course?" The primary focus of this pilot is on the analysis of the evaluative questions

The form used in the pilot did have minor formatting issues that were corrected with the help of the computer programmer. Several of the forms needed to be hand scored because the directions on how to fill in the bubbles were unclear. The programmer tackled this time intensive task and checked and hand scored forms when necessary. The software company was alerted to this problem and modifications in the instructions on scoring have now been added to the SOTE form. The comment section is placed well – immediately following each evaluative question.

The SOTE report form given to instructors, and available to evaluators, also underwent a variety of minor editing after the pilot data was collected. One small issue in the form, is that the student comments for each question are grouped together, but are separated from the rating. With the current version of the software, it is not possible to link a specific student's comment with that student's numerical rating unless the actual forms are examined. It should be noted however, that the actual forms will be available, in the usual way, should an instructor or evaluator want to link a comment with a rating. One particularly nice feature of the SOTE process is that the student forms can be accessed from via computer. During the first phase of SOTE implementation, each faculty will have access to his/her own SOTE results, including student comments. Later, it may be that evaluators will be allowed to do the same.

SOTE Pilot Sample Recruitment Method

Initial letters explaining the study and requesting participation in the pilot were sent to the faculty teaching the randomly selected courses. The instructors were asked to contact Jodie Ullman if they chose not participate. These faculty were removed from the database and additional classes were randomly selected and request letters were sent to the replacement sample.

During the 9th week, the SOTE implementation committee delivered packets of evaluation forms and pencils to the appropriate departments for distribution. Participating faculty were asked to administer the SOTE as they would administer the SETE. The forms were then either returned to Dr. Ullman or sent directly for processing.

The response rate was acceptable; 200 classes were selected, 13 faculty declined to participate prior to SOTE packet distribution, and 43 faculty did not return the SOTE forms. Therefore the final sample comprised data from 157 classes and 3828 students (78.5% class level response rate).

The distribution of the rank of the instructors who were randomly selected and who returned the survey is presented in Table 2 below. The instructor rank was determined by searching the University directory and using the rank given in the online directory. If an instructor was not listed, the instructor was coded as Lecturer. Due to potential classification errors, caution should be used interpreting the data associated with the lectures in the table below. The data in the last column (percentage of faculty by rank at the University) was taken from Academic personnel website for 2005-2006. *Note the 8.4% for lectures in the University data includes only full-time lectures (not part-time).

Table 2. Rank of Participating Faculty in SOTE pilot

Rank	Final Sample	Percentage returned within rank	Percentage returned of final pilot sample	Percentage of faculty by rank at the University
Lecturer	39	80%	25%	8.4%*
Assistant	21	74%	13%	21.1%
Associate	31	67%	20%	21.6%
Full	64	78%	41%	48.7%
Emeriti	2	67%	1%	
Total Final Sample	157			

Procedures of selection of SOTE comments to be analyzed. A random sample of the SOTE comments was analyzed. Expert judges were trained on the qualitative scoring form in Appendix 2. Responses to the open ended questions were classified into the following categories (command of subject matter, organization of instructional materials, effectiveness in instruction/feedback/instructor characteristics and/or qualities, and academic assessment of students). These categories were selected as they broadly reflect the categories that instructors are evaluated on from the FAM. Of interest also were not only the type of comments, but also the frequency of comments; therefore, a count of responses for each question was also recorded.

Procedures for SETE comparison group.

An anonymous dataset of all courses that were evaluated with the in SETE Winter 2006 was obtained from Institutional Research (IR). Given that this dataset was anonomized it was not possible to record data on demographics about the instructors (such as rank) who were teaching the specific courses, nor was it possible to qualitative code student SETE comments.

Analysis Plan

The goals of the analysis were to; 1) evaluate some of the psychometric properties of the SOTE form and the SETE form, and 2) compare the psychometric properties of the SOTE and SETE so that recommendations could be made for optimal use of the SOTE form. For the primary analyses, the SETE and SOTE datasets were analyzed separately. The data from both the SETE and SOTE violate the assumption of normality (the data are not normally distributed, "follow a normal curve") therefore statistical methods that account for the nonnormality were used whenever available.

The results section is structured in the following manner. First, an analysis of the psychometric properties of the SETE is presented. This is followed by the analysis of the SOTE, and comparison of the quantitative portions of the SETE and the SOTE. Following this quantitative section, the results of the qualitative SOTE analysis are discussed. After the technical results section is presented, the results are interpreted in nontechnical language, and the implications of the these results are presented.

Analysis of SETE classes

Data from 1087 classes (404 lecture, 474 seminar, 112 lab, 81 activity, 12 supervision, and 4 rec) from 34,819 students was used in the analysis. This is data is multilevel (students nested within classrooms). Therefore multilevel analyses were used throughout. Analysis of data without consideration of the nested structure is associated with substantial increases in Type I error rates (Raudenbush & Bryk, 2002).

Table 3 presents the medians, means, standard deviations, skewness coefficients, and z score for skewness for the questions on the SETE. Due to the multilevel nature of the data, these summary statistics were calculated for each class individually and then combined across classes. As is seen in Table 3, the skewness and z score for skewness columns of the data are highly nonnormal; therefore, statistical estimation methods that do not require normality were employed (Ullman, 2001).

This significant nonnormality ($p < .001$) indicates that with the exception of question 5, the mean is not an appropriate measure of central tendency and will be biased downward: it will be spuriously small. Thus, in this dataset the median is the appropriate measure of central tendency.

Table 3. Descriptive Statistics for the SETE questions Winter 2006

SETE Question	Median	Mean	Standard Deviation	Skewness	Z skewness
1. Rate how well your instructor knew the subject matter of this course.	4.00	3.78	0.37	-1.74*	-4.48
2. Rate how well prepared your instructor was for class sessions.	4.00	3.68	0.46	-1.61*	-3.7
3. Rate how well your instructor organized the course.	4.00	3.57	0.54	-1.4*	-3.09
4. Consistent with class size, how well the instructor encouraged and was responsive to student questions and comments.	4.00	3.68	0.48	-1.66*	-3.91
5. Consistent with class size, rate the usefulness of the instructor's feedback on your performance.	4.00	3.54	0.58	-1.39	-2.95
6. Rate how well the graded material (tests, papers, projects, etc.) reflected the course objectives).	4.00	3.55	0.58	-1.4*	-3.01
7. Rate how well the objectives and requirements were explained.	4.00	3.55	0.56	-1.38*	-3.06
8. Rate the instructor's ability to make the course material understandable.	4.00	3.56	0.56	-1.44*	-3.3
9. Rate how well the instructor stimulated interest in the subject.	4.00	3.56	0.57	-1.49*	-3.35
10. Rate the overall quality of instruction in this class	4.00	3.60	0.51	-1.48*	-3.34

*$p < .001$

Intraclass Correlation Coefficient. Because the scores on the SETE questions are from students within classrooms, it is possible to partition the variability in student ratings into two parts: the portion of variance in the ratings due to differences among classes (e.g. different instructors, levels, Colleges, days of week, times of day) and the portion of the variability in the ratings due to differences among students within classes (e.g. gender, age, ethnicity, motivation, personality factors).

The intraclass correlation coefficient is used to partition the variability,(technically, the intraclass correlation coefficient is the ratio of the between group variance divided by total variance). Values of the intraclass correlation coefficient (ranging from 0 to 1) indicate the proportion of variability that is attributable to differences in ratings due to differences among groups (classes in this case) relative to the total variability (variability in ratings due to differences among classes and differences among students within classes). An intraclass correlation coefficient equal to zero would indicate that none of the variability in student ratings was accounted for by differences among classes. An intraclass coefficient equal to one would indicate that all of the variance in student ratings of teaching effectiveness is due to differences among classes. This is a squared correlation coefficient that can be converted a percentage of variance. The percentage is interpreted as the percent of variance accounted for by differences among classes (between group variance). One minus the intraclass coefficient is interpreted as the percent of variance that is accounted for by differences among students within classes (within group variance) **(**Raudenbush & Bryk, 2002**).** Ideally, a question that is used to evaluate instructors based on differences in student ratings between classes (differences between instructors) should have an extremely high intraclass correlation (>.80) indicating that underlying mechanism responsible for the variability in the ratings is due to characteristics of the class (for example the instructor), not characteristics of students within the class.

The intraclass correlations coefficients for the SETE questions are given in Table 4. The percent of variability in ratings as a result of classes differing (for example variability because instructors differ, class levels differ, colleges differ, degree of difficulty of course differs) ranges from 16.20% for question 1 to 24.08% question 10. For the sake of clarity, considering only question 10, “Rate the overall quality of instruction in this class,” 24.08% of the variance in student ratings on this question is due to differences between classes. Of that 24.08% of the variance, some of it is due to differences in instructors, and some of it is due to other “between course” characteristics. The majority of the variance in the scores (75.92%) is due to idiosyncratic characteristics of the students. Said another way, 24% of the variability in student ratings is due to the course, and 76% of the variability in student ratings is due to students being individuals, i.e., responding uniquely regardless of the course being rated. And, again, it must be emphasized that of the 24% of the variability in student ratings in this example, it is impossible to know how much of it is based on the instructor alone.

Table 4. SETE Intraclass Correlations given in percentage form. Percentage of variance explained by differences among classes and differences among students within classes.

SETE Question	**Percent of variance explained by differences between classes (based on intraclass correlation coefficient).**	**Percent of variance explained by differences between students within classes.**
1. Rate how well your instructor knew the subject matter of this course.	16.20%	83.80%
2. Rate how well prepared your instructor was for class sessions.	21.72%	78.28%
3. Rate how well your instructor organized the course.	22.68%	77.32%
4. Consistent with class size, rate how well the instructor encouraged and was responsive to student questions and comments.	20.42%	79.58%
5. Consistent with class size, rate the usefulness of the instructor's feedback on your performance.	20.99%	79.01%
6. Rate how well the graded material (tests, papers, projects, etc.) reflected the course objectives).	20.41%	79.59%
7. Rate how well the objectives and requirements were explained.	21.44%	78.56%
8. Rate the instructor's ability to make the course material understandable.	23.74%	76.26%
9. Rate how well the instructor stimulated interest in the subject.	21.40%	78.60%
10. Rate the overall quality of instruction in this class.	24.08%	75.92%

Correlations among questions on the SETE – How much unique information about teaching effectiveness is available using the SETE? One concern about use of the SOTE evolves around the number of evaluative questions in the SOTE (2) relative to the SETE (10). The concern has been that the SOTE might provide less numeric information than the SETE. This concern was addressed in multiple ways using the SETE data and SOTE data.

Using a multilevel approach, the variability in the student ratings for each question was divided into the variability due to differences between classes and the variability due to differences between students within classes. For evaluation of teaching effectiveness, the appropriate variance to analyze is the between class variance. This is variability in student ratings that is explained by differences among classes, and of course, one important difference among classes is the instructor.

Using the between score variability, as a first step in examining these relationships, the bivariate correlations among the questions were examined and are presented in Table 5. These correlations measure the degrees of linear relationship between questions and are very high (ranging from .71 to .95). As it is can be somewhat difficult to conceptualize the importance of the degree of linear relationship, Table 6 presents the correlations squared. These squared correlations can be interpreted as the percent of shared, or common, variance between two questions. A value of one would indicate that the two questions are identical and a value of 0 would indicate that the questions were entirely independent (conceptually completely different) of one another. If the SETE is tapping into separate, unique, aspects of teaching effectiveness with the 10 questions, then the percent of shared variance between questions should be zero or very low. This is not the case. The percent of shared variance ranges from 50% to 91%. Of particular interest is the percent of variance shared between question 10, "Rate the overall quality of instruction in this class" and the other questions (with common, shared, variance ranging from 68% to 91%). This provides strong evidence that SETE measures one construct: overall quality of instruction.

From this data it can be concluded that it is spurious to interpret the results of each question as unique, or independent of the other questions. However, to further probe this issue a multilevel confirmatory factor analysis was performed.

Table 5 SETE - Correlations Between Questions Based on Between Class Variance

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Question 1	1.00									
Question 2	0.82	1.00								
Question 3	0.78	0.91	1.00							
Question 4	0.75	0.76	0.82	1.00						
Question 5	0.74	0.77	0.84	0.91	1.00					
Question 6	0.71	0.74	0.84	0.84	0.90	1.00				
Question 7	0.77	0.84	0.91	0.87	0.89	0.89	1.00			
Question 8	0.79	0.80	0.88	0.89	0.90	0.87	0.93	1.00		
Question 9	0.78	0.77	0.83	0.86	0.88	0.83	0.87	0.92	1.00	
Question 10	0.83	0.86	0.92	0.91	0.93	0.90	0.95	0.95	0.93	1.00

Table 6 Percent of Common (shared) Variance Between Questions on the SETE based on Between Class Variance

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Question 1										
Question 2	68%									
Question 3	61%	83%								
Question 4	56%	58%	67%							
Question 5	54%	59%	70%	83%						
Question 6	50%	55%	71%	70%	81%					
Question 7	59%	70%	84%	76%	79%	79%				
Question 8	62%	65%	77%	79%	81%	75%	86%			
Question 9	61%	60%	69%	75%	77%	69%	76%	85%		
Question 10	**68%**	**74%**	**85%**	**83%**	**86%**	**81%**	**89%**	**91%**	**86%**	

Multilevel Factor Analysis – Is there evidence for one underlying construct of teaching effectiveness? From the bivariate correlations based on the between group variability it can be concluded that it is spurious to interpret the results of each question as unique, or independent of the other questions. However, to further probe this issue, a multilevel confirmatory factor analysis was performed to test the hypothesis that one general latent construct drives the observed bivariate correlations. Due the extreme non-normality of the data the models were estimated using maximum likelihood estimation and evaluated with the Satorra-Bentler Scaled Chi square (Satorra & Bentler, 2001). The standard errors of the path coefficients were also adjusted to the extent of the nonnormality (Bentler& Dijkstra, 1985**)**. Further statistical details of the model are not reported in this here, but are available from Dr. Ullman. A one factor model fit the data extremely well, accounting for approximately 90% of the variance in ratings between classes. This indicates that there is one dimension of teaching effectiveness that underlies the SETE.

Summary of Results for SETE data Winter 2006

1. Questions 1-4, 7-10 are significantly skewed and mean is biased downward – the values of the mean are too small. The median is appropriate measure of central tendency

2. The SETE does not contain 10 unique dimensions (aspects) of teaching effectiveness but instead measures the general concept of teaching effectiveness. Question 10 adequately measures the teaching effectiveness, the other questions add very little, if any, new information.

3. Most (more than 75%) of the variability in the student ratings of teaching effectiveness is due to differences in student characteristics and less than 25% of the variability in student ratings of teaching effectiveness is due to differences among classes.

4. From this analysis it can be concluded that there is one construct that underlies the SETE. The amount variability in student ratings due to differences in classes is extremely small (less than 25%) so caution should be used evaluating numerical results.

Analysis of SOTE Classes data

Data from 157 classes and 3,828 students was used in this analysis. This data is multilevel (students nested within classrooms)in structure; therefore multilevel analysis were used throughout. Analysis of data without consideration of the nested structure is associated with substantial increases in Type I error rates (Raudenbush & Bryk, 2002).

Table 7 presents the medians, means, standard deviations, skewness coefficients, and z score for skewness for the evaluative questions on the SOTE. Due to the multilevel nature of the data these summary statistics were calculated for each class individually and then combined across classes. As is seen in Table 7 in the skewness and z score for skewness columns this data is normally distributed for question 4, "How would you rate the overall quality of instruction in this course?" and nonnormally distributed for question 5, "How would you rate the instructor's contributions to your learning in this course?" Therefore, the mean is an appropriate measure of central tendency for question 4 but not for question 5. For question 5, the significant ($p < .001$) nonnormality indicates that the mean is not an appropriate measure of central tendency and will be biased upward – too large. (Note that the SOTE evaluative questions are coded on a 1- 6 scale where 1 means excellent and 6 means unsatisfactory. This is the different from the SETE). In this pilot study the median is the appropriate measure of central tendency for question 5.

Table 7 Distributional Characteristics of the SOTE evaluation questions.

SOTE Question	Median	Mean	Standard Deviation	Skewness	Z skewness
4. How would you rate the overall quality of instruction in this course?	2.00	1.78	0.78	0.99	2.17
5. How would you rate the instructor's contributions to your learning in this course?	2.00	1.38	0.47	1.35*	3.05

*p < .001

Intraclass Correlation Coefficient. Because the scores on the SOTE questions are from students within classrooms it is possible to partition the variability in these scores into two parts: the portion of the variability due to differences in between classes and the portion of variability accounted for by differences among students within classes. This partitioning is calculated through the intraclass correlation coefficient,(a ratio of between group variance divided by total variance). Values of the intraclass correlation coefficient (ranging from 0 to 1) indicate the proportion of variability to differences between groups (classes in this case) relative to the total variability (variability due to different classes and different students within classes). This

squared correlation coefficient can be converted to the percent of variance that is explained by differences between classes (between group variance) versus the percent of variance that is explained by differences within classes (within group variance) **(**Raudenbush & Bryk, 2002**).** Ideally, a question that is used to identify and evaluate differences between classes (differences between instructors) should have an extremely high intraclass correlation (>.80) indicating that underlying mechanism responsible for the variability in the scores is due to characteristics of the class not characteristics of the students within the class.

The intraclass correlations coefficients for the SOTE questions are given in Table 8. 21% of the variability in student ratings of question 4 is explained by differences in among classes and 79% of the variability is explained by differences among students within classes. 15% of the variability in student ratings question 5 is explained by difference in classes and 85% of the variability is explained by differences among students. In both questions the majority of the variance in the student ratings is due to idiosyncratic characteristics of the students not characteristics of the course.

Table 8. Percent of variance in student ratings of evaluative SOTE questions accounted for by differences among courses and differences among students.

SOTE Question	**Percent of Variance in student ratings explained by classes differing (based on intraclass correlation coefficient)**	**Percent of Variance in student ratings explained by students differing**
4. How would you rate the overall quality of instruction in this course?	21%	79%
5. How would you rate the instructor's contributions to your learning in this course?	15%	85%

Correlation between the evaluative SOTE questions. How much unique information is available with these two questions? Very little. Using only the variability in student ratings due to class differences the correlation (linear relationship) between these questions = .93, 86% of the variance is shared common variance between these two questions.

Summary of Quantitative Results for SOTE sample

1. Question 4 is normally distributed and therefore allows for interpretation of the mean as well as the median.

2. The SOTE measures one dimension of teaching effectiveness. The evaluative questions are extremely highly correlated.

3. Most (more than 78%) of the variability in the student ratings of teaching effectiveness is due to differences among students and less than 21% of the variability in student ratings is due to differences in the classes.

4. From this analysis it can be concluded that there is only one construct that underlies the SOTE. The amount variability in student ratings due to differences in classes is extremely small (less than 21%) so caution should be used evaluating numerical results.

Comparison of the Psychometric Properties of the SETE and the SOTE

In this section the SETE and the SOTE will be compared in three different areas: 1) distribution properties and appropriate measures of central tendency, 2) amount of unique information available, 3) proportion of variability in student ratings that is attributable to differences in classes as opposed to differences among raters (the students).

1. In the SETE all questions, with the exception of question 5, are significantly skewed at $p <$ .001. Given this severe skewness the mean is significantly biased and will be consistently too small. The median is the appropriate measure of central tendency. Question 4 in the SOTE (How would you rate the overall quality of instruction in this course?) is normally distributed and therefore the mean is an appropriate measure of central tendency. Question 5 (How would you rate the instructor's contributions to your learning in this course?) is significantly skewed thus the mean is biased upward (too large) and the median is the appropriate measure of central tendency.

2. Given the bivariate correlations, percents of variance accounted for, and the multilevel confirmatory factor analysis of the SETE. It can be concluded that both the SETE and SOTE yield information on the general construct of teaching effectiveness and separate aspects of teaching cannot be disentangled for either. This finding seems to indicate the use of the SETE may lead to over-interpretation and spurious conclusions.

3. In both the SETE and the SOTE, most of the variability in student ratings for teaching effectiveness is due to the students' idiosyncratic characteristics. The majority of variability in student ratings is not due to differences among classes. This result argues strongly for cautious use of any numerical rating. This also argues for careful use of qualitative comments on teaching effectiveness.

4. There is no a loss of information between the SOTE and SETE.

Recommendations On Use of SOTE Based Psychometric Analysis

1. Use the SOTE in place of the SETE. As is evidenced in the quantitative analysis, there is no loss of information by assessing teaching effectiveness with two questions instead of ten. Additionally evaluators may prefer to interpret the mean instead of the median and this is feasible only with the SOTE.

2. Revise SOTE question 5. The pilot showed that only one dimension of teaching effectiveness is being measured with the SOTE. SOTE question 4 and SOTE question 5 are highly correlated. Only one question is needed to measure the construct of teaching effectiveness. SOTE question 4 is normally distributed and allows for use of the mean and the median. Although the intraclass correlation is low for SOTE question 4 it is substantially larger than SOTE 5. Therefore SOTE question 4 is preferred on statistical grounds. If two questions are desired, then an attempt should be made to differentiate question 5 from question 4. Due to the good psychometric properties of question 4, I would not alter it. Question 5 could be reworded as, "How would you rate your *professor's specific* contributions to your learning in this course?" The italics indicate the suggested word changes. This change would serve to differentiate the two questions more, with a student's more likely to be able to discern a difference between Question 4 and Question 5, with 4 being more general and 5 more specific. See the discussion below on student comments for further amplification.

3. Most of variability in student ratings of teaching effectiveness in both the SETE and the SOTE are due to idiosyncratic characteristics of the students and not characteristics of the instructor and course in general. This argues for extreme caution in the use of student course ratings in the evaluation process. One recommendation is to add the intraclass correlation coefficient and verbal interpretation to the summary form given to faculty and evaluators.

Qualitative Analysis of SOTE Open-ended Questions

Immediately below both question 4 and question 5, students are asked to give reasons for their ratings. One goal of this pilot was to assess the written student comments. This is a time intensive task; therefore, due to time and staffing constraints, a small portion of the SOTE forms were randomly selected, and the student comments were coded into four categories, corresponding conceptually to the areas for instructor evaluation in the FAM, (command of subject matter, organization of instructional materials, effectiveness in instruction/feedback/Instructor characteristics and/or qualities, and academic assessment of students). The coding form is attached in Appendix 2. Three members of the SOTE implementation committee coded responses from a total of 35 SOTE classes. The results are summarized in Table 9 below. The entries in the table represent the median number of responses, and the minimum and maximum number of responses coded into each category per class. The last row indicates the median percentage and minimum and maximum percentage of student SOTE forms that contained one or more comment.

Table 9. Distribution of coded open-ended comments in four categories

Area of Evaluation	**Median number (and range) of responses for question 9 per class** **"Rate the overall instruction in this course"**	**Median number (and range) of responses for question 10 per class** **"Rate the instructor's contribution to your learning in the class".**
Command of subject matter	3 (0 – 9)	1 (0 – 6)
Organization of instructional materials	1 (0 – 13)	0 (0 – 4)
Effectiveness in instruction/feedback/Instructor characteristics and/or qualities	11 (0 – 32)	7.5 (0 – 22)
Academic assessment of students	1 (0 – 7	0 (0 – 8)
Median (and range) percentage of students contributing comments	81.82% (28% - 100%)	51.72% (14% - 100%)

For both questions, most students wrote comments about their instructor's "Effectiveness in instruction/feedback/Instructor characteristics and/or qualities". Students made significantly more comments explaining their responses to Question 4 ("Rate the overall instruction in this course") than question 5 ("Rate the instructor's contribution to your learning in the class").

As in the quantitative analysis there were indications that students perceived these two questions as highly similar. The coders were asked to rate, on a 0 – 100% scale, how similar students' responses were to both question. This rating serves as an indicator of perceived similarity between the questions. The median rating of similarity was 80% with a range from 10% to 100% similar. Additionally, further evidence of the students' perception of similarity, as seen on question 5, when some of the students also commented "see above answer" or "see answer to question 4." The sharp drop in response rate on question 5 may also be due to students' perceptions that they were answering the same question twice.

This was an extremely time consuming task and the SOTE qualitative coders were unable to also do a word count to measure the length of responses. However, the informal perception among the coders was the student comments were longer and more detailed than those often read on the SETE. If this is true, it may simply be because students perceived this as a novel task. However, it may also be due to the structure of the SOTE. Students are asked to explain their ratings in the space immediately following the question. This placement may encourage more detailed responding. It also may be the case that the students' perceive less burden answering fewer questions and are more willing to write comments

Summary and Recommendations for Student Comments

1. Students seem to be making comments across a range of evaluation areas. The majority of comments for both questions are categorized as comments on effectiveness in Instruction, Feedback, and Instructor characteristics and/or qualities.

2. The qualitative analysis provides further evidence that the students do not perceive a difference between the two evaluative questions. If the decision is made to keep two evaluative questions then an effort should be made to reword question 5 to help students perceive the difference in content in the two questions.

3. The potential increase in the time it takes to evaluate a file due to reading more student comments may well be offset by the possible, future availability of the class SOTE forms and comments on evaluators' own computers.

4. Revise SOTE question 5 so that students perceive a stronger differentiation between it and SOTE question 4.

Overall Summary and Conclusions

The SOTE is feasible to implement and is an improvement over the SETE. The pilot test was an extremely important rehearsal for the technical processing aspect of the SOTE. Many glitches in the ultimate implementation will be avoided as a result of this pilot, and for picking up that students did not make as strong a distinction between SOTE questions 4 and 5 as was desired.

Distributional properties of the SOTE allow for interpretation of both the mean and median. Both the SETE and SOTE measure only overall teaching effectiveness despite ten questions probing different aspects of the teaching on the SETE and two questions on SOTE. Therefore the potential to spuriously over-interpret the data is minimized by using the SOTE. Using the SOTE offers students an enhanced opportunity to give instructors written feedback. Given the problems with the low variability (generally less than 25%) attributable to instructor/course, this improvement in written feedback is a strong component of the SOTE.

Caution is warranted in using any student rating data for evaluation purposes. Over 75% of the variability in student ratings of effectiveness is attributed to the student characteristics within a classroom rather than characteristics of the course, i.e., the instructor. It is difficult to develop teaching evaluation questions in which the variability in student rating is primarily a function of the differences among courses. Therefore, it seems important to add the intraclass

coefficient and a verbal interpretation of it to the SOTE report form. This will help provide perspective for interpreting the quantitative portion of the SOTE.

The SOTE evaluative questions are highly correlated. This was seen both statistically and in the written student comments. A recommendation is to revise question 5 on the SOTE form to try to differentiate it from question 4.

References

Bentler, P. M. & Dijkstra, T.(1985). Efficient estimation via linearization in structural models. In P.R. Krishnaiah (ed.), *Multivariate analysis VI* (pp. 9-42). Amsterdam: North-Holland.

Raudenbush, S. Bryk, A, (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd Ed.). Thousand Oaks CA: Sage Publications.

Satorra, A. & Bentler, P. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 4, 507-514.

Ullman, J. B. (2001). *Structural Equation Modeling*. In B.G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, (4th Ed.). Boston: Allyn & Bacon. (pp. 653-771).

Appendix 1. SOTE Standard Form

CALIFORNIA STATE UNIVERSITY, SAN BERNARDINO

Student Opinion of Teaching Effectiveness

5500 University Parkway, San Bernardino, CA 92407-2397

Name of faculty being rated

Class Description

Call Number

---

1. Rate your interest in the subject matter of this course before you took the class.

○ Very High ○ High ○ Moderately High ○ Moderately Low ○ Low ○ Very Low

2. How many class sessions did you attend?

○ All ○ Almost all ○ More than half ○ Less than half

3. Why did you take this course? Choose all that apply.

○ The course fulfills a general education requirement.

○ The course fulfills a requirement in my major.

○ The course will improve job/career opportunities.

○ I have enjoyed the professor's class(es) in the past.

○ The class sounded interesting to me.

○ Other: (Please specify)

4. How would you rate the overall quality of instruction in this course?

○ Excellent ○ Very Good ○ Good ○ Fair ○ Poor ○ Unsatisfactory

Please provide reasons why you gave the above rating.

5. How would you rate the instructor's contributions to your learning in this course?

○ Excellent ○ Very Good ○ Good ○ Fair ○ Poor ○ Unsatisfactory

Please provide reasons why you gave the above rating.

Appendix 2. SOTE Qualitative Coding Form

Comment Rating Sheet for SOTE Pilot Test

Course Number _____ How many students are enrolled_____ Rater Name _____

Part I. Please look at the first set of comments. These apply to the question that asks, "Rate the overall instruction in this course". Read through each comment and using the categories below categorize the comment. Comments may fit in more than one category – if so place tally marks in all appropriate categories. When you are done with this question please total your tally marks. Both positive and negative comments should be classified.

1. Command of Subject Matter Total Number of comments___

2. Organization of Instructional Materials Total Number of comments___

3. Effectiveness in Instruction/Feedback/Instructor characteristics and/or qualities Total Number of comments___

4. Academic Assessment of Students Total Number of comments___

5. How many students supplied comments for this question? _____ (Count the number of different comments)

Part II. Now look at the second set of comments. These apply to the question that asks, "Rate the instructor's contribution to your learning in the class". Read through each comment and using the categories below categorize the comment. Comments may fit in more than one category – if so place tally marks in all appropriate categories. When you are done with this question please total your tally marks. Both positive and negative comments should be classified.

6. Command of Subject Matter Total Number of comments___

7. Organization of Instructional Materials Total Number of comments___

8. Effectiveness in Instruction/Feedback/Instructor characteristics and/or qualities Total Number of comments___

9. Academic Assessment of Students Total Number of comments___

10. How many students supplied comments for this question? _____ (Count the number of different comments)

Part III. Use your judgment and compare the responses to questions 4 and 5. Do their responses indicate that they were answering two conceptually different questions or one overall question?

11. Please provide a rough percentage (0 –100%) score where 0 means that overall, students responses indicated that they were responding to two different questions – asking about different parts of the course and 100% means that the students essentially answered the same question twice (or just elaborated on the first response).

_____% (1 – 100%)

12. Generally were the comments generally POSITIVE MIXED NEGATIVE
13. Is there any other information on this set of forms that you would like to comment on?