S86-2 RATING SCALE: STUDENT EVALUATION OF TEACHING EFFECTIVENESS

Legislative History:

Document dated March 27, 1986.

At its meeting of March 17, 1986, the Academic Senate approved the following Policy Recommendation presented by Charles Whitcomb for the Instruction and Research Committee.

ACTION BY THE UNIVERSITY PRESIDENT:

"Approved and accepted as University Policy. Effective Immediately." Signed: Gail Fullerton, March 27, 1986.

RATING SCALE: STUDENT EVALUATION OF TEACHING EFFECTIVENESS

S 86-2

WHEREAS, The Rating Scale: Student Evaluation of Teaching Effectiveness (S 81-1) was adopted by the Academic Senate for a three year period ending in February, 1984, and

WHEREAS, The Senate extended the time during which the Scale could be used until June, 1986 (S 84-12), and

WHEREAS, The Senate recommended that faculty attitudes be surveyed concerning the instrument (S 84-12), and

WHEREAS, The Senate requested the Student Evaluation Review Board to make recommendations for further use of the form, and

WHEREAS, SERB has conducted the requested survey and is recommending certain modifications in the instrument, be it therefore

RESOLVED: That the Academic Senate accept SERB's recommendations (attached).

SERB FACULTY SURVEY:

SUMMARY OF STATISTICAL DATA AND WRITTEN COMMENTS

In Fall, 1984 the Student Evaluation Review Board (SERB) was directed to survey faculty opinion regarding the Student Evaluation of Teaching Effectiveness (SETE) form which has been in use the past three years. The questionnaire developed by SERB requested faculty to rate the adequacy of the 14 SETE items and to respond to 10 additional questions about the SETE form and process. In addition, faculty were invited to make written comments and suggestions for the evaluation questionnaire and rating process.

The first week in March of this year 1724 questionnaires were sent to departments for distribution to faculty. The 1724 figure was provided by the Academic Senate Office (the number included 22 administrators in the various Schools). Scorable documents were returned by 474 faculty (27.4% of those mailed). The small return may indicate that the majority of the faculty does not have strong concerns about the student rating program or process.

Statistical Data

The Sample

To determine the representativeness of the respondents, the percentage of faculty in each school and the percentage of respondents within each school was determined. Of the 474 respondents 374 (78.9% of the returned surveys) identified their School affiliation. These data are listed in Table 1. Information about numbers of faculty in each School was obtained from the Academic Senate Office.

TABLE 1:

[Original policy contains a table on file at the Academic Senate Office]

The respondents were fairly representatively spread across the Schools but slightly higher percentages of respondents from Applied Arts and Social Science and slightly lower percentages in Humanities & Arts and Science. The fact that fewer than 302 of the faculty returned the survey is probably much more important than minor discrepancies in proportional return among the Schools. Only 345 of the respondents (73% of the sample) indicated rank. Among these 345 respondents, 99 (28.7%) were Lecturers or Instructors, 30 (8.7%) were Ass't. Professors, 61 (17.7%) were Assoc. Professors, and 155

(44.9%) were Professors. The percentages of faculty by rank in the university are: 35% Professors, 11% Assoc. Professors, 4% Ass't. Professors, and the remaining 50% temporary Lecturers/Instructors. Proportional to their numbers in the university, the greatest response rate was from the Ass't. and Assoc. Professors.

Responses to the Survey Questions

Faculty used a 5-point scale (where 1 = outstanding and 5 = poor) to rate each of the 14 items on the Student Evaluation of Teaching Effectiveness (SETE) form. The percentages of responses for each category and means and standard deviations for these 14 rated items are listed in Table 2. For example, for item 1, 70.7% of the respondents reported the item to be outstanding, 2.7% rated it as poor, and the overall item mean was 1.5. In addition, faculty were asked to respond "yes," "no", or "no opinion" to 10 additional items about the SETE form (see Table 3).

TABLE 2

Percent of Faculty Responses to the 14 Rating Scale Items (N = 474)

[Original policy contains a table on file at the Academic Senate Office]

Item 1: Made course requirements clear.

Item 2: Collected enough information to measure what I learned.

Item 3: Was accessible for student inquiries outside of class.

Item 4: Used a variety, of teaching methods.

Item 5: Explained grading criteria.

Item 6: Played an important role in helping me learn.

Item 7: Was regularly available during office hours.

Item 8: Made presentations which were well organized and clear.

Item 9: Used fair methods of grading.

Item 10: Showed concern for students.

Item 11: Promptly returned tests, papers, and other assignments.

Item 12: Gave interesting presentations.

Item 13: Increased my interest in the subject.

Item 14: The overall effectiveness of the instructor was:

Inspection of the data in Table 2 reveals that items 1, 5, 8, 9, and 14 have the lowest means - all below a 2.0 which indicates high endorsement of the item (i.e., the average response was "very good" to "outstanding"). Item 4, on the other hand, has a mean of 3.2, indicating that the average response about the item was that it was only "fair." Variability in response is least for items 1 and 8 and greatest for items 3 and 4. This agrees with the percentage figures where items 1, 5, 8, 9, and 14 show more than 50% of the respondents rated the item "outstanding." Similarly, for item 4 more than one-quarter rates the item "poor."

When the Outstanding and Very Good categories are aggregated, all but item 4 are endorsed by more than 50% of the respondents. Items 1, 5, 8 ant 14 are seen as the best items; more than 552 of the respondents rated these items as "outstanding." The least preferred items were 3, 4, 11, 12, and 13. These items were rated as "outstanding" by 40% or fewer of the respondents.

It appears that the general attitude of the responding faculty is that the current set of rating items are more or less acceptable with the exception of item 4: "Uses a variety of teaching methods." Somewhat troublesome, however, is that 10% or more of the respondents reported five of the items (3, 4, 6, 12, and 13) to be "poor." Exactly what percentage is a "significant minority," warranting some action or response, is unclear; however, the 102 or more "poor" responses came from an average of 70 faculty - approximately 4% of the total faculty.

In an effort to determine more about the relationships among the 14 rating items according to the responding faculty, several other data analyses were completed. Among the generalizations that can be made from these analyses are:

1. All intercorrelations were positive indicating that if one item was rated by a person in the positive direction then, on the average, all items were in the positive direction ant vice versa.

2. Items 1, 2, and 5, dealing with making things clear, (whether requirements, grading criteria , or presentations) appear to the faculty as the most appropriate rating items. Clustering with these items is the one regarding fair (perhaps meaning clear?) grading practices (item 9).

3. Items 3, 4, 12, and 13 (was accessible, used a variety of methods, gave interesting presentations, and increased interest) were seen as the least acceptable items. Perhaps these items are not as objective or behavioral as other items and therefore are more susceptible to the "good guy/bad guy" halo effect.

The 14 SETE items were inspected taking into account faculty rank to see whether there was differential endorsement of items by rank. No obvious trends were apparent.

In addition to the 14 SETE items, 10 supplemental questions were asked to gather further information about faculty opinions about the SE instrument or the rating process. For each of these 10 questions respondents were asked to use the alternatives of "Yes", "No," or "No Opinion." A listing of these supplemental questions and faculty response to them are listed in Table 3. (Note: Some respondents answered only some of the supplemental questions so the number of responses to each question differ.)

The first supplemental question in the survey asked faculty to indicate whether they believed it is appropriate for students to evaluate faculty in any manner. This item was included in an effort to sample the proportion of faculty who might disapprove of any set of questions on a student evaluation survey. There were 43 (9% of the respondents) faculty who responded "no" to question 15; among the 39 who also reported academic rank were 25 Professors, 4 Assoc. Professors, 3 Ass't. Professors, and 7 Lecturers/Instructors. Looking at this subgroup's ratings. items 1, 5, and 8 were reported to be the best items (only 6-8 people rated these items as fair/poor). Items 4 and 12 were least liked by this subgroup where 29 and 20 of the 43 raters reported these items to be either fair or poor, respectively. On the average, those respondents who said no to question 15 tended to dislike all SETE items more than those who answered yes to the question. For example, of the 12 people in the total sample who rated item 1 as "poor," 6 were in the subgroup of 43 who reported that no evaluation is appropriate.

TABLE 3

Percentage Responses to the Additional 10 Questions

[Original policy contains a table on file at the Academic Senate Office]

Q15: Is it appropriate for students to evaluate your teaching in any manner?

Q16: Is the present questionnaire a relatively fair and reasonable evaluation instrument?

Q 17: Are class evaluations of this type an appropriate way for students to evaluate teachers?

Q 18: Should student evaluations be included in RTP deliberations?

Q 19: Is too much weight put on evaluations in RTP deliberations?

Q 20: Do we need a question about teacher's subject matter knowledge?

Q 21: Do we need a question about how much the student learned?

Q 22: Do we need a different rating form for seminar/discussion classes?

Q 23: Are the norms now used adequate and useful?

Q 24: Are the response alternatives appropriate?

Asked if the SETE instrument is relatively fair and reasonable (question 16), the opinion is divided where 52% say yes ant 39% say no. Because the item really asks two questions (fair and reasonable) it is impossible to say whether the SETE form is disliked by 39% because it is unfair or because it is not a reasonable instrument. The high percentage indicating the SETE form is unfair or unreasonable is not evident in the ratings of the 14 SETE items (see Table 2). This discrepancy may be due to dissatisfaction with one or more of the SETE items which then makes the form as a whole unreasonable or unfair. About two-thirds of the faculty report that the SETE type of evaluation is appropriate (question 17). There is general endorsement (67%) that student evaluations should be included in RTP procedures (question 18), but 41% believe that too much weight is put on these data (question 19). The need for an item about the instructor's comment of the subject matter was endorsed by 64% (question 20) and 51% agreed that there is a need for a question about amount students learned (question 21). There was general agreement (51% said yes) that a separate rating instrument is necessary for seminar /discussion classes. Opinion seems to be evenly divided about the adequacy and usefulness of the present SETE norms (question 23). More than half of the respondents reported the response alternatives of "outstanding - poor" are appropriate (question 24).

Faculty Written Comments

Faculty were also asked to provide comments and suggestions for the rating form and/or rating process. Comments were received by 180 (38%) of the respondents. The comments ranges from single short sentences to 2 typewritten pages, to copies of 11 student essays on the topic of students' evaluation of teaching. These comments can be generally categorized as (a) suggestions for item change, deletion, or insertion, (b) personal stories about some inadequacies or good points of the process or items, (c) general invectives about the process and use of SETE, and (d) questions about some of the survey questions (e.g., what is RTP?).

SERB members were gratified that so many faculty made thoughtful suggestion for new items or modification of the existing ones and recommendations of alternative bases for assessing effective instruction. These suggestions along with the survey responses guided the recommendations listed at the end of this report.

Man, faculty expressed concerns of one kind or another about the validity of student ratings and about characteristics of both the process and the students who complete faculty ratings. Aleamoni (1980) presented a fiat of typical faculty concerns about student ratings that mirror well those listed by our faculty. These are:

1. Students cannot make consistent Judgments concerning the instructor because of their immaturity, lack of experience, and capriciousness.

2. Only colleagues with experience (usually those who are deemed to be excellent instructors) are qualified to evaluate their peers' instruction.

3. Student ratings are nothing more than a popularity contest with the warm, friendly, humorous, easy-grading instructor emerging as the winner.

4. Students are not able to make accurate judgments until they have been away from the university for several years.

5. Student rating forms are both unreliable and invalid.

6. Extraneous variables or conditions affect the ratings. Some of the common ones are: (a) the size of the class; (b) sex of the student and the instructor; (c) time of the day of the course; (d) whether the student was taking the course as a requirement or an elective; (e) whether the student was a major or nonmajor; (f) the semester the course was offered; (g) the level of the course (for freshmen, juniors, etc.); and (h) the rank of the instructor.

7. The grades students receive or expect are highly correlated with their ratings and therefore lenient grading will likely produce good ratings.

Of the seven concerns listed above only part of one (number 6) seems to be substantiated by research findings (see reference list at the end of the report). Research indicates that (a) the higher the proportion of students taking the course as a requirement, the lower the rating, and (b) as students move from lower-level courses (i.e., freshmen) to higher-level courses (i.e., graduates) their ratings become higher. As for the rest of these typical faculty concerns, the research literature is highly supportive of the significance and usefulness of student evaluations as accurate reflections of student attitudes. Several investigations suggest that the correlation between student grades ant student ratings are not necessarily artifacts of lenient grading practices.

It is, of course, impossible to report all written comments received. An effort was made to summarize comments into general content categories. We report here only those comments that seemed to be voiced by a number of faculty. About 30 faculty described personal experiences with the SETE student opinion data about keeping office hours and/or return of papers or exams. Many noted that they had never received an average rating of "outstanding" on either item even though every office hour was kept and/or papers or exams were always returned the next class meeting.

There were 43 people who indicated that items 3 ant 7 (kept office hours and available outside class) were redundant. Further, 15 commented that question 7 cannot be answered accurately given the way the item is phrased. Several others commented that 2 out of 13 content items being addressed to out of class

work with students is disproportionate in terms of importance. One suggestion was to replace the two questions with "Were you able to contact the instructor when you needed?"

Some of the most frequent (23) comments of the faculty was directed to item 4 - used a variety of teaching methods. Generally speaking these faculty felt there was an emphasis in this item upon quantity rather than quality of method. One commented that Socrates used but one teaching method ant yet is considered to have been a great teacher. Another person noted that it is better to use one method well than to use several methods poorly. Several suggested that the item might be reworded "Used teaching methods or techniques relevant to (appropriate for) the course content and structure."

Items 12 and 13 (dealing with interesting presentations ant increasing student interest) tended to be paired in faculty comments. There were 26 faculty who addressed one or both of these items. The general theme was that these questions are too subjective and are likely to be very vulnerable to systematic bias from the course type or content and specific student personal characteristics. Several commented that instructors are not obligated to be interesting, increase student interest, or to facilitate student learning.

Another pervasive theme in faculty comments was the need to take into account the nature and characteristics of the students who provide the ratings. Common suggestions dealt with how much the student had learned and how much responsibility the student had taken for learning (e . g., did the student attend the class regularly, did the student study, etc.). While these concerns about student involvement may be valid, it is unlikely that there is any way to deal with the issue.

At least a dozen faculty made specific comments about the use of student ratings for retention, tenure, and promotion decisions. The general theme was that too much weight is placed on the rating data and/ or the evaluation committees misuse the rating data (see Table 3, questions 18 and 19). Several comments are germane to this issue and include: "The most important factor is the competence of the RTP Committee members in making judgments. If the group is sophisticated and thorough, they should be able to judge teaching performance fairly without regard to the detail of the student evaluation forms." If students assess someone's teaching as 'good' then the university shouldn't fire the person because his ratings are 'not good enough.'" "I would like us to return some sanity into evaluations, allowing us to give more weight to peer evaluations as well as evaluations by objective professionals who could be invited by the chair to sit in on lectures-in-progress."

Other more generalized negative comments include: "These regular evaluations are time-consuming and a waste of time." "It would make me feel quite uncomfortable to have an important career decision influenced by a minority (or perhaps even a sizeable plurality) of students who never should have been let out of high school." "Evaluations have always been blackmail lists, contributors of grade inflation. Students expect to be entertained - not taught: sometimes to get by - not learn. "The way SJSU orchestrates them, student evaluations are useless, invalid, irresponsible, and worthless for everything except personal, voluntary, and confidential use by the instructor involved." "...this universal form is in its own inflexibility and inattentiveness to differences of importance, practically advocating an attitude

of formal correctness while ignoring the difficulties of genuine importance. That is, it advocates the attitudes of fraud and sloth." "The common Joke about the norms" is "that everyone must be better than the mean.'" "My ratings on average are close to 'very good.' The University Committee somehow translates this into 'acceptable.'"

Several people suggested that written comments from students would be preferable to the "statistical information or should regularly augment the multiple choice rating data, especially for RTP purposes. Apparently there is little awareness of Senate Policy F 83-2 which encourages departmental use of open-ended written evaluation forms. The policy states that departments shall prepare a form which includes space for free student comment and that these comments can be retained for personal use or be included as data in performance review or periodic evaluations. SERB recommends that faculty, and especially RTP Committee members, be reminded that the summative and comparative statistical data derived from the SETE form is to be but one of several sources of evaluation of effectiveness of instruction and that departments systematically collect information from students by following Policy F 83-2.

SERB Recommendations

1. The Interpretation Guide, prepared by SERB in 1984, should be read by RTP Committees each year prior to reviewing and interpreting faculty rating data. In addition, said committees should be reminded that student opinion data need to be augmented by other systematic evidence of teaching effectiveness.

2. Departments should be encouraged to prepare and distribute at each rating cycle an open-ended form for student comment, consistent with Senate Policy F 83-2.

3. Items 4 and 12 on the SETE form should be deleted.

4. Items 3 and 7 (outside class accessibility and office hours) are redundant and only item 3 should be used.

5. The wording of items 2, 5, 6, 8, 9, 11, and 13 should be modified (see SOTE form).

6. Two new items should be included (see SOTE form).

7. The set of questions (SOTE) listed below was tested during Fall 1985 along with the SETE form in 70 classes to determine the advisability of making the recommended changes. SERB recommends that the Academic Senate adopt the new SOTE form beginning Fall, 1986.

8. A form for laboratory and activity classes should be introduced for norming purposes in Spring or Fall, 1987. SERB is currently constructing a laboratory/activity rating scale.

9. A reanalysis of normative data should be done to determine whether different norms may be necessary for lower division, upper division, and graduate classes.

10. The title of the rating form should be changed to reflect that student opinion information rather than evaluation information is elicited by the rating questions.

RECOMMENDED NEW RATING FORM

Below is a revised student opinion form which was developed by SERB in response to the results of the faculty survey on SETE. The new form (titled "Student Opinion of Teaching Effectiveness ant herein referred to as the new SOTE form) was tested in the Fall 1985 semester by intermingling the new SOTE forms with SETE forms in 70 classes which had populations of 35 or more students. A total of 910 students in these classes responded to the new SOTE and 2103 students in these classes responded to the SETE form. All instructors whose classes participated in this initial testing effort had agreed to the testing process. Selected data from this testing are presented in Tables A, B, C, D, E, F, G.

The recommended new SOTE rating form is presented in Table A. The presently used SETE form is presented in Table B. The items from the new SOTE ant the old SETE are presented in parallel format in Table C.

Tables D and E list the SETE and SOTE items means, standard deviations, and measures of skewness. Table D data are based on individual student responses. Data in Table E are based on class mean responses. Four item. on the two forms are identical - items 1, 3, 5, and 14 on SETE and items 1, 3, 4, ant 14 on SOTE. In all cases the items means are very slightly higher for the SOTE form than the SETE form. At the student level the item standard deviations are slightly smaller on the SOTE form (indicating slightly greater agreement among the students as to the "most appropriate" category for the given instructors). At the class level the SOTE and SETE standard deviations are virtually the same. Both forms produce a negatively skewed distribution (i e., relatively few very low ratings are given relative to the frequency of high ratings). The similarity of the ratings for these items suggests that the changes in the directions and five descriptive anchors for the items do not result in grossly different information from that which we are now getting. It should be noted that some additional rewording occurred between the items presented in Table A and the items to which students in the classes responded.

Data in Tables F and G are intercorrelations among items based either on the SOTE form (Table F) or the SETE form (Table G). All intercorrelation coefficients are positive and relatively high. This is desirable in that if all items are truly assessing the same broad construct (teaching effectiveness), we would expect to see a positive relationship among the various items. For the SOTE form it is obvious that items 3 ant 9 have relatively less overlap with the other items. This is similar to data from the SETE form where the parallel items are 3 and 11. Those who might be alarmed about the magnitude of the intercorrelation coefficients should be aware that these data are based on class level data rather than student level data and therefore are more reliable and should be expected to be quite high. On the whole the class level item intercorrelations from the SOTE form, which are slightly lower than those coefficients from the SETE form, indicate that the SOTE items are slightly less redundant than the SETE items.

While it is certainly possible to carry out more testing of the suggested SOTE items before adoption, the cost of so doing should be considered. The current evidence is that the SOTE items are as valid as the SETE items in terms of evaluating teaching effectiveness. However, the SOTE form was prepared in response to the opinions of the faculty and addresses most of the concerns expressed in the Spring 1985 faculty survey.

SERB recommends that the Senate adopt the SOTE form. If this is done expeditiously, materials can be purchased and necessary computer programming can be carried out so that the new form can be used effective Fall, 1986.

References

Some of the research evidence that accumulated about student ratings is presented here to provide readers with some of the bases, other than survey responses and faculty comments, used by SERB in determining recommendations listed above.

When student, peer, ant self-ratings of teaching are compared it is generally -the case that peer ratings are generous and highly related regardless of the characteristics being rated. Student ratings tend to be slightly lower than self-evaluations but there is high correspondence between self and student ratings (Doyle & Crichton, 1978).

In a large study of loudness students from a CSU campus where all faculty were rated at the end of course and again 1 year after graduation, it was found that course type (i.e., management, quantitative analysis, accounting, etc.), time the rating was done (end of course or after graduation) and their interactions were less related to magnitude of ratings than ratings were related to the individual instructors (Marsh & Overall, 1981). The mean rating on an "overall" item at the end of course was 6.61 on a 9-point scale and, for the same sample of students a year after graduation the mean rating was 6.63 (the median correlation for the-two sets of ratings was +.83). Thus, the attitudes of students may not change as much as might be expected over a several year period.

When effective teaching is measured by student achievement or learning, three rating factors appear to be important: student accomplishment, presentation clarity, and organization-planning (Frey, Leonard, & Beatty, 1975). Data based on calculus and educational psychology classes in three universities were used to reach this conclusion: at the same time the students' achievement level (e.g.. GPA or SAT Math scores) were not systematically related to the student ratings.

Cohen (1981) completed a meta-analysis of 41 separate studies on the relationship between student achievement ant ratings and found an average correlation of +.43 between achievement and overall rating of the instructor and an average correlation of +.50 between student achievement ant rating of instructors' "skill" (e.g., gave clear explanation).

Finally, most researchers agree with Marsh and Hocevar (1984) that student ratings are multi-dimensional and stable, and several factors must be included in order to have a valid and comprehensive rating scale. Thus, a single "overall" rating item is not sufficient.

Aleamoni, L. M. (1980). Students can evaluate teaching effectiveness. National Forum, 60(4), 41.

Cohen, P. A (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. Review of Educational Research, 6, 281-309.

Doyle, K. O., & Crichton, L. I. (1978). Student, peer, and self evaluations of college instructors. Journal of Educational Psychology, 70, 815-826

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. American Educational Research Journal, 435-434.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. American Educational Research Journal, 21, 341-366.

Marsh, H. W., & Overall, J.U. (1981). The relative influence of course level, course type, and instructors on students' evaluation of college teaching. American Educational Research Journal, 16, 57-70.

Original policy contains the following tables on file at the Academic Senate Office:

Table A: Student Opinion of Teaching Effectiveness

Table B: Student Evaluation of Teaching Effectiveness

Table C: SETE (old form), SOTE (proposed new form)

Table D: Item Means, Std. dev. & skewness (student based)

Table E: Item Means and Std. dev. (class based)

Table F: New form class item means intercorrelations

Table G: Class item intercorrelations old form