STUDENTS' EVALUATIONS OF UNIVERSITY INSTRUCTORS: THE APPLICABILITY OF AMERICAN INSTRUMENTS IN A SPANISH SETTING

HERBERT W. MARSH

The University of Sydney, Australia

JAVIER TOURÓN

Universidad de Navarra, Spain

and

BARBARA WHEELER

The University of Sydney, Australia

Abstract -Items from two American instruments designed to measure students' evaluations of teaching effectiveness were translated into Spanish and administered to a sample of Spanish university students. Most of the items were judged by the students to be appropriate, every item was chosen by at least a few as being a most important item, and all but the orkload/Difficulty items clearly differentiated between lecturers whom students indicated to be "good," "average," and "poor." A series of factor analyses clearly identified the factors that the instruments were designed to measure and that have been identified in previous research. Finally, a multitrait-multimethod analysis demonstrated that there was good agreement between factors from the two instruments which were hypothesized to measure the same components of effective teaching, and provided support for both the convergent and discriminant validity of the ratings. The findings illustrate the feasibility of evaluating effective teaching in a Spanish university and the appropriateness of the two American instruments.

Students' evaluations of teaching effectiveness are commonly collected at North American universities and colleges, and their use is widely endorsed by students, faculty, and administrators (Centra, 1979; Leventhal, Perry, Abrami, Turcotte, & Kane, 1981). The purposes of these evaluations are variously to provide (a) diagnostic feedback to faculty about the effectiveness of their teaching; (b) a measure of teaching effectiveness to be used in tenure or promotion decisions; (c) information for students to use in the selection of courses and instructors; and (d) an outcome or process description measure for research on teaching. While the first purpose is nearly universal, the next two are not. At many universities systematic student input is required before faculty can even be considered for promotion, while at others the inclusion of students' evaluations is optional. Likewise, the results of students' evaluations are made public at some universities while at others the results are considered to be confidential. The fourth purpose of students ratings their use in research on teaching, has unfortunately not been systematically examined. The use of students' evaluations, especially for tenure or promotion decisions, has not been without opposition, and in the last decade this has been one of the most frequently studied areas in American

educational research (for reviews see Aleamoni, 1981; Centra, 1979; Cohen, 1980, 1981; Costin, Greenough, & Menges, 1971; de Wolfe, 1974; Doyle, 1975; Feldman, 1976a, 1976b, 1977,1978,1979,1982; Kulik & McKeachie, 1975; Marsh, 1980a, 1982b, 1984; Murray, 1980; Overall & Marsh, 1982). In contrast to the wide use of students' evaluations in North America, they apparently have not been systematically collected in universities in other parts of the world, and there has been little attempt to test the applicability of instruments developed in the United States or the generalizability of findings from American settings in other countries. This article describes two such American instruments and reports upon an investigation of their applicability in a Spanish setting.

The Endeavor Instrument

The endeavor instrument measures seven components of effective teaching+omponents that have been identified through the use of factor analysis in different settings (Frey, Leonard, & Beatty, 1975). The seven factors are Presentative Clarity, Workload, Personal Attention, Class Discussions, Organization-Planning, Grading, and Student Accomplishment. In validating the ratings obtained from this instrument, Frey has shown that the ratings on Endeavor are correlated with student learning (Frey, 1973, 1978; Frey, Leonard, & Beatty, 1975). In these studies, as well as in similar studies described below, student ratings are collected in large multisection courses (i.e., courses in which the large group of students is divided into smaller groups or sections and all instruction is delivered separately to each section). Each section of students in the same course is taught throughout by a different lecturer, but each is taught according to a similar course outline, has similar goals and objectives and, most important, is tested with the same standardized final examination at the end of the course (for further discussion see Cohen, 1981; Marsh, 1982b, 1984; Marsh & Overall, 1980). Frey concluded that those sections of students that rate teaching to be most effective are also the sections that learn the most as measured by performance on the final examination, thus supporting the validity of ratings on the Endeavor instrument.

Frey (1978) further argued that it is important to recognize the multidimensionality of evaluations of effective teaching. In an examination of the relationships between students' evaluations and a variety of other variables, he demonstrated that the size, and even the direction, of the correlations varies with the particular component of effective teaching considered. The failure to recognize this multidimensionality is an important weakness in much of the American research.

The SEEQ instrument

SEEQ (Students' Evaluations of Educational Quality) and the research that led to its development have recently been summarized (Marsh, 1982b, 1983, 1984). Numerous factor analyses have identified the nine SEEQ factors in responses from different poipulations of students (e.g., Marsh, 1982b, 1982c, 1983), and also in lecturer self-evaluations of their own teaching effectiveness when they were asked to complete the same instrument as their students (Marsh, 1982~; Marsh & Hocevar, 1983). The nine SEEQ factors are Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty.

Marsh (1982c, 1984), like Frey, argued that students' evaluations, like the effective teaching they are designed to reflect, should be multidimensional (e.g., a lecturer can be well organized and still lack enthusiasm). He supported this common-sense assertion with empirical results and also demonstrated that the failure to recognized this multidimensionality has led to confusion and misinterpretation in student- evaluation research.

The reliability of responses to SEEQ, based upon correlations among items designed to measure the same factor and correlations among responses by students in the same course, is consistently high (Marsh, 1982b). To test the long-term stability of responses to SEEQ, students from 100 classes were asked to re-evaluate teaching effectiveness several years after their graduation from their university program, and their retrospective evaluations correlated 0.83 with those the same students had given at the end of each class (Overall & Marsh, 1980). Ratings on SEEQ have successfully been validated against the ratings of former students (Marsh, 1977), student achievement as measured by an objective examination in multisection courses (Marsh & Overall, 1980; Marsh, Fleiner, & Thomas, 1975), lecturers' evaluations of their own teaching effectiveness (Marsh, Overall, & Kesler, 1979; Marsh, 1982c), and affective course consequences such as applications of course materials and plans to pursue the subject further (Marsh & Overall, 1980). None of a set of 16 potential sources of bias (e.g., class size, expected grade, prior subject interest) could account for more than 5% of the variance in SEEQ ratings (Marsh, 1980b, 1983), and many of the relationships were inconsistent with a simple bias explanation (e.g., harder, more difficult courses were evaluated more favorably). SEEQ ratings are primarily a function of characteristics of the person who teaches a course, rather than of the particular course which he or she teaches (Marsh, 1981b, 1982a; Marsh & Overall, 1981). Finally, feedback from SEEQ, particularly when coupled with a candid discussion with an external consultant, led to improved ratings and better student learning (Overall & Marsh, 1979).

The Present Study

The purposes of the present study are to test the applicability of the SEEQ and Endeavor instruments in a Spanish setting, and to replicate the results of a similar study conducted in an Australian setting where the factors which these surveys are designed to measure were empirically demonstrated and judged by Australian students to be appropriate and important (Marsh, 1981a).* In the present study, items from both the SEEQ and Endeavor instruments were translated into Spanish and administered to a sample of Spanish university students. Students were asked to select a representative "good," "average," and "poor" lecturer, to evaluate each with the same set of items, to indicate inappropriate items, and to select the most important items. These criteria, in addition to factor analyses of the ratings, were used to test the applicability of these American instruments in a Spanish setting.

Method

Sample and Procedures

The evaluation instrument was administered to a total of 209 students currently enrolled in the Universidad de Navarra. The subjects were second-, third-, and fourth-year university students, primarily between 19 and 21 years of age, who were in the process of completing degrees in education, architecture, or law. Instructions about the study were read to the students, who volunteered to participate, after which they completed the instrument. Students were asked not to put their name on the instrument, and the confidentiality of their responses was guaranteed. There was no time limit for completing the instrument, but most students had completed it within about 30 minutes. All instruments were administered by the second author of the study.

Each evaluation instrument contained a cover page with instructions and items calling for demographic information and requested that students select a "good," and "average," and a "poor" lecturer from their university experience. They were asked to try to limit their choices to lecturers who were in charge of an instructional sequence which lasted at least one term, and who taught courses that employed a lecture or discussion format. Students were then asked to fill out three separate questionnaires, one each for the "good," "average," and "poor" lecturers. The items, in paraphrased form, and the components of effective teaching which they are hypothesized to measure, appear in Table 2. Students responded to each item on a nine-point response scale which varied from "l-very poor, very low, or almost never" to "9-very good, very high, or almost always." An additional "not appropriate" response was provided for items not relevant to the course being evaluated (responses to items left blank were also counted as "not appropriate"). After completing the ratings for a given' lecturer, students were asked to select up to five questions that they felt were "most important in describing either positive or negative aspects of the overall learning experience in this instructional sequence."

Statistical Analysis

Each item was initially tested in terms of: (a) its ability to discriminate among the good, average, and poor instructors; (b) its appropriateness (i.e., the lack of "not appropriate" responses); and (c) its importance (i.e., the number of "most important" nominations). Items were categorized as representing 10 dimensions on an a priori basis (support for these dimensions was found in the Australian study described by Marsh, 1981a), and a factor analysis of responses to all items was used to testthe ability of the responses to differentiate among these hypothesized components of teaching effectiveness. Separate factor analyses were also performed on responses to items from SEEQ and the Endeavor instruments, and factor scores derived from these analyses were used to determine the relationship between SEEQ and Endeavor factors.

All the statistical analyses were conducted with the commercially available SPSS statistical package (Hull & Nie, 1981). A separate oneway analysis of variance (ANOVA) was used to test the ability of each item to discrimate between "good," "average," and "poor" teachers, and differences between the three groups were then broken into linear and nonlinear components (Nie, Hull, Jenkins, Steinbrenner & Bent, 1976, p. 425). The factor analyses were performed with interated communality estimates, a Kaiser nomalization, and an oblique rotation, also using the SPSS procedure.

For purposes of this study, blank and "not appropriate" responses were considered to be missing values. Each of the factor analyses was performed on correlation matrices constructed with "pair-wise deletion" for missing data. Factor scores derived from these analyses were used to represent the SEEQ and Endeavor factors, and consisted of weighted averages of responses to each item. Factor scores, based upon weighted averages of nonmissing values, were computed for each student so long as at least 75% of the responses were completed. Factor scores were derived and missing data were handled in the ways described by Nie et al. (1976, p. 496).

Results

Evaluation of individual Items

Preliminary inspection of the content of the SEEQ and Endeavor instruments revealed considerable overlap in the dimensions defined by each. Five SEEQ factors (Learning/Value, Group Interactions, Individual Rapport, Examinations/Grading, and Workload/Difficulty) appear to correspond closely to five Endeavor factors (Student Accomplishments, Class Discussion, Personal Attention, Grading, and Workload) as shown in Table 1.

Table 1
Pairs of corresponding factors in SEEQ and Endeavor

SEEQ factors	Endeavor factors
1. Learning/Value	1. Student Accomplishments
2. Group Interaction	2. Class Discussion
3. Individual Rapport	3. Personal Attention
4. Examinations/Grading	4. Grading
5. Workload/Difficulty	5. Workload
6. Organization/Clarity	6. Presentation Clarity 7. Organization/Planning

A sixth SEEQ factor, Organization/Clarity, seems have been divided into two factors for the Endeavor instrument (Presentation Clarity and Organization/Planning). Three SEEQ factors, Instructor Enthusiasm, Breadth of Coverage, and Assignments/Readings, do not appear to correspond to any of the Endeavor factors. On the basis of this preliminary inspection and results of the Australian study, 32 of the 34 SEEQ items (MI-M32), the 21 Endeavor items (FI-F21), and seven additional items (AI-A7) were each classified into one of 10 dimensions (see Table 2). Two other SEEQ items, not specifically designed to measure a particular factor, are overall ratings of the instructor (M31) and the course (M30).

With the exception of Workload/Difficulty items, all items differentiated significantly ($p<0.001$; see Table 2) among the "good," "average," and "poor" instructors in the predicted direction (i.e., "average" instructors were evaluated significantly lower than "good" instructors and significantly higher than "poor" instructors).

Table 2
Hypothesized factors, individual items and their characteristics discrimination among lecturers

Factors and items	Mean responses for lecturers chosen as: Good	Average	Poor	Variance explained by: Linear component	Nonlinear component	Number of (not appropriate) responses	Number of (most important) nominations
Learning/Value							
M1 Course challenging and stimulating	7.5	5.7	2.3	61.4	1.7	16	110
M2 Learned something valuable	7.5	6.1	3.6	39.3	1.1	16	89
M3 Increase subject interest	7.2	5.6	2.7	49.3	1.2	12	70
M4 Learned and understood subject matter	7.5	6.4	4.5	33.1	0.9	7	23
F19 Understood the advanced material	7.4	6.2	3.9	35.5	0.1	14	19
F20 Ability to analyze issues	7.4	5.9	3.4	46.7	1.0	17	51
F21 Increased knowledge and competence	7.5	6.2	3.7	43.7	1.4	12	53
Instructor Enthusiasm							
M5 Enthusiastic about teaching	8.1	6.6	4.2	43.7	8.9	5	163
M6 Dynamic and energetic	7.4	5.7	2.7	53.6	0.3	17	43
M7 Enhanced presentation with humor	7.2	5.2	3.0	39.7	0.0	22	72
M8 Teaching style held your interest	7.9	5.4	2.0	68.7	0.4	15	121
A1 Seems to enjoy teaching	8.2	6.4	3.9	48.4	0.5	9	82
Presentation Clarity (Organization)							
M9 Lecturer explanations clear	7.9	6.3	2.6	62.9	3.6	4	184
M10 Materials well explained and prepared	7.9	6.1	2.4	67.2	2.7	12	92
M12 Lecturers facilitated taking notes	7.4	5.7	2.4	52.7	0.2	21	64
F1 Presentations clarified materials	7.6	6.1	2.7	61.0	2.6	43	50
F2 Presented clearly and summarized	7.8	5.9	2.6	64.8	1.5	10	86
F3 Made good use of examples	7.7	6.4	3.8	45.4	1.2	12	58
Planning/Objectives (Organization)							
M11 Course objectives stated and pursued	7.6	6.2	3.9	37.8	0.5	31	42
F13 Presentations planned in advance	8.2	6.8	4.1	45.1	1.5	12	87
F14 Provided detailed course schedule	7.0	6.0	3.9	25.9	1.3	22	18
F15 Activities orderly scheduled	7.2	5.8	3.6	39.0	0.6	18	30
A2 Time distributed over topics	7.1	5.7	3.9	39.0	0.9	21	47
A3 Announced goals and/or criteria	7.4	5.9	3.4	43.9	0.9	17	44
Group Interaction/Discussion							
M13 Encouraged class discussion	6.3	5.8	3.6	20.3	2.5	20	48
M14 Students shared knowledge/ideas	6.6	5.4	3.3	30.8	0.6	47	28
M15 Encouraged questions and gave answers	7.0	5.5	3.1	42.4	0.7	12	19
M16 Encouraged expression of ideas	6.6	5.2	3.4	29.0	0.1	31	8
F10 Class discussion was welcome	8.0	6.5	4.5	38.7	0.4	6	34
F11 Students encouraged to participate	6.9	5.6	3.5	30.1	0.6	21	35
F12 Encouraged students to express ideas	6.5	5.2	3.2	32.3	0.5	34	20

Individual Rapport/Personal Attention							
M17 Friendly towards individual students	7.4	6.4	4.1	31.0	1.5	7	86
M18 Welcomed students seeking help/advice	7.1	5.4	3.0	41.1	0.5	49	36
M19 Interested in individual students	6.4	4.8	2.6	38.6	0.3	36	52
M20 Accessible to individual students	6.6	5.6	4.1	16.2	0.3	35	31
F7 Listened and was willing to help	7.2	5.7	3.6	33.1	0.3	36	42
F8 Able to get personal attention	7.7	6.0	4.2	34.9	0.0	97	36
F9 Concerned about student difficulties	6.8	5.0	3.2	35.0	0.0	22	64
Breadth of Coverage							
M21 Contrasted various implications	7.0	6.0	4.0	28.9	1.1	45	41
M22 Gave background of ideas/concepts	7.0	6.0	3.8	31.8	2.0	33	15
M23 Gave different points of view	6.8	5.7	3.3	35.0	1.4	60	28
M24 Discussed current developments	6.9	5.9	3.9	29.7	1.3	115	22
Grading/Examinations							
M25 Examination feedback valuable	6.5	5.2	2.8	36.3	1.1	89	28
M26 Evaluation methods fair/appropriate	6.9	5.8	3.6	31.9	1.1	38	76
M27 Tested course content as emphasized	6.7	5.6	3.5	30.1	1.0	57	28
F16 Grading was fair and impartial	7.0	5.8	3.7	31.6	0.7	54	87
F17 Grading reflected student performance	6.5	5.5	3.4	29.0	1.5	54	39
F18 Grading indicative of accomplishments	6.7	5.5	3.4	33.3	1.1	54	20
Assignments/Readings							
M28 Readings/texts were valuable	6.8	5.8	4.3	19.5	0.4	132	16
M29 They contributed to understanding	6.8	5.7	3.6	33.9	1.2	99	14
A4 They encouraged further exploration	6.4	5.4	3.3	27.0	1.1	98	32
A5 They were integrated into course	7.3	5.8	4.5	28.0	0.0	121	17
A6 Appropriate in length and difficulty	6.9	5.7	4.1	27.7	0.2	134	11
A7 They were related to class work	7.4	6.1	4.3	30.2	0.2	125	12
Workload/Difficulty							
M32 Course difficulty (easy-hard)	6.0	5.4	5.4	1.4	0.8	0	19
M33 Course workload (light-heavy)	5.9	5.5	0.5	0.2	0	19	
M34 Course pace (slow-fast)	6.0	5.3	4.9	0.2	0.2	1	21
F4 Students had to work hard	7.4	6.3	5.6	11.0	0.2	2	33
F5 Course required a lot of work	6.9	6.1	5.6	6.9	0.2	5	19
F6 Course workload was heavy	6.1	5.8	5.5	1.4	0.0	7	33
Overall Rating Items							
M31 Overall Instructor Rating	8.2	5.9	2.5	81.8	1.0	12	49
M30 Overall Course Rating	8.1	5.9	2.3	79.6	1.8	13	28

Note: The Endeavor factors of Presentation Clarity and Planning/Objectives are represented by a single factor called Organization in SEEQ. The Overall Rating Items on SEEQ were not specifically designed to measure a particular factor.

Furthermore, nearly all of the differences among the three groups are explained by the linear component (i.e., the "variance explained" by the linear component is generally 3C-60 times as large as the remaining variance, which is explained by nonlinear components). The differences are particularly large for the Instructor Enthusiasm, Presentation Clarity, and Learning/Value/Accomplishment dimensions. Workload/Difficulty items do not differentiate among the three groups as clearly. Nevertheless, "good" instructors tend to teach courses which are judged to be more difficult and require more work.

Students were specifically asked to indicate items that were inappropriate. Nine of the 62 items were judged to be inappropriate by more than 10% of the students (see Table 2). These included all six of the Assignments/Reading items, and items related to feedback from examinations, ability to get individual attention, and discussion of current developments. The number of inappropriate responses to the Assignments/Reading items suggests that outside assignments are not necessarily a part of courses in this Spanish university. Nevertheless, a majority of the items were judged to be appropriate by 95% or more of the students, and indicate that most of the items are generally appropriate in this Spanish setting.

Students selected as many as five items that they felt were most important in describing positive or negative aspects of the overall learning experience. Each of the 62 items, even those seen as inappropriate by 10% or more of the students, received at least eight nominations, and at least one item from each of the 10 categories received 32 or more nominations (see Table 2). Four items received over 100 nominations: course challenging and stimulating (MI), lecturer enthusiastic about teaching (M5), teaching style held your interest (M8), and lecturer explanations were clear (M9). Items in the Learning/Value/Accomplishments, Instructor Enthusiasm, and Presentation Clarity categories were nominated most frequently. While some of the items and some of the dimensions were seen as most important, the nominations were spread widely over the entire set of items. This suggests that each of the dimensions measures a potentially important component of effective teaching.

Factor Analyses of the Combined Set of Items

On the basis of an a priori examination of the content of each item and the results of the Australian study (Marsh, 1981a), it was hypothesized that the 62 items would measure 10 components of teaching effectiveness. This hypothesis was empirically tested through the application of factor analysis. The results (see Table 3) demonstrate that each of the 10 factors is identified with remarkable clarity. With the exception of two items (F3 and F13), each item loads substantially on the factor it was designed to measure (target loadings) and less substantially on the other nine factors (nontarget loadings). A majority of the target loadings are greater than 0.55, and only three are less than 0.30. A majority of the nontarget loadings are less than 0.1, 95% are less than 0.2, and less than 1% are greater than 0.3. The overall ratings of the instructor and course are not specifically designed to measure a particular dimension, but results from North American studies indicate that they load most highly on the Instructor Enthusiasm and Learning/ Value dimensions, respectively (Marsh, 1983). In the Australian study, however, the Overall Instructor Rating loaded most highly on the Presentation Clarity dimension, though the Overall Course rating still loaded most substantially

on the Learning/Value dimension. In the Spanish setting, both the Overall Course and Overall Instructor ratings load most highly on the Clarity dimension, and to a lesser extent on the Instructor Enthusiasm dimension.

Table 3
Factor analysis of responses to all items (n = 627 sets of ratings)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Learning/Value											
M1	Course challenging and stimulating	**34**	27	25	−03	06	03	14	10	08	07
M2	Learned something valuable	**63**	04	09	01	07	07	12	00	05	02
M3	Increased subject interest	**47**	22	06	06	05	09	15	03	08	−01
M4	Learned and understood subject-matter	**70**	01	02	15	00	04	−04	10	08	−08
F19	Understood the advanced material	**53**	09	15	17	04	05	−04	03	07	−13
F20	Ability to analyze issues	**58**	10	03	07	05	00	15	05	04	10
F21	Increased knowledge and competence	**66**	03	04	01	−01	14	12	10	07	10
Instructor Enthusiasm											
M5	Enthusiastic about teaching	05	**56**	08	−01	07	13	20	−04	05	01
M6	Dynamic and energetic	14	**31**	20	06	07	02	13	12	16	14
M7	Enhanced presentation with humor	07	**49**	03	04	18	09	08	05	03	07
M8	Teaching style held your interest	22	**45**	22	06	07	−01	00	10	01	04
A1	Seems to enjoy teaching	07	**61**	00	05	08	15	19	01	04	04
Presentation Clarity (Organization)											
M9	Lecturer explanations clear	16	24	**50**	19	02	−02	04	02	06	−01
M10	Materials well explained and prepared	13	22	**50**	21	−02	00	11	04	08	04
M12	Lectures facilitated taking notes	02	25	**35**	27	04	05	01	12	09	−09
F1	Presentations clarified materials	15	19	**46**	15	−01	04	13	07	05	03
F2	Presented clearly and summarized	09	29	**41**	21	00	00	06	11	11	−02
F3	Made good use of examples	07	37	**11**	02	07	02	19	20	07	04
Planning/Objectives (Organization)											
M11	Course objectives stated and pursued	14	−07	22	**58**	02	−02	09	08	04	01
F13	Presentations planned in advance	06	16	29	**13**	−11	12	30	09	−02	09
F14	Provided detailed course schedule	−02	−11	08	**58**	16	02	09	10	03	14
F15	Activities orderly scheduled	08	−04	25	**44**	−01	05	18	14	01	07
A2	Time distributed over topics	14	15	−08	**57**	03	06	−05	10	16	01
A3	Announced goals and/or criteria	08	25	−09	**50**	12	04	08	10	13	−01
Group Interaction/Discussion											
M13	Encouraged class discussion	07	06	02	01	**70**	06	04	02	05	00
M14	Students shared knowledge/ideas	03	06	05	02	**66**	09	15	07	08	−01
M15	Encouraged questions and gave answers	01	11	11	16	**51**	12	15	07	08	00
M16	Encouraged expression of ideas	00	08	−01	09	**69**	16	09	05	06	−02
F10	Class discussion was welcome	04	04	21	03	**44**	23	15	08	03	00
F11	Students encouraged to participate	06	08	−03	10	**73**	09	06	03	08	03
F12	Encouraged students to express ideas	03	07	−02	15	**66**	14	13	06	02	−01

Individual Rapport/Personal Attention											
M17	Friendly towards individual students	00	06	18	−06	29	**39**	09	12	08	−06
M18	Welcomed students seeking help/advice	02	02	24	04	21	**50**	−01	17	09	00
M19	Interested in individual students	03	14	07	06	17	**52**	03	15	06	00
M20	Accessible to individual students	11	01	−04	07	−01	**65**	08	−04	11	05
F7	Listened and was willing to help	01	−03	23	−02	28	**52**	05	10	04	04
F8	Able to get personal attention	06	05	12	−01	09	**60**	09	11	06	04
F9	Concerned about student difficulties	−02	10	08	07	26	**51**	−01	13	10	04
Breadth of Coverage											
M21	Contrasted various implications	00	01	11	−02	12	01	**67**	03	04	01
M22	Gave background of ideas/concepts	05	15	07	08	10	−01	**49**	09	04	00
M23	Gave different points of view	−04	03	08	08	10	10	**57**	04	10	−03
M24	Discussed current developments	12	12	−14	09	−04	16	**57**	03	13	02
Grading/Examinations											
M25	Examination feedback valuable	03	−06	19	10	04	20	18	**25**	19	01
M26	Evaluation methods fair/appropriate	01	04	09	05	06	04	04	**77**	01	−02
M27	Tested course content as emphasized	10	07	−09	18	−01	13	18	**38**	15	−07
F16	Grading was fair and impartial	01	08	00	07	−02	11	03	**78**	03	01
F17	Grading reflected student performance	05	−02	03	09	03	01	03	**82**	06	−02
F18	Grading indicative of accomplishments	08	01	04	09	04	01	05	**77**	02	00
Assignments/Readings											
M28	Readings/texts were valuable	07	−05	12	−11	07	−11	25	22	**35**	01
M29	They contributed to understanding	24	−01	04	−06	13	−02	14	16	**48**	09
A4	They encouraged further exploration	30	−02	08	−12	14	−02	13	07	**43**	01
A5	They were integrated into course	−05	05	03	08	04	03	03	00	**84**	00
A6	Appropriate in length and difficulty	08	08	−08	18	−03	20	−01	00	**58**	10
A7	They were related to class work	03	06	03	12	−01	14	11	01	**60**	05
Workload/Difficulty											
M32	Course difficulty (easy-hard)	−02	07	−05	02	04	−07	−01	−06	02	**83**
M33	Course workload (light-heavy)	−06	04	−06	05	−05	02	−01	−05	−02	**86**
M34	Course place (slow-fast)	09	10	00	10	−06	−04	04	07	−04	**49**
F4	Students had to work hard	03	01	12	−02	03	01	00	04	10	**83**
F5	Course required a lot of work	00	−08	10	00	04	02	02	03	09	**83**
F6	Course workload was heavy	−02	−04	00	−02	02	05	02	02	03	**83**
Overall Rating Items											
M31	Overall Instructor Rating	17	30	40	12	05	09	08	13	10	08
M30	Overall Course Rating	12	29	40	17	02	08	07	10	12	05

Note: The factor loadings in bold type, the target loadings, are for items designed to measure the factor.

Analyses of Responses to SEEQ and Endeavor Instruments

Two separate factor analyses, that is, factor analyses of responses to the 34 SEEQ items (see Table 4) and to the 21 Endeavor items (see Table 5), each clearly identify the factors which those instruments were designed to measure. For both analyses every target loading (those in bold type in tables 4 and 5) is at least 0.3, and a majority are greater than 0.5. Few nontarget loadings in either analysis are as large as 0.3, and most are less than 0.1. Factor scores used in the analysis described below were based upon these factor analyses.

Correlations between the nine SEEQ and seven Endeavor factors (see results of Spanish study in Table 6) are presented in a form somewhat analogous to a multitrait-multimethod (MTMM) matrix, where the dimensions of teaching effectiveness are the multiple traits and the two different instruments correspond to the multiple methods. Convergent validity refers to the correlations between SEEQ and Endeavor dimensions that are hypothesized to measure the same construct, while discriminant validity refers to the distinctiveness of the different dimensions and provides a test of the multidimensionality of the ratings. Typical MTMM analyses (see Marsh & Hocevar, 1983) would require that the same dimensions be assessed by the two instruments. But, with minor modifications, the criteria developed by Campbell and Fiske (1959) can be applied to test for convergent and discriminant validity in these data.

1. Convergent validities, correlations between SEEQ and Endeavor factors that are hypothesized to match (correlations in bold type in Table 6), should be substantial. Here the convergent validities vary between 0.71 and 0.93 and clearly satisfy this criterion.

2. One criterion of discriminant validity is that correlations between these matching factors should be higher than the correlations between nonmatchng SEEQ and Endeavor factors in the same row or column of the rectangular submatrix. The application of this criterion requires that each of the seven convergent validities be compared with 14 other correlations. This test is met for 97 of the 98 comparisions, and clearly satisfies the second criterion.

3. Another criterion of discriminant validity is that correlations between these matching factors should be higher than correlations in the same row or column of the triangular submatrices. The application of this criterion requires that each convergent validity be compared with eight correlations involving other SEEQ factors and six correlations involving other Endeavor factors. This test is met for all 98 of these comparisons, and clearly satisfies the third criterion.

4. The pattern of correlations among SEEQ factors should be similar to the pattern of correlations among Endeavor factors (e.g., because the two SEEQ factors of Group Interaction and Individual Rapport are highly correlated, then so should be the two Endeavor factors of Class Discussions and Personal Attention). A visual inspection of the correlations in Table 6 demonstrates the similarity in the patterns of correlations.

Table 4
Factor analysis of SEEQ items (n = 627 pairs of ratings)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Learning/Value										
M1	Course challenging and stimulating	**33**	42	05	06	08	09	08	13	03
M2	Learned something valuable	**69**	08	00	09	08	05	−05	10	03
M3	Increase subject interest	**49**	25	01	06	13	14	01	02	−02
M4	Learned and understood subject-matter	**68**	−01	15	03	07	00	08	−04	−08
Instructor Enthusiasm										
M5	Enthusiastic about teaching	06	**45**	05	13	14	18	−08	12	02
M6	Dynamic and energetic	11	**47**	02	14	03	12	27	00	10
M7	Enhanced presentation with humor	03	**42**	04	18	19	14	−01	−04	06
M8	Teaching style held your interest	23	**54**	14	09	05	07	07	00	02
Organization/Clarity										
M9	Lecturer explanations clear	19	36	**48**	01	02	12	−05	07	00
M10	Materials well explained and prepared	17	35	**44**	01	01	14	06	07	06
M11	Course objectives stated and pursued	27	−07	**45**	08	00	09	20	−09	09
M12	Lectures facilitated taking notes	07	30	**40**	08	08	10	12	−06	−09
Group Interaction/Discussion										
M13	Encouraged class discussion	05	02	−01	**79**	03	−03	−06	09	00
M14	Students shared known knowledge/ideas	01	02	−02	**80**	00	13	07	−01	00
M15	Encouraged questions and gave answers	04	05	14	**59**	11	14	06	00	02
M16	Encouraged expression of ideas	02	−01	−01	**85**	08	08	03	−06	−02
Individual Rapport										
M17	Friendly to individual students	−06	09	08	38	**34**	04	07	15	−05
M18	Welcomed students seeking advice	−04	10	15	28	**48**	−02	17	06	00
M19	Interested in individual students	−01	09	09	19	**61**	03	08	07	04
M20	Accessible to individual students	13	−01	−06	02	**70**	08	−02	−03	−01
Breadth of Coverage										
M21	Contrasted various implications	−05	05	04	08	−01	**70**	−01	16	02
M22	Gave background of ideas/concepts	08	09	09	12	00	**48**	05	13	01
M23	Gave different points of view	−02	−02	12	11	06	**73**	02	−04	−01
M24	Discussed current developments	19	04	−13	00	14	**65**	08	−07	01
Grading/Examinations										
M25	Examination feedback valuable	−01	01	21	04	27	16	**31**	14	03
M26	Evaluation methods fair/appropriate	01	01	23	12	11	04	**54**	04	−07
M27	Tested course content as emphasized	12	01	06	03	20	19	**45**	−05	−08
Reading/Assignments										
M28	Readings/texts were valuable	12	06	−09	04	−06	12	38	**44**	14
M29	They contributed to understanding	32	−01	−02	12	10	12	21	**34**	06
Workload/Difficulty										
M32	Course difficulty (easy-hard)	01	−04	02	04	−04	00	−07	05	**89**
M33	Course workload (light-heavy)	−07	−02	02	−07	08	00	−05	03	**88**
M34	Course pace (slow-fast)	08	18	−07	03	−09	02	23	24	**47**
Overall Rating Items										
M31	Overall Instructor Rating	09	42	29	06	14	09	11	13	08
M30	Overall Course Rating	18	40	30	04	12	09	09	10	06

Note: The factor loadings in bold type, the target loadings, are for items designed to measure the factor.

For purposes of comparison, the corresponding correlations from the Australian study also appear in Table 6. Results described above for the present study are similar to those in the Australian data with one major exception; in the Australian study the correlation between the SEEQ Grading/Examinations factor and the Endeavor Exam factor is not nearly so high as in the Spanish data. This exception is primarily due to the poor definition of the SEEQ Grading/Examinations factor in the Australian study. Nevertheless, with this exception, there is a striking similarity between the results in the two countries.

Table 5
Factor analysis of Endeavor items (n = 627 sets of ratings)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Presentation Clarity								
F1	Presentations clarified materials	**60**	06	05	04	16	06	15
F2	Presented clearly and summarized	**63**	03	02	08	15	11	10
F3	Made good use of examples	**50**	08	13	11	−03	15	10
Workload/Difficulty								
F4	Students had to work hard	11	**86**	03	−02	01	01	03
F5	Course required a lot of work	−03	**87**	00	02	04	03	03
F6	Course workload was heavy	−03	**85**	−02	01	00	−03	−01
Individual Rapport/Personal Attention								
F7	Listened and was willing to help	06	01	**64**	20	06	07	02
F8	Able to get personal attention	03	01	**70**	04	06	07	15
F9	Concerned about student difficulties	05	02	**48**	30	08	12	03
Class Discussion								
F10	Class discussion was welcome	21	00	33	**44**	01	06	04
F11	Students encouraged to participate	01	04	07	**81**	06	01	06
F12	Encouraged students to express ideas	05	−01	10	**78**	05	06	05
Organization/Planning								
F13	Presentations planned in advance	35	07	20	−08	**30**	04	11
F14	Provided detailed course schedule	−04	12	00	19	**41**	15	06
F15	Activities orderly scheduled	08	00	04	−01	**81**	04	03
Grading								
F16	Grading was fair and impartial	05	01	14	−04	01	**78**	04
F17	Grading reflected student performance	02	−01	−02	03	04	**91**	02
F18	Grading indicative of accomplishments	03	01	−02	05	08	**83**	04
Learning/Value								
F19	Understood the advanced material	27	−12	−07	10	11	07	**41**
F20	Ability to analyze issues	02	05	09	06	06	01	**79**
F21	Increased knowledge and competence	02	06	05	00	00	08	**83**

Note: The factor loadings in bold type, the target loadings, are for items designed to measure the factor.

Discussion and Implications

Items from two American instruments designed to measure students' evaluations of teaching effectiveness were translated into Spanish and administered to a sample of Spanish university students. Most of the items were judged to be appropriate by the students, every item was chosen by at least a few as being most important, and all but the Workload/Difficulty items clearly differentiated between lectures whom students indicated to be "good," "average," and "poor." A series of factor analyses clearly identified the factors which the instruments were designed to measure and which have been identified in previous research. Finally, factors on the SEEQ and Endeavor instruments hypothesized to measure similar dimensions of effective teaching were found to be substantially correlated, while correlations between nonmatching factors were substantially smaller.

An important purpose of the present study was to determine whether components of effective teaching identified in responses by American university students could also be identified in responses by Spanish students. The identification of distinct components suggests that students differentiate among various components of teaching effectiveness and do not merely judge lecturers on a general good-bad dimension.

Table 6
(Multitrait-multimethod) matrix of correlations among SEEQ and Endeavor factors from responses by Spanish students (n = 627 sets of ratings) and Australian students (n = 316 sets)

SEEQ evaluation factors		1	2	3	4	5	6	7	8	9
1 Group Interaction	Aust	(94)								
	Span	(94)								
2 Learning/Value	Aust	26	(92)							
	Span	39	(92)							
3 Workload/Difficulty	Aust	−05	06	(91)						
	Span	04	08	(79)						
4 Exams/Grading	Aust	33	46	20	(81)					
	Span	42	50	13	(85)					
5 Individual Rapport	Aust	54	31	−03	32	(93)				
	Span	68	43	−05	39	(90)				
6 Organization/Clarity	Aust	24	52	−15	48	55	(93)			
	Span	39	50	02	46	43	(91)			
7 Enthusiasm	Aust	39	55	−04	52	47	60	(95)		
	Span	47	65	22	40	43	64	(92)		
8 Breadth of Coverage	Aust	42	39	−01	46	40	47	49	(88)	
	Span	62	55	12	52	55	45	57	(89)	
9 Assignments/Readings	Aust	22	37	07	39	18	35	37	33	(84)
	Span	32	25	−05	26	29	18	24	36	(84)

		SEEQ factors								*Endeavor factors*							
Endeavor factors		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
10 Class Discussion	Aust	**88**	29	−03	33	57	20	45	39	22	(85)						
	Span	**93**	37	03	38	69	38	43	59	29	(92)						
11 Student Accomplishments	Aust	33	**80**	−10	56	37	70	63	49	39	29	(85)					
	Span	46	**86**	11	52	51	55	66	60	31	44	(87)					
12 Workload	Aust	05	14	**75**	32	02	−02	06	05	20	05	03	(94)				
	Span	15	13	**82**	20	09	12	25	20	06	13	20	(91)				
13 Grading	Aust	28	39	04	**34**	39	43	31	36	50	25	35	06	(90)			
	Span	46	48	01	**80**	49	57	41	52	28	42	50	13	(94)			
14 Personal Attention	Aust	63	40	−05	43	**81**	41	56	57	29	60	44	04	40	(90)		
	Span	75	37	08	44	**81**	40	48	56	32	72	46	18	45	(91)		
15 Presentation Clarity	Aust	23	47	−13	55	35	**82**	71	49	32	23	60	00	31	43	(92)	
	Span	47	69	08	51	48	**79**	79	63	30	41	71	14	51	51	(89)	
16 Organization/Planning	Aust	26	58	06	51	35	**68**	59	56	39	21	60	16	41	43	67	(85)
	Span	48	59	18	60	48	**71**	57	59	24	46	58	28	60	45	67	(78)

Note: Coefficients in parentheses are coefficient alpha estimates of reliability, while those in bold type are the convergent validities.

Furthermore, earlier discussion proposes that students' evaluations cannot be adequately understood if this multidimensionality is ignored. The demonstration of a clearly defined factor structure which corresponds to that found in the Australian study as well as in American settings, argues that Spanish students do differentiate among different components and that the specific components have a remarkable generality across quite different nationalities. Similarly, the MTMM analysis of responses to the SEEQ and the Endeavor instruments shows that students differentiate among dimensions of effective teaching in a similar manner with both instruments. Despite the strong evidence for the separation of the various dimensions of effective teaching, there still existed substantial correlations among some of the factors in both the Australian and Spanish studies. For the SEEQ factors, correlations among the Learning/ Value, Instructor Enthusiasm, and Organization/ Clarity factors were all high, as was the correlation between Group Interaction and Organization/Clarity factors. Among the Endeavor factors, Organization/Planning and Clarity were highly correlated, while correlations between Personal Attention and Group Discussion, Organization/Planning and Student Accomplishments, and Clarity and Student Accomplishments were also high. Several points are relevant, however, in interpreting these high correlations. First, these correlations were substantially lower than the reliabilities of the factors and even lower than the convergent validities observed in the MTMM analysis. Second, these correlations are based upon responses by individual students where halo or method effects are likely to have a relatively large impact. Students' evaluations are typically summarized by the average response by all the students under a given instructor, and halo effects specific to particular students are likely to cancel out. Third, because the students were specifically asked to select "good," "average," and "poor" teachers, their ratings are likely to be stereotypic and biased against differentiation among dimensions (e.g., there would be a tendency to rate "bad" lecturers as poor on all items). Finally, some of the differentiation among components may be lost when students are asked to make retrospective ratings of former lecturers rather than to evaluate current lecturers. These findings clearly demonstrate that teaching effectiveness can be measured in a Spanish setting, that evaluation instruments developed at American universities apply in a Spanish setting and that the same components that underlie evaluations of teaching effectiveness at American universities apply in Spanish settings. The same conclusions also resulted from the similar study conducted at an Australian university. Taken together, these two studies suggest the possibility that students' evaluations of teaching effectiveness and components such as those contained in the SEEQ and Endeavor instruments apply to many university settings.

An important and provocative question raised by these findings is: Why are students' evaluations so widely employed at North American universities but not at universities in other countries? The conclusions from this study and the Australian study suggest that teaching effectiveness can be measured by students' evaluations in different countries and that perhaps other findings from research conducted in North America may generalize as well, so lack of empirical evidence is not the reason. A more likely explanation is the political climate in American universities. While the study of students' evaluations has a history dating from the 1920s in the United States, it was only in the late 1960s and 1970s that they became widely used in that country. During this period there was in the United States a marked increase in

student involvement in university policy making and also an increased emphasis on "accountability" in universities.

While the impetus for the increased use of students' evaluations of teaching effectiveness in North American universities may have been the political climate, subsequent research has shown such evaluations to be reliable, valid,relatively free from bias, and useful to students, lecturers, and administrators. Future research in the use of students' evaluations in different countries needs to take three directions. First, to test the generality of the conclusions from the present study, the study described here should be replicated in other countries. Second, the validity of the students' evaluations must be tested against a wide variety of indicators of effective teaching in different countries, as has been done in American research described earlier. Third, perhaps employing the instruments used in this study, investigators should examine and document the problems inherent in the actual implementation of broad, institutionallybased programs of students' evaluations of teaching effectiveness in different countries.

References

Aleamoni. L. M. (1981). Student ratings of instruction. In J. Millman (Ed.). Handbook of teacher evaluatton (pp. 110-145). Beverly Hills, CA: Sage.

Campbell, D. T.. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Buttetin, 56, 81-105.

Centra, J. A. (1979). Determining faculty effectiveness. San Francisco: Jossey-Bass.

Cohen. P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. Research in Higher Education, 13,321-341.

Cohen. P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.

Costin. F.. Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Resarch, 41, 511-536

de Wolf. W. A. (1974). Student ratings of instruction in post-secondary institutions: A comprehensive annotated bibliography of research reported since 1968, Vol. 1. Seattle: University of Washington, Educational Assessment Center.

Doyle, K. 0. (1975). Student evaluation of instruction. Lexington. MA: D. C. Heath.

Feldman. K. A. (1976a). Grades and college students' evaluations of their courses and teachers. Research in Higher Education, 4. 69-111.

Feldman. K. A. (1976b). The superior college teacher from the student's view. Research in Higher Education, 5, 243-288.

Feldman. K. A. (1977). Consistency and variability among college students in rating their teachers and courses. Research in Higher Education, 6, 223-274.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers and courses: What we know and what we don't. Research in Higher Education, 9, 199-242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. Research in Higher Education, 10, 149-172.

Feldman, K. A. (1982). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. Stony Brook, NY: State University of New York.

Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. Science, 182, 83-85.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. Research in Higher Education, 9, 69-91.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. American Educational Research Journal, 12. 327-336.

Hull. C. H., & Nie, H. H. (1981). SPSS update 7-9. New York: McGraw-Hill.

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. Kerlinger (Ed.), Review of research in education (Vol. 3). Itasca, IL: Peacock.

Leventhal, L., Perry, R. P., Abrami, P. C., Turcotte, S. J. C., & Kane, B. (1981, April). Experimental investigation of tenure promotion in American and Canadian universities. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 14. 441-447.

Marsh, H. W. (1980a). Research on students' evaluations of teaching effectiveness. Instructional Evaluation, 4, 5-13.

Marsh, H. W. (1980b). The influence of student, course and instructor characteristics on evaluations of university teaching. American Educational Research Journal, 17, 219-237.

Marsh, H. W. (1981a). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. Australian Journalof Education, 25, 177-192.

Marsh, H. W. (1981b). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. Applied Psychological Measurement, 6, 47-60.

Marsh. H. W. (1982a). Factors affecting students' evaluations of the same course taught by the same instructor on different occasions. American Educational Research Journal, 19, 485-497.

Marsh, H. W. (1982b). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. British Journal of Educational Psychology, 52,77-95.

Marsh, H. W. (1982c). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. Journal of Educational Psychology, 74, 264-279.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology, 75, 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

Marsh, H. W. (in press). Students as evaluatorsof teaching. In T. Hustn & T. N. Postlethwaite (Eds.), International encyclopedia of education: Research and studies. Oxford: Pergamon Press.

Marsh, H. W., Fleiner, H., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology, 67, 833-839.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 20,231-248.

Marsh, H. W., & Overall, J. U. (1980). Validityofstudents' evaluations of teaching effectiveness: Cognitive and affective criteria. Journal of Educational Psychology, 72, 468-475.

Marsh, H. W.. & Overall. J. U. (1981). The relative influence of course level, course type, and instructor on students' evaluations of college teaching. American Educational Research Journal, 18, 103-112.

Marsh, H. W., Overall, J. U., & Kesler. S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Journal of Educational Psychology, 71, 149-160.

Murray, H. G. (1980). Evaluating university teaching: A review of research. Toronto: Ontario Confederation of University Faculty Associations.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill.

Overall, J.-U., & Marsh, H. W. (1979). Midterm feeback from students: Its relationship & instructional improvement and students' cognitive and affective outcomes. Journal of Educational Psychology, 71,856-865.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. Journal of Educational Psychology, 72, 321-325.

Overall, J. U., & Marsh, H. W. (1982). Students' evaluations of teaching: An update. American Association for Higher Education Bulletin, 35(4). 9-13.