Developing a Valid and Reliable Student Evaluation of Teaching (SET) Instrument

Brenda S. Seevers[1], Thomas J. Dormody[2], and Dawn M. VanLeeuwen[3]
Department of Agricultural and Extension Education, New Mexico State University
Las Cruces, NM 88003

Abstract

Student evaluations of teaching results have been used as an indicator of course success and instructor performance. Information obtained has been used to make course changes and improvements as well as to document instructor performance for annual performance evaluation and promotion and tenure packets. The purpose of this study was to develop a valid and reliable scale to assess student evaluation of teaching. The first phase of this study involved conceptualizing effective teaching. Effective teaching was operationalized with a Likert scale of 27 indicators. Face and content validity were assessed by a panel of experts. In the second phase, reliability and dimensionality were assessed. The target population was students enrolled in the College of Agriculture and Home Economics at New Mexico State University during the Fall 1997 semester. Inter-item consistency of the final summated scale of 27 indicators was determined with a Cronbach's alpha reliability of .97. Generalizability Theory was used to estimate the reliability for relative assessments on instructors or classes. Reliability assessments based on class size ranged from .80 to .96.

Introduction

Student evaluation of teaching (SET) instruments are commonly used in higher education to assess quality of instruction or other aspects of a course. The data generated by SETs can be used to assist an instructor to improve instruction or a course (Worley and Casavant, 1995; Boice, 1990-91). Administrators and tenure and promotion committees often use the data to make tenure and promotion decisions. Administrators may rely on SET data for helping make annual performance and salary decisions. SET data are also used to help provide evidence of teaching excellence when faculty are nominated for teaching awards.

Given the ways that SET data are used in higher education, it is imperative that SETs be valid and reliable measures of quality teaching and course development. Many different types of SETs are used at colleges and universities. How many are rooted to a sound theoretical base in the dimensions of exemplary teaching? How many were conceptualized and operationalized so that each question can be traced back to one or more of these dimensions? How reliable are the data produced by the SETs?

As an example of a SET indicator that has no basis in research on exemplary teaching, we chose a question from a SET used widely at our home institution. The clause "Annoying mannerisms of the teacher" has three response categories - seldom exhibits, average, and often exhibits. The follow-up question "Please list annoying mannerisms if present" accompanies the indicator. Our experience with this indicator and follow-up question is that students feel compelled to come up with something they did not like about the instructor. Many answers mention personal and unchangeable qualities: such answers are seldom helpful for improving teaching and are often demoralizing to the instructor. One can only guess the origin of this indicator, but a review of the literature did not uncover "refraining from annoying mannerisms" as a dimension of exemplary teaching.

We embarked on a process to develop a valid and reliable SET for our college that is rooted in a strong theoretical base. Our review of related literature yielded three useful studies (Rosenshine and Furst, 1971, and Feldman, 1988, 1989) that looked at a number of other studies on dimensions of exemplary teaching. Rosenshine and Furst (1971) and Feldman (1988, 1989) attempted to make sense of the results of these many studies that educators can draw upon in developing SETs. These studies provided the framework for developing our SET instrument.

The following steps were followed to develop the instrument:

1. conceptualize effective teaching behaviors
2. operationalize effective teaching behaviors
3. assess face and content validity
4. assess reliability of the student evaluation of teaching
5. assess dimensionality of the student evaluation of teaching.

Methods

Population/Sample

The target population for the study was students

[1] Associate Professor
[2] Professor
[3] Assistant Professor

enrolled in the College of Agriculture and Home Economics classes at New Mexico State University during the Fall 1997 semester (N = 1596). A report available from the Academic Programs office (space utilization report) was used to select the study sample. The report, organized by department, provided the following information: course prefix and number, course title, section number, instructor's name, number of credits and student enrollment. Thirty classes were selected for this study in an attempt to obtain a representative sample. The following criteria were used in the selection of the classes: a) every academic department in the college was represented, b) lower division, upper division and graduate classes from every department were selected, c) all classes were three credits, d) all classes had a minimum enrollment of seven students, e) a variety of instructors was sought within each department - faculty selected were asked to only use the instrument for one section of one class taught, and f) proportional stratification in enrollment in lower division, upper division, and graduate courses was based on departmental enrollment.

Data Collection.

Data were collected during the Fall semester 1997. Most of the instructors for the selected classes were contacted by phone prior to the study and asked to participate. Each instructor was sent a packet that included a cover letter, enough copies of the instrument for each student enrolled in their class, directions for administration and a return envelope. Adhering to university policy, the instrument was administered and collected by someone other than the instructor during regularly scheduled class time during the last two weeks of the semester. Only 26 of the 30 selected classes actually participated in the study. The primary reason given by the four instructors not following through with the study was a lack of time. Students in the 26 classes completing the SET yielded a usable return rate of 531. Every department in the college was represented.

Conceptualizing and Operationalizing Effective Teaching

The theoretical frameworks of Feldman (1988 and 1989) and Rosenshine and Furst (1971) were used as the basis for SET instrument development. Feldman (1989) undertook a metanalysis of "46 studies having information about the relationships between student learning and evaluation of instruction along one or more specific instructional dimensions" (p. 587). Feldman identified 31 instructional categories or dimensions for classifying "specific ratings and their associations with student achievement" (p. 587) from 46 studies. In a previous study, Feldman (1988) also determined the relative importance of 22 of these instructional dimensions as perceived by students and faculty in 31 studies. He also determined the relative importance of these dimensions by their correlations with overall scores on SETs. For the purpose of conceptualizing our SET, we used a criterion of developing questions from only those dimensions ranked in the top 10 by either their correlation with student achievement (Feldman, 1989), or their correlation with overall scores on SETs (Feldman, 1988). The dimensions that met this criterion are listed in Table 1 with Feldman's ranks on each variable and the SET numbers of the corresponding question(s) operationalized for the dimension.

Another analysis of multiple studies on teaching behaviors related to student achievement was conducted by Rosenshine and Furst (1971). Their analysis of 50 studies generalized relatively strong relationships between student achievement and clarity, variability, enthusiasm, task-oriented and/or business-like behaviors and student opportunity to learn criterion material (basically how well material on the tests or assignments was covered in class). These five and another six less strongly related behaviors are listed in Table 2 with SET numbers of the corresponding questions developed for each behavior.

Strengths of the Feldman (1988 and 1989) and Rosenshine and Furst (1971) analysis are their quantitative approaches to summarizing results from other studies and the large number of studies they analyzed. Another strength of the Feldman analyses is that they were conducted on studies from higher education while Rosenshine and Furst analyzed studies from adolescent education. A strength of both the Feldman (1989) and Rosenshine and Furst (1971) analyses is that they looked for relationships between student achievement and selected teacher behaviors. Although their research suggests numerous relationships, educators are cautioned against inferring that the teaching behaviors are causing student achievement to increase.

A comparison of the Feldman (1988 and 1989) and Rosenshine and Furst (1971) frameworks shows that both have an enthusiasm dimension. Overlap also occurs between dimensions that include clarity, organization and understandableness. In total, 20 of the 27 indicators developed for our SET can be matched with the dimensions of exemplary teaching from both Feldman (1988 and 1989) and Rosenshine and Furst (1971). Dimensions unique to Feldman are stimulation of interest in the course and its subject matter and teacher's elocutionary skills. A dimension unique to Rosenshine and Furst is variability. When writing indicators, we made sure each dimension is addressed by at least one indicator. Multiple indicators were written for dimensions that have underlying subdimensions. Because some of the dimensions overlap, we felt some indicators aligned with more than one dimension.

Table 1: Feldman's Dimensions of Exemplary Teaching (1988 and 1989)

	Instructional Dimension	Importance ranking derived from correlations with student achievement	Importance ranking derived from correlations with overall evaluation	Corresponding questions(s) on SET
1.	Teacher's Preparation: Organization of the course	1	6	4,25
2.	Clarity and understandableness	2	2	5, 16, 19
3.	Perceived outcome or impact of instruction	3	3	22,26
4.	Teacher's stimulation of interest in the course and it's subject matter	4	1	8,24
5.	Teacher's encouragement of questions and discussion, and openness to opinion of others	5.5	11	6, 11, 12, 27
6.	Teacher's availability and helpfulness	5.5	16	9
7.	Teacher's elocutionary skills	7.5	10	7
8.	Clarity of course objectives and requirements	7.5	7	18,23
9.	Teacher's knowledge of the subject	9	9	2
10.	Teacher's sensitivity to, and concern with, class level and progress	10	5	9, 13, 19, 21
11.	Teacher's enthusiasm (for subject and for teaching)	11	8	10, 17
12.	Intellectual challenge and encouragement of independent thought (by teacher & course)	13	4	14, 15, 20

Table 2: Rosenshine and Furst's Effective Teaching Behaviors (1971)

	Teaching Behavior	Corresponding question(s) on SET
1.	Clarity	4, 5, 16, 18, 19, 23, 25
2.	Variability	3, 19, 20
3.	Enthusiasm	10, 17
4.	Task-oriented/businesslike behavior	1, 13, 14, 21
5.	Student opportunity to learn criterion material	1, 9, 13, 18, 26
6.	Use of student ideas and general indirectness	6, 11, 12
7.	Criticism (less is better)	11, 16
8.	Use of structuring comments	16
9.	Types of questions	15, 20
10.	Probing	15
11.	Level of difficulty of instruction	14

Face and Content Validity

Although the 27 indicators of effective teaching in our SET can be traced back to the dimensions of exemplary teaching from the Feldman (1988 and 1989) and Rosenshine and Furst (1971) studies, the instrument was further assessed for face and content validity by a panel of six experts. The panel was comprised of faculty in the College of Agriculture and Home Economics who are teacher educators or who have been recognized as outstanding teachers. Feedback from panel members resulted in rewording some items to increase clarity. The goal of administering the instrument to students was to assess reliability and dimensionality of the instrument.

Final Set Instrument

The final instrument (Appendix 1) consisted of 27 Likert items where the scale consisted of 5 = strongly agree, 4 = agree, 3 = undecided, 2 = disagree, and 1 = strongly disagree. The instrument has two subscales. Items 1 - 21 measure instructor behaviors and items 22-27 evaluate the course. Data can be summated and analyzed by subscale or for the whole instrument or by individual indicators. Additionally, we chose to add three open ended questions to the final instrument to allow for student comments.

Results and Discussion

Reliability

Inter-item consistency of the final summated scale of 27 indicators was determined with a Cronbach's alpha reliability of .97. Generalizability Theory was used to estimate the reliability for relative assessments on instructors or classes (Brennan, 1975; Kane and Brennan, 1977). Generalizability Theory requires reporting the variance components, then interpreting them in terms of their impact on the reliability of measurement. For these data, the variance components are 0.16028 for classes (or instructors), 0.02569 for items, 0.25784 for students nested within classes, 0.03511 for classes by item interaction, and 0.31438 for error. Since a score for an instructor involves averaging across both items and students within the class, variance components for students nested within class, class by item interaction, and error all contribute to the error variability. Thus for the smallest class size of seven the reliability for relative assessment of instructors is 0.80 while the largest class size of 47 has a reliability of 0.96. The average class size was 23 students and for a class of this size the reliability is estimated to be 0.92.

Dimensionality

Although we did not conduct the study to

determine if the Feldman (1988 and 1989) and Rosenshine and Furst (1971) dimensions of exemplary teaching really exist independently in the minds of students, we did want to determine if our set of 27 indicators had multiple dimensions. A factor analysis was conducted on the data. Principle factor extraction (using SAS PROC FACTOR Version 6.12) yielded evidence of a single factor with the possibility of up to five others. The first six factors account for 95.2% of the variability. The first factor alone had the largest eigenvalue of the correlation matrix with a value of 21.42 and accounted for 79.3% of the total variability. The second eigenvalue was 1.62, accounting for 6.0% of the generalized variance while the third factor accounted for 3.6%, the fourth 3.3%, the fifth 1.6%, and the sixth 1.4%.

A six-factor solution was used to check for factor interpretability, as the single dominant eigenvalue suggested any interpretable factors may be correlated. An orthogonal (varimax) prerotation was used with a promax (oblique) rotation to produce the final factor solution. The factor loading produced was incoherent; no reasonable interpretations of the factor loading were found. We concluded that a single primary or overall factor existed in the SET with this population.

Summary

1. Twelve of Feldman's (1988 and 1989) instructional characteristics indicative to good teaching and effective instruction and Rosenshine and Fursts' (1971) eleven teacher behaviors related to student achievement were used as the theoretical framework for developing a SET.

2. The instrument developed is a valid and reliable student evaluation of teaching. Because the instrument is a measure of effective teaching behaviors and is not subject matter specific, it can be used in almost any higher education classroom setting.

3. The instrument should continue to be assessed for reliability and dimensionality in other colleges and at other institutions. Questions should not be grouped into dimensions of exemplary teaching unless further research supports such groupings.

4. The scale is designed to be used as an evaluation tool to assess effective teaching. Data can be used both formatively and summatively.

5. Data can be reported on each item, summated for the two subscales or summated for the whole instrument. The instrument is well suited for an electronic scanning format.

Literature Cited

Brennan, R. L. 1975. The calculation of reliability from a split-plot factorial design. Educational and Psychological Measurement 35 (4): 779-788.

Boice, R. 1990-91. Considering common misbeliefs about student evaluations of teaching. Teaching Excellence 2(2) Center for Educational Development, Milton Hall, room 50, New Mexico State University.

Feldman, K.A. 1988. Effective college teaching, from the students' and faculty's view: Matched or mismatched priorities. Research in Higher Education 28(4): 291-344.

Feldman, K.A. 1989. The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education 30 (6); 583-632.

Kane, M.T. and R.L. Brennan, 1977. The generalizability of class means. Review of Educational Research. 47 (1): 267-292.

Rosenshine, B. and M. Furst. 1971. Research on teacher performance criteria. In: B.O. Smith (ed.). Research in teaching education. pp. 27-72. Englewood Cliffs, NJ: Prentice Hall

Worley, T. and K. Casavant. 1995. Student evaluations of teaching: A tool for directing and measuring course improvement efforts. NACTA Jour. 39 (1): 37-38.

Appendix

STUDENT EVALUATION OF INSTRUCTION

COURSE _____ **SECTION** _____ **INSTRUCTOR** _____

Directions: Please use the five point scale shown below to rate the instructor and course. Circle the number that best indicates your level of agreement to the statements below.

5 = Strongly Agree 4 = Agree 3 = Undecided 2 = Disagree 1 = Strongly Disagree

THE INSTRUCTOR....	**SA**	**A**	**U**	**D**	**SD**
1. Devoted appropriate amounts of class time to each topic.	5	4	3	2	1
2. Knew the subject well.	5	4	3	2	1
3. Used a variety of teaching methods.	5	4	3	2	1
4. Was well prepared for class.	5	4	3	2	1
5. Had the ability to get ideas across in an effective manner.	5	4	3	2	1
6. Encourage students to ask ideas.	5	4	3	2	1
7. Was an effective lecturer.	5	4	3	2	1
8. Stimulated my interest in the subject matter.	5	4	3	2	1
9. Was willing to assist students outside of class time.	5	4	3	2	1
10. Demonstrated enthusiasm in class.	5	4	3	2	1
11. Showed respect to the student.	5	4	3	2	1
12. Used students' ideas.	5	4	3	2	1
13. Was sensitive to students' progress.	5	4	3	2	1
14. Challenged me to reach high standards.	5	4	3	2	1
15. Encouraged thinking by asking probing questions.	5	4	3	2	1
16. Provided constructive feedback.	5	4	3	2	1
17. Was enthusiastic toward the subject.	5	4	3	2	1
18. Clearly explained how students would be evaluated.	5	4	3	2	1
19. Used examples to simplify complex concepts.	5	4	3	2	1
20. Varied the difficulty of questions asked in class.	5	4	3	2	1
21. Was concerned with student learning.	5	4	3	2	1

THE COURSE....	**SA**	**A**	**U**	**S**	**SD**
22. Increased my knowledge of the subject area.	5	4	3	2	1
23. Objectives were clearly stated at the beginning of the course.	5	4	3	2	1
24. Was interesting to me.	5	4	3	2	1
25. Was well organized.	5	4	3	2	1
26. Assignments contributed to the achievement of course objectives.	5	4	3	2	1
27. Was sensitive to diversity.	5	4	3	2	1

What did you like best about the course?

What are the qualities of the instructor you feel were most effective?

What suggestions do you have for improving the course?