Date: August 6, 1997

To: Members, Senate Council

From: Terry Engelder, Chair, Special Committee on Faculty Teaching Development and Evaluation

Attached is a report entitled “An Analysis of the Penn State Student Rating of Teaching Effectiveness.” This report was prepared by Michael Dooris from the Center for Quality and Planning in response to a request from the Special Committee on Faculty Teaching Development and Evaluation of which he is a member. Michael Dooris will present his report to the Senate. The Special Committee finds that there has been much discussion among Penn State faculty concerning the effectiveness of the SRTE but there have been little analyses of this instrument. Dr. Dooris’ report presents a preliminary version of such an analysis which will serve as part of the groundwork for a future report coming from the Special Committee on Faculty Teaching Development and Evaluation.

Thank you.

SPECIAL COMMITTEE ON FACULTY TEACHING DEVELOPMENT AND EVALUATION

John Bagby
Jeremy Cohen
Michael Dooris
James Eisenstein
Diane Enerson
Terry Engelder, Chair
Elizabeth Hawthorne
Robert Heinsohn
William Kelly
John Lowe
John Moore
Judy Ozment
Patrick Terenzini
Harry West
Fern Willits
AN ANALYSIS OF THE PENN STATE
STUDENT RATING OF TEACHING EFFECTIVENESS
JUNE 1997
Michael J. Dooris
Center for Quality and Planning

Abstract

This report reviews the Penn State Student Rating of Teaching Effectiveness, especially with respect to the ratings on the quality of instructors, in the context of the research literature on student ratings. That literature is extensive, with hundreds of studies dating from the 1920s.

The SRTE appears to be a dependable instrument that probably produces consistent student ratings of individual instructor quality, when enough observations are taken. Expected grades statistically explain only about five percent of the variation in student ratings of instructor quality. The effects of other factors (such as class size and course level) on student ratings of instructor quality are even weaker. The findings on the Penn State SRTE are mostly similar to the results of other research on student ratings. Various measures of quality (quality of the course, quality of the instructor, department-selected items) on the SRTE tend to confirm one another.

A number of questions cannot be addressed on the basis of SRTE data alone. What is the effect of student gender? Instructor gender? Instructor rank? Time of day in which a course occurs? Student major? The research literature suggests that these variables typically do not have substantive effects on student ratings.

The question of validity remains. Does the locally developed SRTE in fact measure the quality of courses and instructors? This is impossible to answer on the basis of SRTE data alone, and the question still appears to be unanswered. The extensive research literature shows that professionally developed student ratings are reliable and valid indicators of the quality of instruction. A handful of such instruments have been thoroughly tested, are nationally normed, are widely used, and are available.
Introduction

Student ratings are not the only possible source of information about teaching effectiveness. They may not be the best source of that information. However, they are certainly the most widely used and visible means of teaching evaluation. Nationally, researchers have conducted hundreds of studies on the reliability and validity of student ratings. This analysis summarizes some of the findings from that research literature.

In April 1985, Penn State's University Faculty Senate mandated the use of what has come to be known as the SRTE – the locally developed Student Rating of Teaching Effectiveness. A complete psychometric evaluation of the Penn State SRTE is beyond the scope of this report. This report does provide new information, and reviews other information, about the Penn State SRTE. The emphasis of this analysis is upon the use of the SRTE to evaluate instructors.

Research on Student Rating Instruments

There have been decades of extensive research on student ratings. That research can answer many persistent questions and dispel some widely held misconceptions. For example, many faculty and administrators believe that student ratings are only popularity contests that have little to do with learning. Such a hypothesis can be readily tested – for example, by analyzing student ratings and grades in large, multiple section courses (such as introductory chemistry or physics) taught by several instructors who use common textbooks and give identical examinations. Many such studies have, in fact, been conducted. Arreola (1995) reviewed hundreds of studies dating as far back as 1924 on this topic; among his conclusions (pp. 83-89) were the following.

1. Student evaluations are not simply popularity contests. It is true that students prefer friendly, humorous instructors. Nonetheless, separate ratings of traits such as instructor warmth have not correlated well to student ratings of teaching effectiveness. Much evidence shows that students are discriminating judges of instructional effectiveness. This is especially true for professionally developed (versus homegrown) ratings.

2. There is little support for the proposition that students can only make accurate judgments after they have been away from the course and/or the university for several years. This is a difficult hypothesis to test conclusively, but studies tend to show that five to ten years after graduation, alumni rate instructors about the same as do current students.

3. Professionally developed instruments are both valid and reliable. Rigorous psychometric procedures have been applied to develop and test a handful of readily available, nationally normed, widely used student rating forms. Reliability – the extent to which a scale produces the same results on repeated trials – can be readily tested and is quite high for such instruments; it is often in the range of .8 to .9. Validity – the extent to
which ratings measure what they are intended to measure – is more difficult to test. However, students’ ratings have been positively correlated with other indicators of teacher competence, such as peer reviews and measures of student learning.

4. **Locally developed** student rating forms are often of questionable reliability and validity, according to the same body of research. Validating such instruments is a significant undertaking.

5. It is a misconception that because students prefer smaller classes, instructors of large classes receive poorer ratings. Students do prefer smaller classes. However, various studies have not shown a consistent relationship between class size and student ratings.

6. There is mixed evidence on the effects of students’ gender and instructors’ gender on student ratings. Some studies have found no differences; some studies have found female students to give lower ratings to male instructors than to female instructors; some studies have found female instructors to receive higher ratings regardless of students’ gender. No consistent pattern has emerged in the literature on this topic.

7. Research suggests that the time of day that a course is offered does not influence student ratings, but this evidence is relatively limited.

8. The bulk of the evidence suggests that students who are required to take a course may rate it more poorly than do students taking it as an elective. However, some studies have shown no difference.

9. It appears that whether students are majors or non-majors has no effect on their ratings of a particular course. However, this question has not been as deeply researched as some others discussed here.

10. There is some indication that upper-division and graduate students tend to rate instructors somewhat more favorably than do lower-division students.

11. It is a misconception that students prefer senior faculty, or junior faculty, or experienced teachers. No relationship between student ratings and faculty rank or experience has been consistently demonstrated.

12. There are conflicting indications about how well one or two global items (as opposed to more specific items) reflect instructional effectiveness. The amount of research on this issue is relatively limited, and on balance it is not conclusive.

13. The single most frequently researched issue on student ratings is this: Do **grades** that students receive in a course correlate with their ratings of the course and instructor? Arreola (1995) reviewed over 400 studies on this question. The signals are mixed. Various studies have found mildly positive, zero, and even mildly negative
correlations. The clearest interpretation from the relevant studies is this: the relationship between grades and ratings is weak.

**Single Dimensions, Multiple Dimensions**

As noted above, the research literature does not provide a clear indication about how well one or two global items, as opposed to more specific items, reflect instructional effectiveness. The interpretation here is that both approaches are defensible.

Unless teaching is viewed conceptually as a single behavior, there is theoretical justification for using multiple items to assess the quality of teaching. Also, to the extent that evaluations are to be used formatively, specific measures are necessary to identify particular strengths and behaviors upon which individual instructors can improve. On the other hand, there is evidence that global quality-of-instruction items are statistically stable and, as noted, that they can be psychometrically valid. Also, to the extent that student ratings are one source (among many) of information for personnel decisions, the use of one or two global items is desirable on the basis of practicality.

In short, a case can be made for a multidimensional approach to the evaluation of teaching. A case can also be made for the use of one or two global measures. Neither necessarily precludes the other.
Background on the Penn State SRTE

SRTE History

Faculty Senate documentation (Senate Committee on Faculty Affairs, 1984, 1985, 1989) traces the evolution of what is now the Student Rating of Teaching Effectiveness. It was in 1980 that a joint faculty-administrative committee on promotion and tenure policies made a recommendation for student evaluation of teaching, more or less along the lines of the SRTE. The process from that point forward involved rounds of Senate recommendations, legislation, discussion, comments to the Senate from Presidents Oswald and Jordan, input from external consultants, and various revisions. A review of that entire history would be tedious here, but it probably is worth highlighting the more persistent questions:

- Should each unit should select its own method of evaluation, or should there be a more uniform system across the University?
- Should the University standardize the reporting of evaluation results (a separate issue from whether the evaluations should be standardized)?
- Should and/or can an evaluation system for course or instructor improvement also serve evaluation for promotion and tenure purposes?
- How much weight should student input have relative to peer input?
- How frequently should course and/or instructor evaluations be required?
- Who should be allowed to summarize, interpret, and otherwise use the results of teaching evaluations?

At its meeting on April 30, 1985, the University Faculty Senate endorsed a Senate Committee on Faculty Affairs report that included a statement of recommended practices for the evaluation of teaching effectiveness for promotion and tenure. This mandated the use of in-class student rating services, and outlined what would become the SRTE. For example, the 1985 report recommended a combination of University-wide global items with other questions selected from a cafeteria of items, all using a seven-point scale.

That advisory and consultative report also contained the following statement (Senate Committee on Faculty Affairs, 1985, p. 5) that has, over the years, become part of Penn State's administrative folklore:

Student evaluation surveys shall belong to the unit which administers them. The faculty member shall be furnished with a copy of all survey results. Results shall be included in promotion and tenure documentation.

In a 1989 report, the Faculty Affairs committee could state that the University had created the SRTE, incorporated it into the promotion and tenure policy guidelines, was administering the SRTE, and was using the results widely. However, in that same report, the Faculty Affairs committee also noted that “this in-class survey has not lived up to its promises.” Topics of concern raised in 1989 included the effectiveness of unit input into creation of the form; consistency of standards for distribution and administration; the fair
use of the SRTE in personnel decisions; and the reporting and interpretation of SRTE results. Analysis of (and suggestions for changes to) the SRTE have continued, up to and including this report.

SRTE Structure

Penn State’s Student Rating of Teaching Effectiveness contains the following four University-wide items:

A1. Are you taking this course as an elective? (If uncertain, omit)
A2. What grade do you expect to earn in this course? A B C D F
A3. Rate the overall quality of this course.
A4. Rate the overall quality of the instructor.

The instrument also allows individual departments to select up to 15 items from a bank of 177 items. Instructors may also add 5 items. All items except for A1 and A2 use a response scale of 1 (lowest rating) to 7 (highest rating).

The SRTE data file does not contain data on actual grades received nor demographic data (such as gender, rank, or class standing) about instructors or students. The SRTE data also do not identify individual instructors or students.

SRTE Participation Rates

SRTE participation rates as a percentage of course enrollments from 1991-92 through 1995-96 averaged 70 percent at University Park, and 80 percent at other locations (Office of Undergraduate Education, October 1996). Participation rates ranged from 58 percent in fall 1992 at University Park, to 99 percent in summer 1994 at other campuses. In fall 1995, there were 178,000 responses (out of 238,000 possible), for an overall participation rate of 75 percent (72 percent at University Park, 78 percent at other locations).

SRTE Dependability

A 1993 study by Randy Deike investigated the SRTE in terms of its “dependability” (that is, the extent to which the measure yields the same results on repeated trials). The study included SRTE data for a subset of instructors who had taught two or more different courses, and were rated by five or more students, in Penn State’s College of the Liberal Arts in Spring 1993 – 128 courses with 7,236 SRTE observations. The research question concerned only the scores for the University-wide item that asks students to “Rate the overall quality of the instructor.” Deike’s research did not examine the relationship of those ratings to specific factors such as course level, section size, or gender. Instead, the study analyzed the scores’ variance across students and courses to see how consistent the scores were, given multiple observations of each instructor.
The generalizability coefficients that Deike found for given numbers of courses and students were similar to the results of other studies. For example, the generalizability coefficients of the SRTE were in the range of .45 to .51 when two courses, with five to 20 students, were included. He cited other research that found generalizability coefficients of about .40 to .53, for two courses with five to 20 students.

Although the Penn State data were limited to one semester of observations, those more extensive studies have found that global measures of instructor quality are highly stable when an adequate number of courses and students are included. Deike cited other research that estimated generalizability coefficients in the range of .62 to .84 for ratings covering five to ten courses and 20 students.

The policy implication is twofold. First, to evaluate a single instructor, SRTE quality-of-instructor data should come from at least five different courses and 20 or more students. Second, given a large enough sample, the measure is probably statistically dependable.

It is tempting to conclude that since high generalizability coefficients are possible, student rating scores reflect a single, true, underlying quality: specifically, the overall quality of each instructor. However, as Deike pointed out, his analysis did not address that issue. He stated that different measures of teaching effectiveness might or might not correlate with the SRTE. If teaching is a multi-faceted behavior, then other evaluation approaches (particularly multi-dimensional methods) might be more valid than the SRTE.
Analysis of Fall 1996 SRTE Data

For this report, Penn State's fall 1996 SRTE data were analyzed using basic descriptive statistics and measures of association and dispersion. Results of analyzing the complete file for that semester (with over 186,000 observations) are presented in Tables 1-3 and in Figures 1-4. Quality of course was, on average, rated 5.38; quality of instructor was, on average, rated 5.63 on a seven-point scale.

Quality of Instructor in Relation to Grades and Other Factors

Commonly raised questions about the SRTE involve the quality-of-instructor rating in relation to other factors (such as grades, course level, and so on).

The grades that students said they expected were relatively high, at 3.23, on average, as shown in Table 1. According to a December 1996 Senate report, the average grade point average for all baccalaureate degree students in spring 1996 was 2.91 (Senate Committee on Undergraduate Education, 1996). As a basis for comparison to that figure of 2.91, the SRTE respondents' expected grades in undergraduate courses (not shown in these tables) was 3.12. Fewer than one percent of the respondents in the SRTE file expected to fail (this is true both in total and for 0-499 courses); according to the Senate's data, about four percent of baccalaureate course grades are F's. Figure 3 illustrates the distribution of grades expected by SRTE respondents. As noted, the SRTE file does not contain data on actual grades received.

Table 2 shows that the estimated correlation coefficients (r-values) for expected grades to quality of course and quality of instructor were .30 and .23, respectively. There is no hard-and-fast rule for interpreting r-values; a reasonable interpretation here is that there is a modest, positive linear relationship between expected grades and student ratings of quality. In other words, students earning better grades tend to rate courses and instructors somewhat more positively than do students earning poorer grades.

The proportion of variation in quality-of-course and quality-of-instructor ratings statistically explained by expected grades is the square of the r-values — only about nine percent and five percent, respectively. Figure 4 plots a linear regression of instructor quality and expected grade.

Table 2 also shows convergence among the various quality indicators, with the quality-of-course and quality-of-instructor correlation coefficient of .78, and the department-selected items and quality-of-instructor correlation coefficient of .66.

Department- and Instructor-Selected Items

Departments use the 177-item test bank heavily. Almost every record on the file for fall 1996 (186,808 of 186,832 observations) had at least one department-selected item, and 42 percent (78,744) had the maximum of 15. University-wide, departments
### Table 1
**Descriptive Statistics**

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes (%)</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SRTE observations (University-wide, all levels)</td>
<td></td>
<td>186,832</td>
<td></td>
</tr>
<tr>
<td>&quot;Are you taking this course as an elective?&quot;</td>
<td>26%</td>
<td>3.23</td>
<td>.75</td>
</tr>
<tr>
<td>&quot;What grade do you expect to earn in this course?&quot;</td>
<td></td>
<td>5.38</td>
<td>1.30</td>
</tr>
<tr>
<td>&quot;Rate the overall quality of this course&quot;</td>
<td></td>
<td>5.63</td>
<td>1.38</td>
</tr>
<tr>
<td>&quot;Rate the overall quality of the instructor&quot;</td>
<td></td>
<td>5.92</td>
<td>1.28</td>
</tr>
</tbody>
</table>

**Department-selected items**

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Are you taking this course as an elective?&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;What grade do you expect to earn in this course?&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Rate the overall quality of this course&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Rate the overall quality of the instructor&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2
**Correlation Coefficients (Pearson)**

<table>
<thead>
<tr>
<th>Quality of Course</th>
<th>Quality of Instructor</th>
<th>Department Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of instructor</td>
<td>.78</td>
<td>.66</td>
</tr>
<tr>
<td>Department items</td>
<td>.56</td>
<td>.23</td>
</tr>
<tr>
<td>Expected grade</td>
<td>.30</td>
<td>-.06</td>
</tr>
<tr>
<td>Section size</td>
<td>-.06</td>
<td>-.01</td>
</tr>
</tbody>
</table>
Table 3
Quality of Instructor
by Expected Grade, Course Level, Section Size, Required vs. Elective, and Location

<table>
<thead>
<tr>
<th>By expected grade</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5.93</td>
<td>1.24</td>
</tr>
<tr>
<td>B</td>
<td>5.57</td>
<td>1.36</td>
</tr>
<tr>
<td>C</td>
<td>5.09</td>
<td>1.52</td>
</tr>
<tr>
<td>D</td>
<td>4.50</td>
<td>1.74</td>
</tr>
<tr>
<td>F</td>
<td>4.62</td>
<td>1.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>By course level</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-299</td>
<td>5.61</td>
<td>1.39</td>
</tr>
<tr>
<td>300-399</td>
<td>5.54</td>
<td>1.42</td>
</tr>
<tr>
<td>400-499</td>
<td>5.72</td>
<td>1.33</td>
</tr>
<tr>
<td>500-699</td>
<td>5.82</td>
<td>1.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>By section size</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 29</td>
<td>5.72</td>
<td>1.34</td>
</tr>
<tr>
<td>30-99</td>
<td>5.59</td>
<td>1.42</td>
</tr>
<tr>
<td>100-299</td>
<td>5.49</td>
<td>1.40</td>
</tr>
<tr>
<td>300-499</td>
<td>5.29</td>
<td>1.47</td>
</tr>
<tr>
<td>500 or more</td>
<td>5.50</td>
<td>1.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>By required versus elective</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elective</td>
<td>5.72</td>
<td>1.35</td>
</tr>
<tr>
<td>Required</td>
<td>5.59</td>
<td>1.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>By location</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Park</td>
<td>5.56</td>
<td>1.39</td>
</tr>
<tr>
<td>Non-University Park</td>
<td>5.74</td>
<td>1.36</td>
</tr>
</tbody>
</table>
Figure 1. Overall Quality of the Course

Course Quality: 1=lowest, 7=highest

Figure 2. Overall Quality of the Instructor

Instructor Quality: 1=lowest, 7=highest
Figure 3. Expected Grade

Figure 4. Instructor Quality & Expected Grade

Appendix "F"
9-9-97
clearly prefer some items over others: 89 percent of the forms used the item, "Rate the clarity of the instructor's presentations," while some of the 177 items in the test bank were never used.

As shown in Table 1, student ratings on average were slightly higher for the departmental items than for quality of instructor, and slightly higher for quality of instructor than quality of course. This study did not analyze individual responses to each of the 177 items in the test bank. The main point is that, as far as was determined, the various quality ratings were consistent with one another.

SRTE and Other Research Findings

The relationship between grades and student ratings is, as noted, basically similar to that found in other research. The extremely weak relationships of the SRTE student ratings to section size, and student ratings to course level, are consistent with the research literature as well. Data in Table 3 show higher ratings for graduate and upper-division courses than for lower-division courses. Table 3 also suggests a possible curvilinear effect of section size, as students generally preferred smallish or very large sections; however, there was only one section (with about 800 respondents) in the largest category. Penn State students taking courses as electives rated the instructors slightly more favorably than did students taking courses as requirements. Consistent with the broader body of research, none of these differences are especially large.

(Table 3 does not show $F$ and $p$ values because for very large data sets, tests of significance mean little. In this case, the $F$ values ranged from about 100 to over 2,000, and all $p$ values were less than .001. Although statistically significant, the results do not address whether the relationships are of any educational importance, nor do they indicate anything about causality.)

Campus Differences

On average, students at non-University Park locations rate the quality of instructors slightly higher (5.74 compared to 5.56) than do students at University Park. A campus-by-campus tabulation, not reproduced in this report, shows quality-of-instructor averages ranging from a low of 5.56 at University Park to a high of 6.11 at Shenango Valley. (41 percent of the SRTE records were from non-University Park locations.)

Quality of Course

This report has focused on the SRTE data for the quality-of-instructor rating. Some analysis was done on the quality-of-course data, as well. The results were generally similar. As already noted, there was a positive relationship to expected grades. There was also a mild positive relationship between quality-of-course and course level, but within the relatively narrow range of 5.33 to 5.61. There was a mild inverse
relationship with section size, but again within a relatively narrow range: 5.04 to 5.46. The other effects were essentially negligible.

Validity

The most important question about the effectiveness of the SRTE is the validity question: *does the SRTE, in fact, measure instructional effectiveness?* It is precisely this question that cannot be answered from SRTE data alone. Other measures of student learning, or other evaluations of instructional effectiveness (such as peer reviews), are necessary to reach strong conclusions about validity of any student-rating instrument. The positive correlations among the several SRTE “quality” items (quality of course, quality of instructor, and department-selected items) provide some evidence that the instrument’s items are at least consistent with one another.
References


Senate Committee on Faculty Affairs (May 1, 1984). Evaluating teaching effectiveness for promotion and tenure. University Park, PA: The Pennsylvania State University.


Senate Committee on Faculty Affairs (February 21, 1989). Revision of statement of practices for the evaluation of teaching effectiveness for promotion and tenure. University Park, PA: The Pennsylvania State University.