Journal of Educational Psychology
1984, Vol. 76, No. 5, 707-754

Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility

Herbert W. Marsh
University of Sydney, Australia

This article provides an overview of findings and research designs used to study students' evaluations of teaching effectiveness and examines implications and directions for future research. The focus of the investigation is on the author's own research that has led to the development of the Students' Evaluations of Educational Quality (SEEQ), but it also incorporates a wide range of other research. Based on this overview, class-average student ratings are (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases; and (f) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions. In future research a construct validation approach should be used in which it is recognized that effective teaching and students' evaluations designed to reflect it are multifaceted, that there is no single criterion of effective teaching, and that tentative interpretations of relations with validity criteria and with potential biases must be scrutinized in different contexts and examine multiple criteria of effective teaching.

Students' evaluations of teaching effectiveness are commonly collected at North American universities and colleges and are 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/promotion decisions; (c) information for students to use in the selection of courses and instructors; and (d) an outcome on a process description for research or teaching. Although the first purpose is nearly universal, the next two are not. At many universities systematic student input is required before faculty are even considered for promotion, whereas at others the inclusion of students' evaluations is optional. Similarly, the results of students' evaluations are published at some universities, whereas at others the results are considered to be strictly confidential.

The fourth purpose of student ratings, their use in research on teaching, has not been systematically examined. This is unfortunate. Research on teaching involves at least three major questions (Gage, 1963, 1972; Dunkin, in press): How do teachers behave? Why do they behave as they do? and What are the effects of their behavior? Dunkin goes on to conceptualize this research in terms of process variables (global teaching methods and specific teaching behaviors); presage variables (characteristics

I would like to thank Wilbert McKeachie, Kenneth Feldman, Peter Frey, Kenneth Doyle, Robert Menges, John Centra, Peter Cohen, Michael Dunkin, Samuel Ball, Jennifer Barnes, Les Leventhal, John Ware, and Philip Abrami for their comments on earlier research that is described in this article, and Jesse Overall, my co-author on many of the earlier studies. I would also like to gratefully acknowledge the support and encouragement that Wilbert McKeachie has consistently given to me from the time I was a graduate student first starting research in this area, and the invaluable assistance offered by Kenneth Feldman in both personal correspondence and the outstanding set of review articles he has authored. Nevertheless, the interpretations expressed in this article are those of the author, and may not reflect those of others whose assistance has been acknowledged.

Requests for reprints should be sent to Herbert W. Marsh, Department of Education, University of Sydney, Sydney, New South Wales 2006, Australia.

of teachers and students); context variables (substantive, physical, and institutional environments); and product variables (student academic/professional achievement, attitudes, and evaluations). Student ratings are important both as a process-description measure and as a product measure. This dual role played by student ratings, as a process description and as an evaluation of the process, is also inherent in their use as diagnostic feedback, as input for tenure promotion decisions, and as information for students to use in course selection.

Particularly in the last decade, the study of students' evaluations has been one of the most frequently emphasized areas in American educational research. Literally thousands of papers have been written, and an exhaustive review is beyond the scope of this article. The reader is referred to reviews by Aleamoni (1981), Centra (1979), Cohen (1980, 1981), Costin, Greenough, and Menges (1971), de Wolf (1974), Doyle (1975), Feldman (1976a, 1976b, 1977, 1978, 1979, 1983), Kulik and McKeachie (1975), Marsh (1980a, 1982b, in press), Murray (1980), Overall and Marsh (1982), and Remmers (1963). Individually, these studies may provide important insights. Yet, collectively the studies cannot be easily summarized, and opinions about the role of students' evaluations vary from "reliable, valid, and useful" to "unreliable, invalid, and useless" (Aleamoni, 1981). How can opinions vary so drastically in an area which has been the subject of thousands of studies? Part of the problem lies in the preconceived biases of those who study student ratings; a second part of the problem lies in unrealistic expectations of what student evaluations can and should be able to do; another part of the problem lies in the plethora of ad hoc instruments based upon varied item content and untested psychometric properties; and part of the problem lies in the fragmentary approach to the design of both student-evaluation instruments and the research based upon them.

Validating interpretations of student responses to an evaluation instrument involves an ongoing interplay between construct interpretations, instrument development, data collection, and logic. Each interpretation must be considered a tentative hypothesis to be challenged in different contexts and with different approaches. This process corresponds to defining a nomological network (Cronbach, 1971; Shavelson, Hubner, & Stanton, 1976) where differentiable components of students' evaluations of teaching effectiveness are related to each other and to other constructs. Within-network studies attempt to ascertain whether students' evaluations consist of distinct components and, if so, what these components are. This involves logical approaches such as content analysis and empirical approaches such as factor analysis and multitrait–multimethod (MTMM) analysis. Clarification of within-network issues must logically precede between-network studies in which students' evaluations are related to external variables. Inherent in this construct approach is the adage that one validates not a test, but the interpretation of data arising from specific applications, as responses may be valid for one purpose but not for another. Construct validity is never completely present or absent, and most studies lead to an intermediate conclusion in which the emphasis is on understanding relationships. A construct validation approach (see Cronbach, 1971; Shavelson et al., 1976 for a more extensive presentation) is used to examine student evaluation research to be described here.

The construct validation approach described here and elsewhere (Marsh, 1982b; 1983) has been incorporated more fully in the design, development, and research of the Students' Evaluations of Educational Quality (SEEQ) than with other student evaluation instruments. Consequently, the focus of this overview will be on my own research with SEEQ. In each section that follows, relevant SEEQ research is described, and methodological, theoretical, and empirical issues are related to other research in the field. The emphasis of this article on my own research with SEEQ can be justified by the nature of the article as an invited lead article, but also because SEEQ has been studied in a wider range of research studies than have other student evaluation instruments.

The purpose of this article is to provide an overview of findings conducted in selected areas of student evaluation research, to examine methodological issues and weaknesses in these areas of study, to indicate implications for the use and application of the rat-

ings, and to explore directions for future research. This research overview emphasizes the construct validation approach described above, and several perspectives about student-evaluation research that underlie this approach follow:

1. Teaching effectiveness is multifaceted. The design of instruments to measure students' evaluations and the design of research to study the evaluations should reflect this multidimensionality.

2. There is no single criterion of effective teaching. Hence, a construct approach to the validation of student ratings is required in which the ratings are shown to be related to a variety of other indicators of effective teaching. No single study, no single criterion, and no single paradigm can demonstrate, or refute, the validity of students' evaluations.

3. Different dimensions or factors of students' evaluations will correlate more highly with different indicators of effective teaching. The construct validity of interpretations based on the rating factors requires that each factor be significantly correlated with criteria to which it is logically and theoretically related, and less correlated with other variables. In general, student ratings should not be summarized by a response to a single item or an unweighted average response to many items. If ratings are to be averaged for a particular purpose, logical and empirical analyses specific to the purpose should determine the weighting each factor receives, so that the weighting will depend on the purpose.

4. An external influence, in order to constitute a bias to student ratings, must be substantially and causally related to the ratings, and relatively unrelated to other indicators of effective teaching. As with validity research, bias interpretations should be viewed as tentative hypotheses to be challenged in different contexts and with different approaches which are consistent with the multifaceted nature of student ratings. Bias interpretations must be made in the context of an explicit definition of what constitutes a bias.

Dimensionality

Information from students' evaluations necessarily depends on the content of the evaluation items. Poorly worded or inappropriate items will not provide useful information. Student ratings, like the teaching they represent, should be unequivocally multidimensional (e.g., a teacher may be quite well organized but lack enthusiasm). This contention is supported by common sense and a considerable body of empirical research. Unfortunately, most evaluation instruments and research fail to take cognizance of this multidimensionality. If a survey instrument contains an ill-defined hodgepodge of items, and student ratings are summarized by an average of these items, then there is no basis for knowing what is being measured, no basis for differentially weighting different components in the way most appropriate to the particular purpose they are to serve, nor any basis for comparing these results with other findings. If a survey contains separate groups of related items derived from a logical analysis of the content of effective teaching and the purposes the ratings are to serve, or a carefully constructed theory of teaching and learning, and if empirical procedures such as factor analysis and multitrait–multimethod analyses demonstrate that the items within the same group do measure separate and distinguishable traits, then it is possible to interpret what is being measured. The demonstration of a well-defined factor structure also provides a safeguard against a halo effect—a generalization from a subjective feeling, an external influence, or an idiosyncratic response mode—which affects responses to all items.

An important issue in the construction of multidimensional rating scale instruments is the content of the dimensions to be surveyed. A logical analysis of the content of effective teaching and the purposes of students' evaluations, coupled with feedback from students and faculty members, is one typical approach. An alternative approach based on a theory of teaching or learning could be used to posit the evaluation dimensions, though such an approach does not seem to have been used in student evaluation research. However, with each approach, it is important to also use empirical techniques such as factor analysis to further test the dimensionality of the ratings. The most carefully constructed instruments combine both logical/theoretical and empirical anal-

yses in the research and development of student rating instruments.

Factor analysis provides a test of whether students are able to differentiate among different components of effective teaching and whether the empirical factors confirm the facets that the instrument is designed to measure. The technique cannot, however, determine whether the obtained factors are important to the understanding of effective teaching; a set of items related to an instructor's physical appearance would result in a Physical Appearance factor that would probably have little to do with effective teaching. Consequently, carefully developed surveys typically begin with item pools based on literature reviews, and with systematic feedback from students, faculty members, and administrators about what characteristics are important and what type of feedback is useful (e.g., Marsh, 1982b; Hildebrand, Wilson, & Dienst, 1971). For example, in the development of SEEQ a large item pool was obtained from a literature review, forms in current usage, and interviews with faculty members and students about characteristics they see as constituting effective teaching. Then, students and faculty members were asked to rate the importance of items, faculty members were asked to judge the potential usefulness of the items as a basis for feedback, and open-ended student comments on pilot versions were examined to determine if important aspects had been excluded. These criteria, along with psychometric properties, were used to select items and revise subsequent versions. This systematic development constitutes evidence for the content validity of SEEQ and makes it unlikely that it contains any trivial factors.

Some researchers, while not denying the multidimensionality of student ratings, argue that a total rating or an overall rating provides a more valid measure. This argument is typically advanced in research where separate components of students' evaluations have not been empirically demonstrated, and so there is no basis for testing the claim. More important, the assertion is not accurate. First, there are many possible indicators of effective teaching and many possible uses for student ratings; the component that is "most valid" will depend on the criteria being considered (Marsh & Overall, 1980). Second, reviews of different validity criteria show that specific components of student ratings are more highly correlated with individual validity criteria than an overall or total rating (e.g., student learning, Cohen, 1981; instructor self-evaluations, Marsh, 1982c; Marsh, Overall, & Kesler, 1979b; effect of feedback for the improvement of teaching, Cohen, 1980). Third, the influence of a variety of background characteristics suggested by some as "biases" to student ratings is more difficult to interpret with total ratings than with specific components (Marsh, 1980b; 1983). Fourth, the usefulness of student ratings, particularly as diagnostic feedback to faculty, is enhanced by the presentation of separate components. Finally, even if it were agreed that student ratings should be summarized by a single score for a particular purpose, the weighting of different factors should be a function of logical and empirical analyses of the multiple factors for the particular purpose; an optimally weighted set of factor scores will automatically provide a more accurate reflection of any criterion than will a non-optimally weighted total. Hence, no matter what the purpose, it is logically impossible for an unweighted average to be more useful than an optimally weighted average of component scores.

Still other researchers, while accepting the multidimensionality of students' evaluations and the importance of measuring separate components for some purposes such as feedback to faculty, defend the unidimensionality of student ratings because, according to such an argument, when student ratings are used in personnel decisions, only one decision is made. However, such reasoning is clearly illogical. First, the use to which student ratings are put has nothing to do with the multidimensionality of student ratings, although it may influence the form in which the ratings are to be presented. Second, even if a single total score were the most useful form in which to summarize student ratings for personnel decisions, and there is no reason to assume that it is, this purpose would be poorly served by a ill-defined total score based on an ad hoc collection of items that was not appropriately balanced with respect to the components of

effective teaching that were being measured. If a single score were to be used, it should represent a weighted average of the different components where the weight assigned to each component was a function of logical and empirical analyses. There are a variety of ways in which the weights could be determined, including the importance of each component as judged by the instructor being evaluated, and the weighting could vary for different courses or for different instructors. However the weights are established, they should not be determined by the ill-defined composition of whatever items happen to appear on the rating survey, as is typically the case when a total score is used. Third, implicit in this argument is the suggestion that administrators are unable to utilize or prefer not to be given multiple sources of information for use in their deliberations, and there is no basis for such a suggestion. At institutions where SEEQ has been used, administrators prefer to have summaries of ratings for separate SEEQ factors for each course taught by an instructor for use in administrative decisions (see description of longitudinal summary report by Marsh, 1982b, pp. 78–79). Important unresolved issues in student evaluation research are how different rating components should be weighted for various purposes and what form of presentation is most useful for different purposes. The continued, and mistaken, insistence that students' evaluations represent a unidimensional construct hinders progress on the resolution of these issues.

Student Evaluation Factors Found With Different Instruments

The student evaluation literature does contain several examples of instruments that have a well-defined factor structure and that provide measures of distinct components of teaching effectiveness. Some of these instruments and the factors that they measure are

1. Frey's Endeavor instrument (Frey, Leonard, & Beatty, 1975; also see Marsh, 1981a): Presentation Clarity, Workload, Personal Attention, Class Discussion, Organization/Planning, Grading, and Student Accomplishments.

2. The instrument developed by Hildebrand, Wilson, and Dienst (1971): Analytic/Synthetic Approach, Organization/Clarity, Instructor Group Interaction, Instructor Individual Interaction, and Dynamism/Enthusiasm.

3. Marsh's SEEQ instrument (Marsh, 1982b; 1983): Learning/Value, Instructor Enthusiasm, Organization, Individual Rapport, Group Interaction, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty.

4. The Michigan State SIRS instrument (Warrington, 1973): Instructor Involvement, Student Interest and Performance, Student-Instructor Interaction, Course Demands, and Course Organization.

The systematic approach used in the development of each of these instruments and the similarity of the facets they measure support their construct validity. Factor analyses of responses to each of these instruments provide clear support for the factor structure they were designed to measure, and demonstrate that the students' evaluations do measure distinct components of teaching effectiveness. More extensive reviews describing the components found in other research (Cohen, 1981; Feldman, 1976b; Kulik & McKeachie, 1975) identify dimensions similar to those described here.

Illustrative evidence comes from research with SEEQ. Factor analyses of responses to SEEQ (Marsh, 1982b, 1982c, 1983) consistently identify the nine factors the instrument was designed to measure. Separate factor analyses of evaluations from nearly 5,000 classes were conducted on different groups of courses selected to represent diverse academic disciplines at graduate and undergraduate levels; each clearly identified the SEEQ factor structure (Marsh, 1983). In one study, faculty were asked to evaluate their own teaching effectiveness in 329 courses on the same SEEQ form completed by their students (Marsh, 1982c, Marsh & Hocevar, 1983). Separate factor analyses of student ratings and instructor self-evaluations each identified the nine SEEQ factors (see Table 1). In other research (Marsh & Hocevar, 1984) evaluations of the same instructor teaching the same course on different occasions demonstrated that even the multivariate pattern of ratings was gener-

Table 1
Factor Analyses of Students' Evaluations of Teaching Effectiveness (S) and the Corresponding Faculty Self-Evaluations of Their Own Teaching (F) in 329 Courses

	Factor pattern loadings																	
	1		2		3		4		5		6		7		8		9	
Evaluation items (paraphrased)	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F
1. Learning/Value																		
Course challenging/stimulating	42	40	23	25	09	−10	04	04	00	−03	15	27	09	05	16	23	29	20
Learned something valuable	53	77	15	02	10	−02	09	04	01	01	10	00	10	04	17	09	16	06
Increased subject interest	57	70	12	05	08	07	08	07	02	−03	18	08	03	−04	19	05	14	−02
Learned/understood subject matter	55	52	12	12	13	12	05	03	03	11	02	−01	19	07	14	−04	−23	−11
Overall course rating	36	33	25	29	16	09	12	08	09	02	12	16	13	−08	14	27	08	16
2. Enthusiasm																		
Enthusiastic about teaching	15	29	55	42	16	00	07	02	21	15	10	00	05	16	01	09	05	06
Dynamic & energetic	08	03	60	70	15	01	11	06	08	05	06	05	07	16	01	05	06	03
Enhanced presentations with humor	10	04	66	58	−04	06	05	01	13	02	12	02	14	07	02	−18	−07	−10
Teaching style held your interest	09	12	59	64	23	20	16	06	06	00	03	14	10	05	06	03	−02	−03
Overall instructor rating	12	27	40	54	23	09	14	08	23	02	11	16	10	−08	05	27	05	16
3. Organization																		
Instructor explanations clear	12	00	07	24	55	42	20	09	05	04	10	06	13	01	06	23	−08	−03
Course materials prepared & clear	06	06	03	−02	73	69	09	01	10	−02	09	04	06	03	10	03	01	12
Objectives stated & pursued	19	12	−05	−08	49	41	03	05	08	05	14	08	25	27	06	05	06	06
Lectures facilitated note taking	−03	02	20	09	58	53	−17	07	−02	05	14	04	15	06	08	01	−04	−05
4. Group Interaction																		
Encouraged class discussions	04	06	10	02	01	03	84	86	03	00	00	00	06	00	06	−05	00	−03
Students shared ideas/knowledge	02	08	06	−07	−04	−01	85	88	05	13	05	01	08	−02	08	−10	−02	01
Encouraged questions & answers	03	−04	06	09	14	06	62	69	16	−02	15	03	07	11	08	21	00	01
Encouraged expression of ideas	07	01	02	06	01	−11	73	75	20	09	05	07	09	12	05	09	00	−02
5. Individual Rapport																		
Friendly towards students	−04	10	17	06	00	−06	13	12	68	78	−01	−05	13	02	10	−05	−07	01
Welcomed seeking help/advice	04	−10	05	02	02	07	06	00	85	75	−04	04	12	06	05	20	03	−04
Interested in individual students	07	10	11	09	00	01	14	07	69	77	−01	−09	14	03	08	−09	03	09
Accessible to individual students	02	−13	−11	−11	16	09	09	−02	62	43	20	25	08	13	00	14	04	07
6. Breadth of Coverage																		
Contrasted implications	−05	02	12	01	05	03	08	01	−03	01	72	84	08	−03	14	02	08	−06
Gave background of ideas/concepts	08	03	08	10	16	07	−03	−02	02	−02	71	78	01	08	11	−01	03	03
Gave different points of view	04	−06	04	09	11	11	08	16	06	01	72	55	07	17	01	−06	04	08
Discussed current developments	23	29	08	−04	−04	−04	05	12	09	00	50	48	06	05	16	10	−01	−02
7. Examinations/Grading																		
Examination feedback valuable	−03	01	08	09	06	−11	09	05	08	12	−04	03	72	62	05	−03	09	03
Eval. methods fair/appropriate	06	02	00	−03	03	14	07	06	14	00	10	17	69	64	11	11	−08	04
Tested emphasized course content	08	00	−01	04	11	21	01	01	06	00	11	−04	70	58	07	10	−02	−03

Table 1 (*continued*)

Evaluation items (paraphrased)	Factor pattern loadings																	
	1		2		3		4		5		6		7		8		9	
	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F
8. Assignments																		
Reading/texts valuable	−06	09	−03	−03	03	07	−01	−06	03	01	07	−07	01	11	91	70	02	04
Added to course understanding	12	01	−01	−12	01	04	09	21	01	17	−02	08	07	05	81	56	06	10
9. Workload/Difficulty																		
Course difficulty (Easy-Hard)	−06	00	06	−01	04	−05	−04	02	−01	00	08	00	−04	08	10	04	85	74
Course workload (Light-Heavy)	14	−04	−09	−01	03	02	07	05	00	04	06	01	00	01	00	−04	88	86
Course pace (Too Slow-Too Fast)	−20	07	12	00	04	18	−12	−09	06	02	−03	−07	04	08	05	−04	62	32
Hours/week outside of class	14	00	07	00	−11	00	07	02	00	02	−04	03	03	−08	05	21	73	46

Note. Factor loadings in boxes are the loadings for items designed to measure each factor. All loadings are presented without decimal points. Factor analyses of student ratings and instructor self-ratings consisted of a principal-components analysis, Kaiser normalization, and rotation to a direct oblimin criterion. The analyses were performed with the commercially available Statistical Package for the Social Sciences (SPSS) routine (see Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

alizable (e.g., a teacher who was judged to be well organized but lacking enthusiasm in one course was likely to receive a similar pattern of ratings in other classes). These findings clearly demonstrate that student ratings are multidimensional, that the same factors underlie ratings in different disciplines and at different levels, and that similar ratings underlie faculty evaluations of their own teaching effectiveness.

In a study designed to test the applicability of North American surveys in an Australian university, Marsh (1981a) asked students to select a "good" and a "poor" instructor from their previous experience and to evaluate these instructors on a survey that contained items from both my SEEQ and Frey's Endeavor instrument that were described earlier. Even though most of these students had never before evaluated teaching effectiveness and the educational setting in Australian universities differs from North American universities, students indicated that virtually all the items in both instruments were appropriate. Separate factor analyses of responses to the SEEQ and Endeavor items identified the factors the respective instruments were designed to measure. All factors (except Workload/Difficulty on SEEQ and Course Demands on Endeavor) significantly differentiated between good and poor teachers. An MTMM analysis was conducted (see Table 2) on correlations between SEEQ and Endeavor factors. The convergent validities—correlations between factors that were hypothesized to be matching (underlined in Table 2; median r = .81)—were much higher than correlations between nonmatching factors (median r = .35), and approached the reliability of the rating factors (.91). A similar study (Marsh, Touron, & Wheeler, in press) was recently conducted at a Spanish university where the SEEQ and Endeavor items were translated into Spanish. The Spanish study also differed from the Australian study in that each student selected three instructors to represent a "good," an "average," and a "poor" instructor. The results of the Spanish study substantially replicated those from the Australian study, and the results of the corresponding MTMM matrix also appear in Table 2. The findings from both studies support the generality of the evalu-

Table 2
Multitrait–Multimethod Matrix of Correlations Among Students' Evaluations of Educational Quality (SEEQ) and Endeavor Factors From Responses by Spanish Students (N = 627 Sets or Ratings) and Australian Students (N = 316 Sets)

	SEEQ									Endeavor						
Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SEEQ																
1. Group Interaction																
Australian	(94)															
Spanish	(94)															
2. Learning/Value																
Australian	26	(92)														
Spanish	39	(92)														
3. Workload/Difficulty																
Australian	−05	06	(91)													
Spanish	04	08	(79)													
4. Exams/Grading																
Australian	33	46	20	(81)												
Spanish	42	50	13	(85)												
5. Individual Rapport																
Australian	54	31	−03	32	(93)											
Spanish	68	43	−05	39	(90)											
6. Organization/Clarity																
Australian	24	52	−15	48	33	(93)										
Spanish	39	50	02	46	43	(91)										
7. Enthusiasm																
Australian	39	55	−04	52	47	60	(95)									
Spanish	47	65	22	40	43	64	(92)									
8. Breadth of Coverage																
Australian	42	39	−01	46	40	47	49	(88)								
Spanish	62	55	12	52	55	45	57	(89)								
9. Assignments/Readings																
Australian	22	37	07	39	18	35	37	33	(84)							
Spanish	32	25	−05	26	29	18	24	36	(84)							
Endeavor																
10. Class Discussion																
Australian	88	29	−03	33	57	20	45	39	22	(85)						
Spanish	93	37	03	38	69	38	43	59	29	(92)						
11. Student Accomplishments																
Australian	33	80	−10	56	37	70	63	49	39	29	(85)					
Spanish	46	86	11	52	51	55	66	60	31	44	(87)					

Table 2 (*continued*)

Factor	SEEQ									Endeavor							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
12. Workload																
Australian	05	14	75	32	02	−02	06	05	20	05	03	(94)				
Spanish	15	13	82	20	09	12	25	20	06	13	20	(91)				
13. Grading																
Australian	28	39	−04	34	39	43	31	36	50	25	35	06	(90)			
Spanish	46	48	01	80	49	57	41	52	28	42	50	13	(94)			
14. Personal Attention																
Australian	63	40	−05	43	81	41	56	57	29	60	44	04	40	(90)		
Spanish	73	37	08	44	81	40	48	56	32	72	46	18	45	(91)		
15. Presentation Clarity																
Australian	23	47	−13	55	35	82	71	49	32	23	60	00	31	43	(92)	
Spanish	47	69	08	51	48	79	79	63	30	41	71	14	51	51	(89)	
16. Organization/Planning																
Australian	26	58	06	51	35	68	59	56	39	21	60	16	41	43	67	(85)
Spanish	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; underlined values are the convergent validities.

ation factors across independently constructed instruments and quite different educational settings.

Frey (1978) argued that two higher-order factors underlie the seven Endeavor dimensions, and he demonstrated quite different patterns of relations between each of the proposed higher-order factors and other variables such as class size, class-average grade, student learning in multisection validity studies, and an index of research citation counts. Frey argued that many of the inconsistencies in the student evaluation literature result from the inappropriate unidimensional analysis of ratings, which should be examined in terms of separate dimensions. Although the thrust of Frey's argument is similar to the emphasis here, his justification for summarizing the seven Endeavor dimensions with two higher-order dimensions is dubious. His analysis was based on responses to only 7 of the 21 Endeavor items, the higher-order factors were not easily interpreted, no attempt was made to test the ability of the two-factor solution to fit responses from the 21 items, other research has shown that responses to the 21 items do identify seven factors (Frey et al., 1975; Marsh, 1981a), and confirmatory factor analytic techniques designed to test higher-order structures were not used. Nevertheless, his findings do demonstrate that the relation between student ratings and other variables does depend on the component of teaching effectiveness being measured.

Implicit Theories of Teaching Behaviors

Abrami, Leventhal, and Dickens (1981), Larson (1979), Whitely and Doyle (1976), and others have argued that dimensions identified by factor analyses of students' evaluations may reflect raters' *implicit theories* about dimensions of teacher behaviors in addition to, or instead of, dimensions of actual teaching behaviors. For example, if a rater assumes that the occurrence of behaviors X and Y are highly correlated and knows that the person being rated is high on X, then the rater may rate the person as high on Y even though the rater does not have an adequate basis for rating Y. Implicit theories are likely to have a particularly

large impact on factor analyses of individual student responses, which further argues against the use of the individual student as the unit of analysis. In fact, if the ratings by individual students within the same class are factor analyzed and it is assumed that the stimulus being judged is constant for different students—a problematic assumption—then the derived factors reflect primarily implicit theories.

Whitely and Doyle (1976) suggest that students' implicit theories are controlled for when factor analyses are performed on class-average responses, and Abrami, Leventhal, and Dickens (1981) warn that it is only when students are randomly assigned to classes that the "computation of class-means cancels out individual student expectations and response patterns as sources of variability" (p. 13). However, Larson (1979) demonstrated that even class-average responses, whether or not based on random assignment, are affected by implicit theories if the implicit theories generalize across students; it is only the implicit theories that are idiosyncratic to individual students, along with a variety of sources of random variation, that are canceled out in the formation of class averages. Larson goes on to argue that the validity of students' implicit theories cannot be tested with alternative factor analytic procedures based on student ratings, no matter what the unit of analysis, and that independent measures are needed. Hence, the similarity of the factor structures resulting from student ratings and instructor self-evaluations shown in Table 1 is particularly important. Although students and instructors may have similar implicit theories, instructors are uniquely able to observe their own behaviors and have little need to rely on implicit theories in forming their self-ratings. Thus, the similarity of the two factor structures supports the validity of the rating dimensions that were identified.

Summary of the Dimensionality of Student Ratings

In summary, most student evaluation instruments used in higher education, both in research and in actual practice, have not been developed using systematic logical and empirical techniques such as those described in this article. The surveys reviewed earlier each provided clear support for the multidimensionality of students' evaluations. The debate about which specific components of teaching effectiveness can and should be measured has not been resolved, though there seems to be consistency in those that are measured by the most carefully designed surveys. Students' evaluations cannot be adequately understood if this multidimensionality is ignored. Many orderly, logical relations are misinterpreted or cannot be consistently replicated because of this failure, and the substantiation of this claim will constitute a major focus of this article. Instruments used to collect students' evaluations of teaching effectiveness should be designed to measure separate components of teaching effectiveness, and support for both the content and construct validity of the multiple dimensions should be demonstrated.

Reliability, Stability, and Generalizability

Reliability

The reliability of student ratings is commonly determined from the results of item analyses (i.e., correlations among responses to different items designed to measure the same component of effective teaching) and from studies of interrater agreement (i.e., agreement among ratings by different students in the same class). The internal consistency among items is consistently high, but it provides an inflated estimate of reliability because it ignores the substantial portion of error due to the lack of agreement among different students, and so it generally should not be used (see Gilmore, Kane, & Naccarato, 1978 for further discussion). It may be appropriate, however, for determining whether the correlations between multiple facets have become so large that the separate facets cannot be distinguished, as in multitrait—multimethod (MTMM) studies.

The correlation between responses by any two students in the same class (i.e., the single-rater reliability) is typically in the .20s, but the reliability of the class-average response depends on the number of students rating the class (see Feldman, 1977, for a review of methodological issues and empirical findings). For example, the estimated reliability for SEEQ factors is about .95 for

the average response from 50 students, .90 from 25 students, .74 from 10 students, .60 from five students, and only .23 for one student. Given a sufficient number of students, the reliability of class-average student ratings compares favorably with that of the best objective tests. In most applications, this reliability of the class-average response, based on agreement among all the different students within each class, is the appropriate method for assessing reliability. Recent applications of generalizability theory demonstrate how error due to differences between items and error due to differences between ratings of different students can both be incorporated into the same analysis, but the error due to differences between items appears to be quite small (Gilmore et al., 1978).

Long-Term Stability

Some critics suggest that students cannot recognize effective teaching until after being called upon to apply course materials in further coursework or after graduation. According to this argument, former students who evaluate courses with the added perspective of time will differ systematically from students who have just completed a course when evaluating teaching effectiveness. Cross-sectional studies (Centra, 1979; Marsh, 1977) have shown good correlational agreement between the retrospective ratings of former students and those of currently enrolled students. In a longitudinal study (Marsh & Overall, 1979a; Overall & Marsh, 1980) the same students evaluated classes at the end of a course and again several years later, at least 1 year after graduation. End-of-class ratings in 100 courses correlated .83 with the retrospective ratings (a correlation approaching the reliability of the ratings), and the median rating at each time was nearly the same. Firth (1979) asked students to evaluate classes at the time of graduation from their university (rather than at the end of each class) and 1 year after graduation, and he also found good agreement between the two sets of ratings by the same students. These studies demonstrate that student ratings are quite stable over time, and argue that added perspective does not alter the ratings given at the end of a course.

In the same longitudinal study, Marsh (see Marsh & Overall, 1979a) demonstrated that consistent with previous research, the single-rater reliabilities were generally in the .20s for both end-of-course and retrospective ratings. (Interestingly, the single-rater reliabilities were somewhat higher for the retrospective ratings.) However, the median correlation between end-of-class and retrospective ratings, when based on responses by individual students instead of class-average responses, was .59. The explanation for this apparent paradox is the manner in which systematic unique variance, as opposed to random error variance, is handled in determining the single-rater reliability estimate and the stability coefficient. Variance that is systematic, but unique to the responses of a particular student, is taken to be error variance in the computation of the single-rater reliability. However, if this systematic variance was stable over the several year period between the end-of-course and retrospective ratings for an individual student, a demanding criterion, then it is taken to be systematic variance rather than error variance in the computation of the stability coefficient. While conceptual differences between internal consistency and stability approaches complicate interpretations, there is clearly an enduring source of systematic variation in responses by individual students that is not captured by internal consistency measures. This also argues that although the process of averaging across the ratings produces a more reliable measure, it also masks much of the systematic variance in individual student ratings, and that there may be systematic differences in ratings linked to specific subgroups of students within a class (also see Feldman, 1977). Various subgroups of students within the same class may view teaching effectiveness differently, and may be differentially affected by the instruction they receive, but there has been surprisingly little systematic research to examine this possibility.

Generalizability: Teacher and Course Effects

Researchers have also asked how highly correlated student ratings are in two different courses taught by the same instructor, and even in the same course taught by dif-

Table 3
Correlations Among Different Sets of Classes for Student Ratings and Background Characteristics

Measure	Same teacher, same course	Same teacher, different course	Different teacher, same course	Different teacher, different courses
Student rating				
Learning/Value	.696	.563	.232	.069
Enthusiasm	.734	.613	.011	.028
Organization/Clarity	.676	.540	−.023	−.063
Group Interaction	.699	.540	.291	.224
Individual Rapport	.726	.542	.180	.146
Breadth of Coverage	.727	.481	.117	.067
Examinations/Grading	.633	.512	.066	−.004
Assignments	.681	.428	.332	.112
Workload/Difficulty	.733	.400	.392	.215
Overall course	.712	.591	−.011	−.065
Overall instructor	.719	.607	−.051	−.059
Mean coefficient	.707	.523	.140	.061
Background characteristic				
Prior subject interest	.635	.312	.563	.209
Reason for taking course (percent indicating general interest)	.770	.448	.671	.383
Class average expected grade	.709	.405	.483	.356
Workload/difficulty	.773	.400	.392	.215
Course enrollment	.846	.312	.593	.058
Percent attendance on day evaluations administered	.406	.164	.214	.045
Mean coefficient	.690	.340	.491	.211

ferent teachers on two different occasions. This research is designed to address three related questions. First, what is the generality of the construct of effective teaching as measured by students' evaluations? Second, what is the relative importance of the effect of the instructor who teaches a class on students' evaluations, compared with the effect of the particular class being taught? If the impact of the particular course is large, then the practice of comparing ratings of different instructors for tenure/promotion decisions may be dubious. Third, should ratings be averaged across different courses taught by the same instructor?

In my 1981 study (Marsh, 1981b) I arranged ratings of 1,364 courses into sets of four such that each set contained ratings of the same instructor teaching the same course on two occasions, the same instructor teaching two different courses, and the same course taught by a different instructor. For an overall instructor rating item the correlation between ratings of different instructors teaching the same course was −.05, whereas correlations for the same instructor in different courses (.61) and in two different offerings of the same course (.72) were much larger (see Table 3). Although this pattern was observed in each of the SEEQ factors, the correlation between ratings of different instructors in the same course (i.e., a course effect) was slightly higher for some evaluation factors (e.g., Workload/Difficulty, Assignments, and Group Interaction) but had a mean of only .14 across all the factors. In marked contrast, correlations between background variables in different sets of courses (e.g., prior subject interest, class size, reason for taking the course) were higher for the same course taught by two different instructors than for two different courses taught by the same instructor (see Table 3). Based on a path analysis of these results, I argued that the effect of the teacher on student ratings of teaching effectiveness is much larger than is the effect of the course being taught, and that there is a small portion of reliable variance that is unique to a particular instructor in a particular course

that generalizes across different offerings of the same course taught by the same instructor. Hence, students' evaluations primarily reflect the effectiveness of the instructor rather than the influence of the course, and some instructors may be uniquely suited to teaching some specific courses. A systematic examination of the suggestion that some teachers are better suited for some specific courses, and that this can be identified from the results from a longitudinal archive of student ratings, is an important area for further research.

Marsh and Overall (1981) examined the effect of course and instructor in a setting where all students were required to take all the same courses, thus eliminating many of the problems of self-selection. The same students evaluated instructors at the end of each course and again 1 year after graduation from the program. For both end-of-course and follow-up ratings, the particular instructor teaching the course accounted for 5 to 10 times as much variance as the course. These findings again demonstrate that the instructor is the primary determinant of student ratings rather than the course he or she teaches.

Marsh and Hocevar (1984) also examined the consistency of the multivariate structure of student ratings. University instructors who taught the same course at least four times over a 4-year period were evaluated by different groups of students in each of the four courses (N = 314 instructors, 1,254 classes, 31,322 students). Confirmatory factor analysis demonstrated not only the generalizability of the ratings of the instructors across the four sets of courses, but also the the generalizability of multivariate structure. For example, an instructor who was evaluated to be enthusiastic but poorly organized in one class received a similar pattern of ratings in other offerings of the same course. The results of this study demonstrate a consistency of the factor structure across the different sets of courses, a relative lack of method/halo effect in the ratings, and a generalizability of the multivariate structure; all of which provide a particularly strong demonstration of the multifaceted nature of student ratings.

Gilmore et al. (1978), applying generalizability theory to student ratings, also found that the influence of the instructor who teaches the course is much larger than that of the course that is being taught. They suggested that ratings for a given instructor should be averaged across different courses to enhance generalizability. If it is likely that an instructor will teach many different classes during his or her subsequent career, then tenure decisions should be based on as many different courses as possible; Gilmore et al. suggest at least five. However, if it is likely that an instructor will continue to teach the same courses in which he or she has already been evaluated, then results from at least two different offerings of each of these courses is suggested. These recommendations require that a longitudinal archive of student ratings be maintained for personnel decisions. These data would provide for more generalizable summaries, the assessment of changes over time, and the determination of which particular courses are best taught by a specific instructor. It is indeed unfortunate that some universities systematically collect students' evaluations but fail to keep a longitudinal archive of the results. Such an archive would help overcome some of the objections to student ratings (e.g., idiosyncratic occurrences in one particular set of ratings), enhance their usefulness, and provide an important data base for further research.

Validity

Student ratings, which constitute one measure of teaching effectiveness, are difficult to validate because there is no single criterion of effective teaching. Researchers who use a construct validation approach have attempted to demonstrate that student ratings are logically related to various other indicators of effective teaching. In this approach ratings are required to be substantially correlated with a variety of indicators of effective teaching and less correlated with other variables. In particular, rating factors are required to be most highly correlated with variables to which they are most logically and theoretically related. Within this framework, evidence for the long-term stability and the generalizability of student ratings described in the Reliability section may be interpreted as support for their va-

lidity. The most widely accepted criterion of effective teaching is student learning, but other criteria include changes in student behaviors, instructor self-evaluations, the evaluations of peers and/or administrators who actually attend class sessions, the frequency of occurrence of specific behaviors observed by trained observers, and the effects of experimental manipulations.

Multisection Validity Studies

It is difficult to validate students' evaluations against student learning measured by objective examination, because examination scores in different courses normally cannot be compared. However, this may be possible in large multisection courses in which different groups of students are presented the same materials by different instructors. In the ideal multisection validity study, (a) there are many sections of a large multisection course; (b) students are randomly assigned to sections or at least enroll without any knowledge about the sections or who will teach them; (c) there are pretest measures available that correlate substantially with final course performance for individual students; (d) each section is taught completely by a separate instructor; (e) each section has the same course outline, textbooks, course objectives, and final examination; and (f) the final examination is constructed to reflect the common objectives by some person who does not actually teach any of the sections, and, if there is a subjective component, is graded by an external person (for further discussion see Cohen, 1981; Marsh, 1980a; Marsh & Overall, 1980). Support for the validity of the student ratings would be demonstrated when the sections that evaluate the teaching as most effective near the end of the course are also the sections that perform best on standardized final examinations, and when plausible counter-explanations are not viable.

Rodin and Rodin (1972) reported a negative correlation between section-average grade and section-average evaluations of graduate students in charge of different quiz sections. Ironically, this highly publicized study did not really constitute a multisection validity study as described above, and contained serious methodological problems. First, the ratings were not of the instructor in charge of the course but of teaching assistants who played a small ancillary role in the actual instruction. Thus, there was no way to separate achievement produced by a teaching assistant from that due to the instructor; a student who put too much reliance on the teaching assistant at the expense of lectures by the instructor might evaluate the assistant highly and perform poorly on the exam. Doyle (1975) also argued that a negative correlation might be expected because it would be the less able students who would have the most need for the supplemental services provided by the teaching assistants. Second, the study was conducted during the third term of a year-long course, and students were free to change teaching assistants between terms. Furthermore, during the third term students were not even required always to attend sections led by the same teaching assistant. Consequently, the effects of different teaching assistants on student achievement were confounded. Third, there was no adequate measure of course achievement; performance was evaluated with problems given at the end of each segment of the course, and students could repeat each exam as many as six times without penalty. Hence, a teaching assistant who engendered resentment by applying added pressure on students to continue retaking the exam might be evaluated poorly even though his or her students eventually got more problems correct. Because there was no final examination, there is no evidence that students who got more problems correct actually knew more at the end of the course. Fourth, these negative findings have not been replicated in any other studies. Hence, even though there are possible explanations for the negative correlation, the serious methodological problems in the study render the findings uninterpretable (also see Frey, 1978). In reviewing this study, Doyle (1975) stated that "to put the matter bluntly, the attention received by the Rodin and Rodin study seems disproportionate to its rigor, and their data provide little if any guidance in the validation of student ratings" (p. 59). In retrospect, the most interesting aspect of this study was that such a methodologically flawed study received so much attention.

Even when the design of multisection validity studies is more adequate, numerous methodological problems may still exist. First, the sample size in any given study is almost always quite small—the number of different sections is generally about 15—and produces extremely large sampling errors. Second, most variance in achievement scores at all levels of education is attributable to student presage variables and researchers are generally unable to find appreciable effects due to differences in teacher, school practice, or teaching method (Cooley & Lohnes, 1976; McKeachie, 1963). In multisection validity studies so many characteristics of the setting are held constant that differences in student learning due to differences in teaching effectiveness are even further attenuated. Hence, although the design is defensible, it is also quite weak for obtaining achievement differences that are systematically correlated with students' evaluations. Third, the comparison of findings across different multisection validity studies is problematic, given the lack of consistency in measures of course achievement and student rating instruments. Fourth, other criteria of teaching effectiveness besides student learning should be considered; Marsh and Overall (1980) found that different criteria of effective teaching were not significantly correlated with each other even though each was significantly correlated with student ratings. Fifth, presage variables such as initial student motivation and particularly ability level must be equated across sections for comparisons to be valid, and even random assignment becomes ineffective at accomplishing this when the number of sections is large and the number of students within each section is small, because by chance alone some sections will have more able students than others. This paradigm does not constitute an experimental design in which students are randomly assigned to treatment groups that vary systematically in terms of experimentally manipulated variables, and so the advantages of random assignment are not so clear cut as in a standard experimental design. For this design the lack of initial equivalence is particularly critical, because initial presage variables are likely to be the primary determinant of end-of-course achievement. For this reason it is important to have effective pretest measures even when there is random assignment. Although this may produce a pretest sensitization effect, the effect is likely to be trivial because (a) the nature of pretest variables will differ substantially from that of posttest measures; (b) there is no intervention other than the normal instruction that students expect to receive; (c) it seems unlikely that the collection of pretest measures will systematically affect either teaching effectiveness or student performance; and (d) pretest measures may already be available from student records without having to actually be collected as part of the study. Also, a no-pretest control group could be included. In summary, the multisection validity design is inherently weak and there are many methodological complications in its actual application.

Cohen (1981) conducted a meta-analysis of all known multisection validity studies, regardless of methodological problems such as found in the Rodin and Rodin study. Across 68 multisection courses, student achievement was consistently correlated with student ratings of Skill (.50), Overall Course (.47), Structure (.47), Student Progress (.47), and Overall Instructor (.43). Only ratings of Difficulty had a near-zero or a negative correlation with achievement. The correlations were higher when ratings were of full-time teachers, when students knew their final grade when rating instructors, and when achievement tests were evaluated by an external evaluator. Other study characteristics (e.g., random assignment, course content, availability of pretest data) were not significantly related to the results. Many of the criticisms of the multisection validity study are at least partially answered by this meta-analysis, particularly problems due to small sample sizes and the weakness of the predicted effect, and perhaps the issue of the multiplicity of achievement measures and student rating instruments, though perhaps not the problem of initial section equivalence. These results provide strong support for the validity of students' evaluations of teaching effectiveness.

In several studies (Marsh, Fleiner, & Thomas, 1975; Marsh & Overall, 1980) I

identified an alternative explanation for positive results in multisection validity studies that I called the "grading satisfaction hypothesis" (also called the "grading leniency effect"). When course grades (known or expected) and performance on the final exam are substantially correlated, then higher evaluations may be due to (a) more effective teaching that produces greater learning and higher evaluations by students; (b) increased student satisfaction with higher grades that causes them to reward the instructor with higher ratings independent of more effective teaching or greater learning; or (c) initial differences in student characteristics (e.g., prior subject interest, motivation, and ability) that affect both teaching effectiveness and performance. The first hypothesis argues for the validity of student ratings as a measure of teaching effectiveness, the second may be interpreted as an undesirable bias in the ratings, and the third is the effects of presage variables that are accurately reflected by the student ratings. Even when there are no initial differences between sections, either of the first two explanations is viable, and Cohen's finding that validity correlations are substantially higher when students know their final course grade makes the grading satisfaction hypothesis a plausible counter-explanation.

Only in two studies (Marsh, Fleiner, & Thomas, 1975; Marsh & Overall, 1980) was the grading satisfaction hypothesis eliminated as a viable alternative. We reasoned that in order for satisfaction with higher grades to affect students' evaluations at the section-average level, section-average expected grades must differ at the time the student evaluations are completed. In both of these studies student ratings were collected before the final examination, and tests used to evaluate student performance prior to the final examination were not standardized across sections. Hence, although each student knew approximately how his or her performance compared with other students within the same section, there was no basis for knowing how the performance of any section compared with that of other sections, and thus there was no basis for differences between the sections in their satisfaction with expected grades. Consequently, section-average expected grades indicated by students at the time the ratings were collected did not differ significantly from one section to the next, and were not significantly correlated with section-average performance on the final examination (even though individual expected grades within each section were). Here, because section-average expected grades at the time the ratings were collected did not vary, they could not be the direct cause of higher student ratings that were positively correlated with student performance, or the indirect cause of the higher ratings as a consequence of increased student satisfaction with higher grades. In most studies section-average expected grades and section-average performance on the criterion measures are positively correlated, and the grading satisfaction hypothesis cannot be so easily eliminated.

The methodologically flawed study by Rodin and Rodin aroused considerable interest in multisection validity studies, and focused attention on the methodological weaknesses of the design. Perhaps more than any other type of study, the credibility of student ratings has rested on this paradigm. Researchers' preoccupation with the multisection validity study has had both positive and negative aspects. The notoriety of the Rodin and Rodin study required that further research be conducted. Despite methodological problems and difficulties in the interpretation of results, Cohen's meta-analysis demonstrates that sections for which instructors are evaluated more highly by students tend to do better on standardized examinations; a finding that has been taken as strong support for the use of the ratings. However, the limited generality of this type of setting, the inherent weakness of the design, and the possibility of alternative explanations all dictate that it is important to consider other paradigms in student-evaluation research.

Instructor Self-Evaluations

Validity paradigms in student evaluation research are often limited to a specialized setting (e.g., large multisection courses) or they use criteria such as retrospective ratings

of former students that are unlikely to convince skeptics. Hence, the validity of student ratings will continue to be questioned until criteria are utilized that are both applicable across a wide range of courses and widely accepted as a indicator of teaching effectiveness. Instructors' self-evaluations of their own teaching effectiveness are a criterion that satisfies both of these requirements. Furthermore, instructors can be asked to evaluate themselves with the same instrument used by their students, thereby testing the specific validity of the different rating factors.

Despite the apparent appeal of instructor self-evaluations as a criterion of effective teaching, it has had limited application. Centra (1973) found correlations of about .20 between faculty self-evaluations and student ratings, but both sets of ratings were collected at the middle of the term rather than at the end of the course. Blackburn and Clark (1975) also reported correlations of about .20, but they only asked faculty to rate their own teaching in a general sense rather than their teaching in a specific class that was also evaluated by students. In small studies with ratings of fewer than 20 instructors, correlations of .31 and .65 were reported by Braskamp, Caulley, and Costin (1979) and .47 by Doyle and Crichton (1978). In large studies with ratings of 50 or more instructors, correlations of .62, .49, and .45 were reported by Webb and Nolan (1955), Marsh, Overall, and Kesler (1979b), and Marsh (1982c), respectively.

Marsh (1982c; Marsh, Overall, & Kesler, 1979b) conducted the only studies where faculty in a large number of courses (81 and 329) were asked to evaluate their own teaching on the same multifaceted evaluation instrument completed by students. In both studies separate factor analyses of teacher and student responses identified the same evaluation factors (see Table 1 for 1982 results); student-teacher agreement on every dimension was significant (median rs = .49 and .45; see Table 4 for 1982 results); and mean differences between student and faculty responses were small and not statistically significant for most items, and were unsystematic when differences were significant (i.e., student ratings were higher than faculty self-evaluations in some areas but lower in others).

In MTMM studies, multiple traits (the student rating factors) are assessed by multiple methods (student ratings and instructor self-evaluations). Consistent with the construct validation approach discussed earlier, correlations (see Table 4 for MTMM matrix from 1982 study) between student ratings and instructor self-evaluations on the same dimension (i.e., convergent validities—median rs = .49 and .45) were higher than correlations between ratings on nonmatching dimensions (median rs = −.04 and .02), and this is taken as support for the divergent validity of the ratings. In the second study, separate analyses were also performed for courses taught by teaching assistants, undergraduate level courses taught by faculty, and graduate level courses. Support for both the convergent and divergent validity of the ratings was found in each set of courses.

This research has important implications. First, the fact that students' evaluations show significant agreement with instructor self-evaluations provides a demonstration of their validity that is acceptable to most researchers and can be examined in all instructional settings. Second, there is good evidence for the validity of student ratings for both undergraduate and graduate level courses. Third, support for the divergent validity demonstrates the validity of each specific rating factor as well as the ratings in general, and argues for the importance of using systematically developed, multifactor evaluation instruments.

Ratings by Peers

Peer ratings, based on actual classroom visitation, are often proposed as indicators of effective teaching (French-Lazovich, 1981; Centra, 1979), and hence a criterion for validating students' evaluations. In studies where peer ratings are *not* based on classroom visitation (e.g., Blackburn & Clark, 1975; Guthrie, 1954; Maslow & Zimmerman, 1956), ratings by peers have correlated well with student ratings of university instructors, but it is likely that peer ratings were based on information from students. Centra

Table 4
Multitrait-Multimethod Matrix: Correlations Between Student and Faculty Self-Evaluations in 329 Courses

	Instructor self-evaluation factor									Student evaluation factor								
Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Instructor self-evaluations																		
1. Learning/Value	(83)																	
2. Enthusiasm	29	(82)																
3. Organization	12	01	(74)															
4. Group Interaction	01	03	−15	(90)														
5. Individual Rapport	−07	−01	07	02	(82)													
6. Breadth	13	12	13	11	−01	(84)												
7. Examinations	−01	08	26	09	15	20	(76)											
8. Assignments	24	−01	17	05	22	09	22	(70)										
9. Workload/Difficulty	03	−01	12	−09	06	−04	09	21	(70)									
Student evaluations																		
10. Learning/Value	46	10	−01	08	−12	09	−04	08	02	(95)								
11. Enthusiasm	21	54	−04	−01	−02	−01	−03	−09	−09	45	(96)							
12. Organization	17	13	30	−03	04	07	09	00	−05	52	49	(93)						
13. Group Interaction	19	05	−20	52	00	−02	−14	−04	−08	37	30	21	(98)					
14. Individual Rapport	03	03	−05	13	28	−19	−03	−02	00	22	35	33	42	(96)				
15. Breadth	26	15	09	00	−14	42	00	09	02	49	34	56	17	15	(94)			
16. Examinations	18	09	01	−01	06	−09	17	−02	−06	48	42	57	34	50	33	(93)		
17. Assignments	20	03	02	09	−01	04	−01	45	12	52	21	34	30	29	40	42	(92)	
18. Workload/Difficulty	−06	−03	04	00	03	−03	12	22	69	06	02	−05	−05	08	18	−02	20	(87)

Note. Values in parentheses in the diagonals of the upper left and lower right matrices, the two triangular matrices, are reliability (coefficient alpha) coefficients (see Hull & Nie, 1981). The underlined values in the diagonal of the lower left matrix, the square matrix, are convergent validity coefficients that have been corrected for unreliability according to the Spearman Brown equation. The nine uncorrected validity coefficients, starting with Learning, would be .41, .48, .25, .46, .25, .37, .13, .36, & .54. All correlation coefficients are presented without decimal points. Correlations greater than .10 are statistically significant.

(1975) compared peer ratings based on classroom visitation and student ratings at a brand new university, thus reducing the probable confounding of the two sources of information. Three different peers evaluated each teacher on two occasions, but there was a relative lack of agreement among peers (mean r = .26), which brings into question their value as a criterion of effective teaching and precluded any good correspondence with student ratings (r = .20).

Morsh, Burgess, and Smith (1956) correlated student ratings, student achievement, peer ratings, and supervisor ratings in a large multisection course. Student ratings correlated with achievement, supporting their validity. Peer and supervisor ratings, although significantly correlated with each other, were not related either to student ratings or to achievement, which suggests that peer ratings may not have value as an indicator of effective teaching. Webb and Nolan (1955) reported good correspondence between student ratings and instructor self-evaluations, but neither of these indicators was positively correlated with supervisor ratings (which the authors indicated to be like peer ratings). Ward, Clark, and Harrison (1981) suggested a methodological problem with the collection of peer ratings in that the presence of a colleague in the classroom apparently affects the classroom performance of the instructor and provides a threat to the external validity of the procedure.

In summary, peer ratings based on classroom visitation do not appear to be substantially correlated with student ratings or with any other indicator of effective teaching. Although these findings neither support nor refute the validity of student ratings, they clearly indicate that the use of peer evaluations of university teaching for personnel decisions is unwarranted (see Scriven, 1981 for further discussion). Other reviews of the peer evaluation process in higher education settings (e.g., Centra, 1979; French-Lazovich, 1981) have also failed to identify studies that provide empirical support for the validity of peer ratings as an indicator of effective college teaching or as a criterion for student ratings. Murray (1980), in comparing student ratings and peer ratings, found peer ratings to be "(1) less sensitive, reliable, and valid; (2) more threatening and disruptive of faculty morale; and (3) more affected by non-instructional factors such as research productivity" (p. 45) than student ratings.

Behavioral Observations by External Observers

At the precollege level, observational records compiled by specially trained observers are frequently found to be positively correlated with both student ratings and student achievement (see Rosenshine, 1971; Rosenshine & Furst, 1973 for a review), and similar studies at the postsecondary level are also encouraging (see Dunkin, in press; Murray, 1980). Murray (1976) found high positive correlations between observers' frequency-of-occurrence estimates of specific teaching behaviors and an overall student rating. Cranton and Hillgartner (1981) examined relations between specific teaching behaviors observed on videotapes of lectures in a naturalistic setting and student ratings; student ratings of effectiveness of discussion were higher "when professors praised student behavior, asked questions and clarified or elaborated student responses" (p. 73); student ratings of organization were higher "when instructors spent time structuring classes and explaining relationships" (p. 73). Murray (1980) concluded that student ratings "can be accurately predicted from outside observer reports of specific classroom teaching behaviors" (p. 31).

In one of the most ambitious observation studies, Murray (1983) trained observers to estimate the frequency of occurrence of specific teaching behaviors of 54 university instructors who had previously obtained high, medium, or low student ratings in other classes. A total of 18–24 sets of observer reports were collected for each instructor. The median of single-rater reliabilities (i.e., the correlation between two sets of observational reports) was .32, but the median reliability for the average response across the 18–24 reports for each instructor was .77. Factor analysis of the observations revealed nine factors, and their content resembled factors in student ratings described earlier (e.g., Clarity, Enthusiasm, Interaction, Rapport, Organization). The observations

significantly differentiated among the three criterion groups of instructors, but were also modestly correlated with a set of background variables (e.g., sex, age, rank, class size). Unfortunately, Murray only considered student ratings on an overall instructor rating item, and these were based on ratings from a previous course rather than the one that was observed. Hence, MTMM-type analyses could not be used to determine whether specific observational factors were most highly correlated with matching student rating factors. The findings do show, however, that instructors who are rated differently by students do exhibit systematically different observable teaching behaviors.

Some observational studies focus on a limited range of teacher behaviors rather than the broad spectrum of behaviors considered in research described above, and the study of teacher clarity has been a particularly fruitful example. In field or correlational research, observers measure (count or rate) clarity-related behaviors (see Land, in press, for a description of the types of behaviors) in natural classroom settings and these are related to student achievement scores. In one such study, Land and Combs (1981) operationally defined teacher clarity as the number of false starts or halts in speech, redundantly spoken words, and tangles in words. More generally, teacher clarity is a term used to describe how clearly a teacher explains subject matter to students and is frequently examined with student evaluation instruments and with observational schedules. Teacher clarity variables are important because they can be reliably judged by students and by external observers, they are consistently correlated with student achievement, and they are amenable to both correlational and experimental designs (see Dunkin, in press; Land, in press; 1979; Rosenshine & Furst, 1973). In experimental settings, lesson scripts are videotaped that differ only in the frequency of clarity-related behaviors, and randomly assigned groups of subjects view different lectures and complete achievement tests. Most studies, whether they use correlational or experimental designs, focus on the positive relation between observations of clarity variables and student achievement. Dunkin (in press), and Rosenshine and Furst (1973) were particularly impressed with the robustness of this effect and its generality across different instruments, different raters, and different levels of education. Although they are not generally the primary focus in this area of research, student ratings of teaching effectiveness and particularly ratings of teacher clarity in conjunction with other variables have been collected in some studies.

Land and Combs (1981; also see Land & Smith, 1981) constructed 10 videotaped lectures in which the frequency of clarity-related behaviors was systematically varied to represent the range of these behaviors observed in naturalistic studies. Ten randomly assigned groups of students each viewed one of the lectures, evaluated the quality of teaching on a 10-item scale, and completed an examination. The group-average variables (i.e., the average of student ratings and achievement scores in each group) were determined, and the experimentally manipulated occurrence of clarity-related behaviors was significantly correlated with both student ratings and achievement, and student ratings and achievement were significantly correlated with each other. Student responses to the item "the teacher's explanations were clear to me" were most highly correlated with both the experimentally manipulated clarity behaviors and results on the achievement test. In an observational study, Hines, Cruickshank, and Kennedy (1982) found that observer ratings on a cluster of 29 clarity-related behaviors were substantially correlated with both student ratings and achievement in college level math courses. In a review of such studies, Land (in press) indicated that although clarity behaviors were significantly related to both ratings and achievement, the correlations with ratings were significantly higher.

Research on teacher clarity, though not specifically designed to test the validity of students' evaluations, offers an important paradigm for student evaluation research. Teacher clarity is evaluated by items on most student evaluation instruments, can be reliably observed in a naturalistic field study, can be easily manipulated in laboratory studies, and is consistently related to student

achievement. Both naturalistic observations and experimental manipulations of clarity-related behaviors are significantly correlated with student ratings and with achievement, and student ratings of teacher clarity are correlated with achievement. This pattern of findings supports the inclusion of clarity on student evaluation instruments, demonstrates that student ratings are sensitive to natural and experimentally induced manipulations of this variable, and supports the construct validity of the student ratings in such research.

Systematic observations by trained observers are positively correlated with both students' evaluations and student achievement, even though research described in the last section reported that peer ratings are not systematically correlated with either students' evaluations or student achievement. A plausible reason for this difference lies in the reliability of the different indicators. *Class-average* student ratings are quite reliable, but the average agreement between ratings by any two students (i.e., the single rater reliability) is generally in the .20s. Hence, it is not surprising that agreement between two peer visitors who attend only a single lecture is even lower. When observers are systematically trained and asked to rate the frequency of quite specific behaviors, and/or there are a sufficient number of ratings of each teacher by different observers, then it is reasonable that their observations will be more reliable than peer ratings and more substantially correlated with student ratings. However, further research is needed to clarify this suggestion. Although peer ratings and behavioral observations have been considered as separate in the present article, the distinction may not be so clear in actual practice; peers can be trained to estimate the frequency of specific behaviors, and some behavior observation schedules look like rating instruments. The agreement between multifaceted observational schedules and multiple dimensions of students' evaluations appears to be an important area for future research. However, a word of caution must be noted. The finding that specific teaching behaviors can be reliably observed and that they do vary from teacher to teacher does not mean that they are important. Here, as with student ratings, specific behaviors and observational factors must also be related to external indicators of effective teaching. In this respect, the research conducted on teacher clarity provides an important model.

Research Productivity

Teaching and research are typically seen as the most important products of university faculty. Research helps instructors to keep abreast of new developments in their field and to stimulate their thinking, and this in turn provides one basis for predicting a positive correlation between research activity and students' evaluations of teaching effectiveness. However, Blackburn (1974) caricatured two diametrically opposed opinions about the direction of the teaching/research relationship: (a) a professor cannot be a first rate teacher if he or she is not actively engaged in scholarship; and (b) unsatisfactory classroom performance results from the professor neglecting teaching responsibilities for the sake of publications. A review (Marsh, 1979; also see Centra, 1981) of 13 empirical studies that mostly used student ratings as an indicator of teaching effectiveness reported that there was virtually no evidence for a negative relation between effectiveness in teaching and research; most studies found no significant relation, and a few studies reported weak positive correlations. Faia (1976) found no relation at research-oriented universities, but a small significant relation where there was less emphasis on research. Centra (1981), in one of the largest studies (N = 4,596 faculty from a variety of institutions), found weak positive correlations (median r = .22) between number of articles published and students' evaluations for social sciences, but no correlation in natural sciences and the humanities. Marsh and Overall (1979b) found that instructor self-evaluations of their own research productivity (see Table 6 in the next section) were only modestly correlated with their own self-evaluations of their teaching effectiveness (rs between .09 and .41), and were less correlated with student ratings (rs between .02 and .21). However, 4 of these 11 correlations with students' evaluations did reach statistical significance, and the largest correlation was with student ratings of

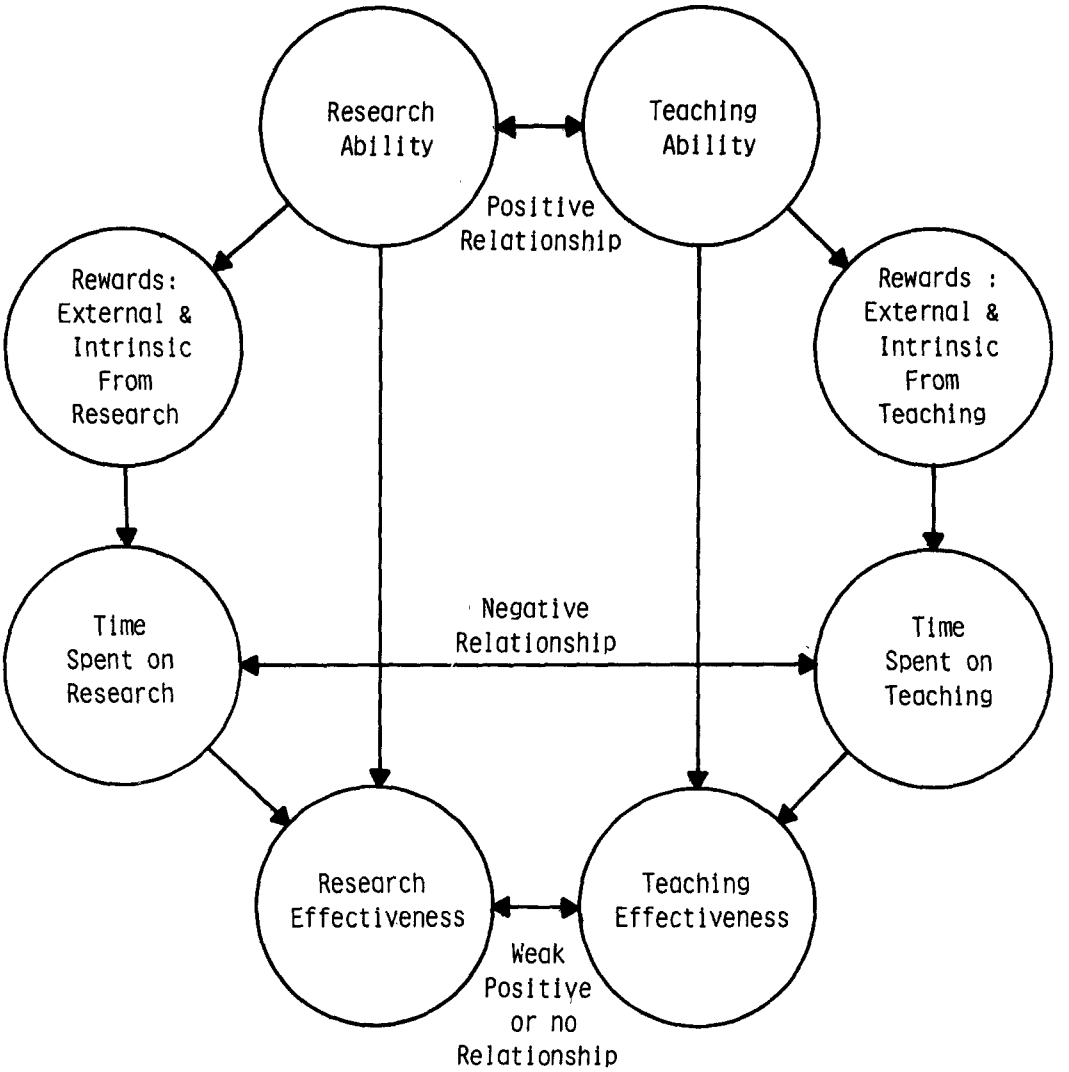

Figure 1. Model of predicted relations among teaching-related and research-related variables.

Breadth of Coverage. The researchers reasoned that this factor, which assesses characteristics such as "instructor adequately discussed current developments in the field," was the factor most logically related to research activity. Linsky and Straus (1975) found research activity was not correlated with students' global ratings of instructors, but did correlate modestly with student ratings of instructors' knowledge (.27). Frey (1978) found that citation counts for senior faculty in the sciences were significantly correlated with student ratings of Pedagogical Skill (.37), but not student ratings of Rapport (−.23, *ns*). Frey emphasized that the failure to recognize the multifaceted nature of students' evaluations may mask consistent relations, and this may account for the nonsignificant relations usually found.

Ability, time spent, and reward structure are all critical variables in understanding the teaching—research relation. In a model developed to explain how these variables are related (see Figure 1) it is proposed that (a) the abilities to be effective at teaching and research are positively correlated (a view consistent with the first opinion presented by Blackburn); (b) time spent on research and time spent on teaching are negatively correlated (a view consistent with the second

opinion presented by Blackburn) and may be influenced by a reward structure that systematically favors one over the other; (c) effectiveness, in both teaching and research, is a function of both ability and time; (d) the positive relation between abilities in the two areas and the negative correlation in time spent in the two areas will result in little or no correlation in measures of effectiveness in the two areas. Jauch (1976) found that research effectiveness was positively correlated with time spent on research and negatively correlated with time spent on teaching, and that time spent on teaching and research were negatively correlated with each other. Hence, although the model in Figure 1 has not been tested, Predictions b, c, and d are consistent with empirical findings.

Research examined in this subsection suggests that there is a zero to low-positive correlation between measures of research productivity and student ratings or other indicators of effective teaching, and that correlations may be somewhat higher for student rating dimensions that are most logically related to research effectiveness. Although these findings seem to neither support nor refute the validity of student ratings, they do demonstrate that measures of research productivity cannot be used to infer teaching effectivenes or vice versa.

Summary and Implications of Validity Research

Effective teaching is a hypothetical construct for which there is no single indicator. Hence, the validity of students' evaluations or of any other indicator of effective teaching must be demonstrated through a construct validation approach. Student ratings are significantly and consistently related to a number of varied criteria including the ratings of former students, student achievement in multisection validity studies, faculty self-evaluations of their own teaching effectiveness, and, perhaps, the observations of trained observers on specific processes such as teacher clarity. This provides support for the construct validity of the ratings. Peer ratings, based on classroom visitation, and research productivity were shown to have little correlation with students' evaluations, and because they are also relatively uncorrelated with other indicators of effective teaching, their validity as measures of effective teaching is problematic.

Additional research not described here has considered other indicators of effective teaching, but does not justify inclusion in an overview; the criteria are idiosyncratic to particular settings, are insufficiently described, or have only been considered in a few studies. For example, in a multisection validity study, Marsh and Overall (1980) found that sections who rated their teacher most highly were more likely to pursue further coursework in the area and to join the local computer club (the course was an introduction to computer programming). Ory, Braskamp, and Pieper (1980) found high correlations between student ratings and summative measures obtained from open-ended comments and a group interview technique, although the ratings proved to be the most cost effective procedure. Marsh and Overall (1979b) asked lecturers to rate how well they enjoyed teaching relative to their other duties, such as research, committees, and so on. Instructor enjoyment of teaching was significantly and positively correlated with students' evaluations and instructor self-evaluations (see Table 6 in the next section), and the highest correlations were with ratings of Instructor Enthusiasm. Surprisingly, little research has been done on the ability of colleagues to evaluate aspects of teaching such as course content, quality of examinations, and of reading lists; nor on the correlation between such evaluations and student ratings.

Nearly all researchers argue strongly that it is absolutely necessary to have multiple indicators of effective teaching whenever the evaluation of teaching effectiveness is to be used for personnel/tenure decisions. This emphasis on multiple indicators is clearly reflected in research described in this article. However, it is critical that the validity of *all* indicators of teaching effectiveness, not just student ratings, be systematically examined before they are actually recommended for use in personnel/tenure decisions. It seems ironic that researchers who argue that the validity of student ratings has not been sufficiently demonstrated, despite the preponderance of research supporting their

validity, are so willing to accept other indicators that have not been tested or have been shown to have little validity.

Relation to Background Characteristics: The Witch Hunt for Potential Biases in Students' Evaluations

The construct validity of students' evaluations requires that they be related to variables that are indicative of effective teaching, but relatively uncorrelated with variables that are not (i.e., potential biases). Because correlations between student ratings and other indicators of effective teaching rarely approach their reliability, there is considerable residual variance in the ratings that may be related to potential biases. Furthermore, faculty members generally believe that students' evaluations are biased by a number of factors they believe to be unrelated to teaching effectiveness. In a survey conducted at a major research university where SEEQ was developed (Marsh & Overall, 1979b) faculty members were asked which of a list of 17 characteristics would cause a substantial bias to student ratings, and over half the respondents cited course difficulty (72%), grading leniency (68%), instructor popularity (63%), student interest in subject before course (62%), course workload (60%), class size (60%), reason for taking the course (55%), and student's grade point average (GPA; 53%). In the same survey faculty members indicated that some measure of teaching quality should be given more emphasis in personnel decisions than was presently the case and that student ratings provided useful feedback to faculty. A dilemma existed in that faculty members wanted teaching to be evaluated, but were dubious about any procedure to accomplish this purpose. They were skeptical about the accuracy of student ratings for personnel decisions but were even more critical of classroom visitation, self-evaluations, and other alternatives. Whether or not potential biases actually affect student ratings, their utilization will be hindered so long as instructors think they are biased.

Marsh and Overall (1979b) also asked instructors to consider the special circumstances involved in teaching a particular course (e.g., class size, content area, students' interest in the subject, etc.) and to rate the "ease of teaching this particular course." These ratings of ease-of-teaching (see Table 6) were not significantly correlated with any of the student rating factors and were nearly uncorrelated with instructor self-evaluations. Scott (1977) asked instructors to indicate which, if any, "extenuating circumstances" (e.g., class size, class outside area of competence, first time taught the course, as well as an "other" category) would adversely hinder students' evaluations. The only extenuating circumstances to actually affect a total score representing the students' evaluations was class size, and this effect was small. These two studies suggest that extenuating circumstances that lecturers think might adversely affect students' evaluations in a particular course apparently have little effect, and also support earlier conclusions that the particular course has relatively little affect on students' evaluations compared with the effect of the lecturer who teaches the course.

Several large studies have looked at the multivariate relation between a comprehensive set of background characteristics and students' evaluations. In such research it is important that similar variables not be included both as items on which students rate teching effectiveness and as background characteristics, particularly when reporting some summary measure of variance explained. For example, Price and Magoon (1971) found that 11 background variables explained over 20% of the variance in a set of 24 student rating items. However, variables that most researchers would consider to be part of the evaluation of teaching (e.g., availability of instructor, explicitness of course policies) were considered as background characteristics and contributed to the variance explained in the student ratings. Similarly, in a canonical correlation relating a set of class characteristics to a set of student ratings Pohlman (1975) found that over 20% of the variance in five student rating items (i.e., the redundancy statistic described by Cooley & Lohnes, 1976) could be explained by background characteristics. However, the best predicted student rating item was course difficulty, and it was substantially correlated with the conceptually

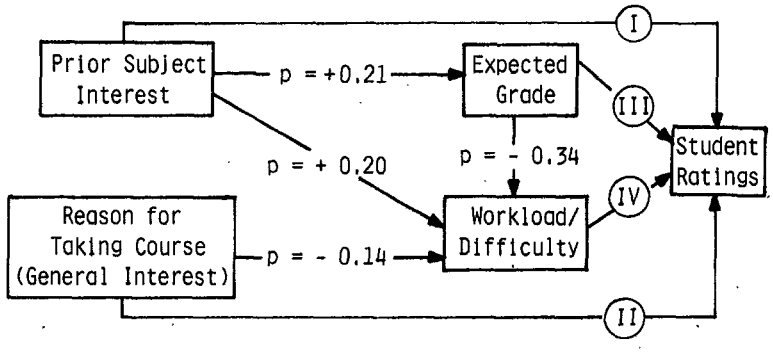

Figure 2. Path analysis model relating prior subject interest, reason for taking course, expected grade, and Workload/Difficulty. (Path coefficients for the student rating factors appear in Table 5.)

similar background characteristics of hours spent outside of class and expected grades.

Other multivariate studies have been more careful to separate variables considered as part of the students' evaluations from background characteristics. Brown (1976) found that 14% of the variance in an average of student rating items could be explained but that expected grade accounted for the most variance. Burton (1975) showed that eight background items explained 8%–15% of the variance in instructor ratings over a seven-semester period, but that the most important variable was student enthusiasm for the subject. Stumpf, Freedman, and Aguanno (1979) found that background variables accounted for very little of the variance in student ratings after the effects of expected grades (which they reported to account for about 5% of the variance) had been controlled.

A few studies have considered both multiple background characteristics and multiple dimensions of students' evaluations. One study (Marsh, 1980b) found that a set of 16 background characteristics explained about 13% of the variance in the set of SEEQ dimensions. However, the amount of variance explained varied from more than 20% in the overall course rating and the learning/value dimension, to about 2% of the organization and individual rapport dimensions. Four background variables were most important and could account for nearly all the explained variance; more favorable ratings were correlated with higher prior subject interest, higher expected grades, higher levels of Workload/Difficulty, and a higher percentage of students taking the course for general interest only. A path analysis (see Figure 2 and Table 5) demonstrated that prior subject interest had the strongest impact on student ratings, and that this variable also accounted for about one-third of the relation between expected grades and student ratings. Another study (Marsh, 1983) demonstrated a similar pattern of results in five different sets of courses (one of which was the set of courses used in the 1980 study) representing diverse academic disciplines at the graduate and undergraduate level, although the importance of a particular characteristic varied somewhat with the academic setting.

Based on a review of large multivariate studies that examine the combined effect of a set of background variables on student ratings, it appears that between 5% and 20% of the variance in student ratings can be explained, depending on the nature of the student rating items, the background characteristics, and perhaps the academic discipline. Prior subject interest, expected grades, and perhaps Workload/Difficulty seem to be the background variables most strongly correlated with students' evaluations of teaching.

A Construct Approach to the Study of Bias

The finding that a set of background characteristics are correlated with students' evaluations of teaching effectiveness should not be interpreted to mean that the ratings are biased, although this conclusion is often inferred by researchers. Support for a bias hypothesis, as with the study of validity,

Table 5
Path Analysis Model Relating Prior Subject Interest, Reason for Taking Course, Expected Grade and Workload/Difficulty to Student Ratings

	Factor											
	I. Prior Subject Interest			II. Reason (General Interest Only)			III. Expected Course Grade			IV. Workload/ Difficulty		
Student ratings	DC	TC	Orig r	DC	TC	Orig r	DC	TC	Orig r	DC	TC	Orig r
Learning/Value	36	44	44	15	13	15	26	20	29	17	17	12
Enthusiasm	17	23	23	09	08	09	20	16	20	11	11	06
Organization	−04	−04	−03	16	16	16	03	02	01	04	04	00
Group Interaction	21	28	29	06	06	07	30	27	31	06	06	−02
Individual Rapport	−05	09	09	−01	−02	−02	18	16	17	06	06	01
Breadth	−07	−03	−03	23	19	19	06	−01	−02	21	21	15
Exams/Grading	−05	03	03	12	10	10	25	18	18	20	20	10
Assignments	11	19	20	21	17	18	19	09	13	30	30	23
Overall course	23	32	33	19	15	16	26	15	22	30	30	23
Overall instructor	12	20	20	13	11	12	24	17	20	17	17	10
Variance components[a]	2.9%	5.1%	5.3%	2.3%	1.5%	1.8%	4.5%	2.6%	4.0%	3.6%	3.6%	1.8%

Note. The methods of calculating the path coefficients (p values in Figure 2), Direct Causal Coefficients (DC), and Total Causal Coefficients (TC) are described by Marsh (1980a). Orig r = original student rating. See Figure 2 for the corresponding path model.
[a] Calculated by summing the squared coefficients, dividing by the number of coefficients, and multiplying by 100%.

must be based on a construct approach. This approach requires that the background characteristics hypothesized to bias students' evaluations be examined in studies that are relatively free from methodological flaws using different approaches and that are interpreted in relation to a specific definition of bias. Despite the huge effort in this area of student-evaluation research, such a systematic approach is rare. Perhaps more than any other area of student-evaluation research, the search for potential biases is extensive, confused, contradictory, misinterpreted, and methodologically flawed. In the subsections that follow, methodological weaknesses common to many studies are presented, theoretical definitions of bias are discussed, and alternative approaches to the study of bias are considered. The purpose of these subsections is to provide guidelines for evaluating existing research and for conducting future research. Finally, within this context, relations between students' evaluations and specific characteristics frequently hypothesized to bias student ratings are examined.

Methodological Weaknesses Common to Many Studies

Important and common methodological problems in the search for potential biases to students' evaluations include the following:

1. Using correlation to argue for causation. The implication that some variable *biases* student ratings argues that causation has been demonstrated, whereas correlation only implies that a relation exists.

2. Neglect of the distinction between practical and statistical significance. All conclusions should be based on some index of effect size as well as on tests of statistical significance.

3. Failure to consider the multivariate nature of both student ratings and a set of potential biases.

4. Selection of an inappropriate unit of analysis. Because nearly all applications of students' evaluations are based on class-average responses, this is nearly always the appropriate unit of analysis. The size and even the direction of correlations based on

class-average responses may be different from correlations obtained when the analysis is performed on responses by individual students. Hence, effects based on individual students as the unit of analysis must also be demonstrated to operate at the class-average level.

5. Failure to examine the replicability of findings in a similar setting and their generalizability to different settings. This is particularly a problem in studies based on small sample sizes or on classes from a single academic department at a single institution.

6. The lack of an explicit definition of bias against which to evaluate effects. If a variable actually affects teaching effectiveness and this effect is accurately reflected in student ratings, then the influence is not a bias.

7. Questions of appropriateness of experimental manipulations. Studies that attempt to simulate hypothesized biases with operationally defined experimental manipulations must demonstrate that the size of the manipulation and the observed effects are representative of results likely to occur in the actual application of students' evaluations (i.e., they must examine threats to the external validity of the findings).

Theoretical Definitions of Bias

An important problem in research that examines the effect of potential biases on students' evaluations is that adequate definitions of bias have not been formulated. The mere existence of a significant correlation between students' evaluations and some background characteristic should not be interpreted as support for a bias hypothesis. Even if a background characteristic is causally related to students' evaluations, there is insufficient evidence to support a bias hypothesis. For example, it can be plausibly argued that many of the validity criteria discussed earlier, the alternative indicators of effective teaching such as student learning and experimental manipulations of teacher clarity, are causally related to students' evaluations, but it makes no sense to argue that they bias students' evaluations. Support for a bias hypothesis must be based on a theoretically defensible definition of what constitutes a bias. Alternative definitions of bias, which are generally implicit rather than explicit, are described below.

One possible example, the simplistic bias hypothesis, is that if an instructor (a) gives students high grades; (b) demands little work of students; and (c) agrees to be evaluated in small classes only; then (d) he or she will be favorably evaluated on all rating items. Implicit in this hypothesis is the assumption that instructors will be rewarded on the basis of these characteristics rather than effective teaching. Two of my studies (Marsh, 1980b, 1983) refute such a hypothesis. The clarity of the factor structure underlying SEEQ demonstrates that students differentiate their responses on the basis of more than just global impressions, so that potential biases, if they do have an effect, will affect different rating dimensions differentially. No background variable was substantially correlated with more than a few SEEQ factors, and each showed little or no correlation with some of the SEEQ factors. The percentage of variance that could be explained in different dimensions varied dramatically. Furthermore, the direction of the Workload/Difficulty effect was opposite to that predicted by the hypothesis, whereas the class size effect was small for factors other than Group Interaction and Individual Rapport. Most importantly, the entire set of background variables, ignoring the question of whether or not any of them represent biases, was able to explain only a small portion of the variance in student ratings.

The simplistic bias hypothesis is a straw man and its rejection does not mean that student ratings are unbiased, but only that they are not biased according to this definition. More rigorous and sophisticated definitions of bias are needed. A more realistic definition, which has guided research based on SEEQ and seems implicit in other research, is that student ratings are biased to the extent that they are influenced by variables unrelated to teaching effectiveness and, perhaps, to the extent that this influence generalizes across all rating factors rather than being specific to the particular factors most logically related to the influ-

ence. For example, even though student learning in muiltisection validity studies is correlated with student ratings, this effect should not be considered a bias. However, this seemingly simple and intuitive notion of bias is difficult to test. It is not sufficient to show that some variable is correlated with student ratings and that a causal interpretation is warranted; it must also be shown that the variable is *not* correlated with effective teaching. This is difficult in that effective teaching is a hypothetical construct so that all the problems involved in trying to show that student ratings are valid come into play, and trying to prove a null hypothesis is always problematic. According to this definition of bias, most claims that students' evaluations are biased by any particular characteristic are clearly unwarranted.

Other researchers infer yet another definition of bias by arguing that ratings are biased to the extent that they are affected by variables not under the control of the instructor. According to this conceptualization, ratings must be fair to be unbiased, even to the extent of not accurately reflecting influences that do affect teaching effectiveness. Such a definition is particularly relevant to a variable like prior subject interest, which probably does affect teaching effectiveness in a way that is accurately reflected by student ratings (see discussion below and Marsh & Cooper, 1981). Ironically, this conceptualization would not classify a grading leniency effect (i.e., students giving better-than-deserved ratings to instructors as a consequence of instructors giving better-than-deserved grades to students) as a bias because this variable is clearly under the control of the instructor. Hence, although the issue of fairness is important, particularly when students' evaluations are to be used for personnel decisions, this definition of bias also seems to be inadequate. Although there is a need for further clarification of the issues of bias and fairness, it is also important to distinguish between these two concepts so that they are not confused. The fairness of students' evaluations needs to be examined separately from, or in addition to, the examination of their validity.

Still other researchers (e.g., Hoyt, Owens, & Grouling, 1973; also see Howard & Bray, 1979) seem to circumvent the problem of defining bias by statistically controlling for potential biases with multiple regression techniques or by forming normative (cohort) groups that are homogeneous with respect to potential biases (e.g., class size). However, underlying this procedure is the untested assumption that the variables being controlled are causally related to student ratings, and that the relation does represent a bias. For example, if inexperienced, less able teachers are systematically assigned to teach large introductory classes, then statistically removing the effect of class size is not appropriate. Furthermore, this approach is predicated on the existence of a theoretical definition of bias and offers no help in deciding what constitutes a bias. Thus, although this procedure may be appropriate and valuable in some instances, it should only be used cautiously and in conjunction with research findings that demonstrate that a variable does constitute a bias according to a theoretically defensible definition of bias or fairness.

Approaches to Exploring for Potential Biases

Over a decade ago McKeachie (1973) argued that student ratings could be better understood if researchers did not concentrate exclusively on trying to interpret background relations as biases but instead examined the meaning of specific relations. Following this orientation, several approaches to the study of background influences have been utilized. The most frequently used approach is simply to correlate class-average students' evaluations with a class-average measure of a background variable hypothesized to bias student ratings. Such an approach can be heuristic, but in isolation it can never be used to demonstrate a bias. Instead, hypotheses generated from these correlation studies should be more fully explored in further research using alternative approaches such as those described below.

One alternative approach (Bausell & Bausell, 1979; Marsh, 1982a) is to examine the relation between differences in background variables and differences in student ratings for two or more offerings of the same course taught by the same instructor. The

rationale here is that because the instructor is the single most important determinant of student ratings, the within-instructor comparison provides a more powerful analysis. I (Marsh, 1982a) found that for such pairs of courses the more favorably evaluated offering was correlated with (a) higher expected grades, and presumably better mastery, since grades were assigned by the same instructor to all students in the same course; (b) higher levels of Workload/Difficulty; and (c) the instructor having taught the course at least once previously, and presumably having benefited from that experience and the student ratings. Other background characteristics such as enrollment, reason for taking a course, and prior subject interest had little effect. Although it provides valuable insight, this approach is limited by technical difficulties involved in comparing sets of difference scores and by the lack of variance in difference scores representing both student ratings and the background characteristics (i.e., if there is little variance in the difference scores, then no relation can be shown).

A second approach is to isolate a specific variable, simulate the variable with an experimental manipulation, and examine its effect in experimental studies in which students are randomly assigned to treatment conditions. The internal validity (see Campbell & Stanley, 1963, for a discussion of internal and external threats to validity) of interpretations is greatly enhanced because many counter-explanations that typically exist in correlational studies can be eliminated. However, this can only be accomplished at the expense of many threats to the external validity of interpretations: the experimental setting or the manipulation may be so contrived that the finding has little generality to the actual application of student ratings; the size of the experimental manipulation may be unrealistic; the nature of the variable in question may be seriously distorted in its operationalization; and effects shown to exist when the individual student is the unit of analysis may not generalize when the class average is used as the unit of analysis. Consequently, although the results of such studies can be very valuable, it is still incumbent upon the researcher to explore the external validity of the interpretations and to demonstrate that similar effects exist in real settings where student ratings are actually used.

A third approach, derived from the construct validation emphasis, is based on the assumption that specific variables (e.g., background characteristics, validity criteria, experimental manipulations, etc.) should logically or theoretically be related to some specific components of students' evaluations and less related to others. According to this approach, if a variable is most highly correlated with the dimensions to which it is most logically connected, then the validity of the ratings is supported. For example, class size is substantially correlated with ratings of Group Interaction and Individual Rapport but not with other SEEQ dimensions (Marsh, Overall, & Kesler, 1979a; see discussion below). This pattern of findings argues for the validity of the student ratings. Many relations can be better understood from this perspective rather than from trying to support or refute the existence of a bias that affects all students ratings.

A related approach, which has guided much of the SEEQ research, is more closely tied to an earlier definition of bias. This approach is based on the assumption that a bias that is specific to student ratings should have little impact on other indicators of effective teaching. If a variable is related both to student ratings and to other indicators of effective teaching, then the validity of the ratings is supported. Using this approach, I asked instructors in a large number of classes to evaluate their own teaching effectiveness with the same SEEQ form used by their students, and the SEEQ factors derived from both groups were correlated with background characteristics. Support for the interpretation of a bias in this situation requires that some variable be substantially correlated with student ratings, but not with instructor self-evaluations of their own teaching. Of course, even when a variable is substantially correlated with both student and instructor self-evaluations, it is still possible that the variable biases both student ratings and instructor self-evaluations, but such an interpretation would require that the variable was not substantially correlated with yet other valid indicators of effective teaching. Also, when the pattern

of correlations between a specific variable and the set of student evaluation factors is similar to the pattern of correlations with faculty self-evaluation factors, there is further support for the validity of the student ratings. Results based on this and each of the other approaches are presented below.

Effects of Specific Background Characteristics

Hundreds of studies have used a variety of approaches to examine the influence of many specific background characteristics on students' evaluations of teaching effectiveness, and a comprehensive review is beyond the scope of this article. Many of the older studies may be of questionable relevance, and may also have been inaccurately described. Reviewers, apparently relying on secondary sources, have perpetuated these inaccurate descriptions and faulty conclusions (some findings commonly cited in "reviews" are based on older studies that did not even consider the variable they are cited to have examined; see Marsh, 1980a, for examples). Empirical findings in this area have been reviewed in an excellent series of articles by Feldman (1976a, 1976b, 1977, 1978, 1979, 1983), other review papers by Aubrecht (1981), Marsh (1983; in press), and McKeachie (1973, 1979), monographs by Centra (1979; Centra & Creech, 1976) and Murray (1980), and a chapter by Aleamoni (1981). Older reviews by Costin, Greenough, and Menges (1971), Kulik and McKeachie (1975), and the annotated bibliography by de Wolf (1974) are also valuable.

Results to be summarized below emphasize the description and explanation of the multivariate relations that exist between specific background characteristics and multiple dimensions of student ratings. This is a summary of findings based on some of the most frequently studied and/or the most important background characteristics, and of different approaches to understanding the relations.

Class size/enrollment. I reviewed previous research on the class size effect and examined correlations between class size and SEEQ dimensions (Marsh, Overall, & Kesler, 1979a; Marsh, 1980b; 1983). Class size was moderately correlated with Group Interaction and Individual Rapport (negatively, *r*s as large as −.30) but not with other SEEQ dimensions or with the overall ratings of course or instructor (absolute values of *r*s < .15). In the class size effect there was also a significant nonlinear function where small and very large classes were evaluated more favorably. These findings appeared also when instructor self-evaluations were considered; the pattern and magnitude of correlations between instructor self-evaluations of their own teaching effectiveness and class size was similar to findings based on student ratings (Marsh, Overall, & Kesler, 1979a; also see Table 6). The specificity of the class size effect to dimensions most logically related to this variable, and the similarity of findings based on student ratings and faculty self-evaluations argue that this effect is not a bias to student ratings; rather, class size does have a moderate effect on some aspects of effective teaching (primarily Group Interaction and Individual Rapport) and these effects are accurately reflected in the student ratings. This discussion of the class size effect clearly illustrates why students' evaluations cannot be adequately understood if their multidimensionality is ignored (also see Frey, 1978).

Superficially, these findings appear to contradict Glass's well-established conclusion that teaching effectiveness, inferred from achievement indicators or from affective measures, suffers when class size increases (e.g., Glass, McGaw, & Smith, 1981, pp. 35–43; Smith & Glass, 1980), but a more careful examination suggests that this might not be the case. Glass also found a nonlinear class-size effect, which he summarized as a logarithmic function where nearly all the negative effect occurred between class sizes of 1 and 40, and he did not present data for extremely large class sizes of several hundred. Within the range of class sizes reported by Glass, the class size/student evaluation relation that I found could also be fit by a logarithmic relation, and the increase in students' evaluations did not occur until class size was 200 or more. However, the suggestion that instruction for these extremely large classes may not suffer, or is even superior, has very important implications, since the offering of just a few of these

very large classes can free up enormous amounts of instructional time that can be used to substantially reduce the average class size in the range where the effect of class size does appear to be negative. However, I (Marsh, Overall, & Kesler, 1979a) argued that my correlational effect should be interpreted cautiously and speculated that the unexpectedly higher ratings for very large classes could be due to (a) the selection of particularly effective instructors with demonstrated success in such settings; (b) students systematically selecting classes taught by particularly effective instructors, thereby increasing class size; (c) an increased motivation for instructors to do well when teaching particularly large classes; and (d) the development of "large class" techniques instead of trying to use inappropriate, "small class" techniques that may produce lower ratings in moderately large classes. Clearly this is an area that warrants further research.

Prior subject interest. Marsh and Cooper (1981) reviewed previous studies of the prior subject interest effect, as did Feldman (1977) and Howard and Maxwell (1980), and examined its effect on SEEQ ratings by students (also see Marsh, 1980b, 1983) and by faculty. The effect of prior subject interest on SEEQ scores was greater than that of any of the 15 other background variables considered (Marsh, 1980b, 1983). In different studies prior subject interest was consistently more highly correlated with Learning/Value (*r*s about .4) than with any other SEEQ dimensions (*r*s between .3 and −.12). Instructor self-evaluations of their own teaching were also positively correlated with both their own and their students' perceptions of students' prior subject interest (see Table 6). The dimensions that were most highly correlated, particularly Learning/Value, were the same as observed with student ratings. The specificity of the prior subject interest effect to dimensions most logically related to this variable and the similarity of findings based on student ratings and faculty self-evaluations argue that this effect is not a bias to student ratings. Rather, prior subject interest is a variable that influences some aspects of effective teaching (particularly Learning/Value) and these effects are accurately reflected in both the student ratings and instructor self-evaluations. Higher student interest in the subject apparently creates a more favorable learning environment and facilitates effective teaching, and this effect is reflected in student ratings as well as faculty self-evaluations.

Workload/Difficulty. The Workload/Difficulty effect on students' evaluations was also one of the largest found (Marsh, 1980b, 1983). Paradoxically, at least based on the supposition that Workload/Difficulty is a potential bias to student ratings, higher levels of Workload/Difficulty were positively correlated with student ratings. I (Marsh, 1982a) found that in pairs of courses taught by the same instructor, the more highly rated course tended to be the one perceived to have the higher levels of Workload/Difficulty. Marsh and Overall (1979b) found that instructor self-evaluations of their own teaching effectiveness were less highly correlated with Workload/Difficulty than were student ratings, but the direction of these correlations was also positive (see Table 6). Because the direction of the Workload/Difficulty effect is opposite to that predicted as a potential bias effect, and this finding is consistent in both student ratings and instructor self-evaluations, Workload/Difficulty does not appear to constitute a bias to student ratings.

Expected grades. Studies based on SEEQ, and literature reviews (e.g., Centra, 1979; Feldman, 1976a; Marsh, Overall, & Thomas, 1976) have typically found class-average expected grades to be positively correlated with student ratings. There are, however, three quite different explanations for this finding. The "grading leniency hypothesis" proposes that instructors who give higher-than-deserved grades will receive higher-than-deserved student ratings, and this constitutes a serious bias to student ratings. The "validity hypothesis" proposes that better expected grades reflect better student learning, and that a positive correlation between student learning and student ratings supports the validity of student ratings. A "student characteristics hypothesis" proposes that preexisting student presage variables such as prior subject interest may affect student learning, student grades, and teaching effectiveness, so that the expected

Table 6
Background Characteristics: Correlations With Student Ratings (S) and Faculty Self-Evaluations (F) of Their Own Teaching Effectiveness (N = 183 undergraduate courses)

Background Characteristic	SEEQ factor[a]										
	Learn	Enthu	Organ	Group	Indiv	Brdth	Exams	Assign	Wrkld	Over Crse	Over Instr
Faculty Rating "Scholarly production in their discipline" (1 = *well below average* to 5 = *well above average*)											
S	17	07	18	04	06	21	04	17	11	14	16
F	28	20	40	09	11	26	25	25	10	40	41
Students Rating Course Workload/Difficulty (1 = *low* to 5 = *high*)											
S	20	08	01	04	06	18	04	23	53	26	16
F	−03	−04	03	03	00	00	21	15	—	17	09
Faculty Rating Course Workload/Difficulty (1 = *low* to 5 = *high*)											
S	08	02	01	−03	04	02	05	12	—	15	08
F	07	03	15	−09	10	−06	21	21	53	29	16
Students Rating expected course grade (1 = F to 5 = A)											
S	28	20	05	38	16	01	28	24	−25	26	27
F	11	−03	−07	17	−10	−11	−11	02	−19	−01	00
Faculty Rating of "Grading Leniency" (1 = *easy/lenient* to 5 = *hard/strict*)											
S	−04	−16	−06	06	−08	−05	−05	−02	26	−06	−10
F	00	04	06	16	14	08	32	19	28	14	03
Class size/enrollment (actual number of students enrolled)											
S	−24	−04	−13	−36	−21	−09	−22	−09	−07	−18	−20
F	−02	03	10	−43	−17	−03	−03	−11	−04	−04	−09
Faculty Rating "Enjoy teaching relative to other duties" (1 = *extremely unenjoyable* to 5 = *extremely enjoyable*)											
S	25	34	18	22	33	00	20	09	03	29	32
F	24	39	01	10	12	−21	−20	03	−03	15	22
Faculty Rating "Ease of teaching particular course" (1 = *very easy* to 5 = *very difficult*)											
S	07	−01	10	11	06	09	09	01	05	03	08
F	−12	−16	−07	17	12	06	05	04	17	−14	−10

Note. SEEQ = Students' Evaluations of Educational Quality. Correlations are presented without decimal points; all those greater than .15 are statistically significant. For more detail, see Marsh and Overall (1979b).
[a] See Table 1 for full factor names.

grade effect is spurious. Although these explanations of the expected grade effect have quite different implications, it should be noted that grades, actual or expected, must surely reflect some combination of student learning, the grading standards used by an instructor, and preexisting presage variables.

I examined the relations among expected grades, prior subject interest, and student ratings in a path analysis (Marsh, 1980b, 1983; also see Aubrect, 1981; Feldman, 1976a). Across all rating dimensions, nearly one-third of the expected grade effect could be explained in terms of prior subject interest. Since prior subject interest precedes expected grades, a large part of the expected grade effect is spurious, and this finding supports the student characteristic hypothesis. I interpreted the results, however, as support for the validity hypothesis, in that prior subject interest is likely to affect student performance in a class but is unlikely to affect grading leniency. Hence, support for the student characteristics hypothesis may also constitute support for the validity hypothesis; prior subject interest produces more effective teaching, which leads to better student learning, better grades, and higher evaluations. This interpretation, however, depends on a definition of bias in which student ratings are not biased to the extent that they reflect variables that actually influence effectiveness of teaching.

In a similar analysis, Howard and Maxwell (1980) found that most of the covariation between expected grades and class-average overall ratings was eliminated by controlling for prior student motivation and student progress ratings. In their path analysis, prior student motivation had a causal impact on expected grades that was nearly the same as I reported, and a causal effect on overall ratings that was even larger, whereas the causal effect of expected grades on student ratings was smaller than the one I found. They concluded that "the influence of student motivation upon student performance, grades, and satisfaction appears to be a more potent contributor to the covariation between grades and satisfaction than does the direct contaminating effect of grades upon student satisfaction" (p. 818).

Marsh and Overall (1979b) examined correlations among student and teacher ratings of teaching effectiveness, student ratings of expected grades, and teacher self-evaluations of their own "grading leniency" (see Table 6). Correlations between expected grades and student ratings were positive and modest (*r*s between .01 and .28) for all SEEQ factors except Group Interaction ($r = .38$) and Workload/Difficulty ($r = -.25$). Correlations between expected grades and faculty self-evaluations were close to zero (*r*s between −.11 and .11) except for Group Interaction ($r = .17$) and Workload/Difficulty ($r = -.19$). Correlations between faculty self-perceptions of their own "grading leniency" (on a scale from *easy/lenient grader* to *hard/strict grader*) and both student and teacher evaluations of effective teaching were small (*r*s between −.16 and .19) except for ratings of Workload/Difficulty (*r*s of .26 and .28) and faculty self-evaluations of Examinations/Grading ($r = .32$). The lack of correlation between grading leniency and student ratings and the similarity in the pattern of correlations between expected grades and ratings by students and by faculty seem to argue against the interpretation of the expected grade effect as a bias. Nevertheless, the fact that expected grades were more positively correlated with student ratings than with faculty self-evaluations may provide some support for a grading leniency bias.

I (Marsh, 1982a) compared differences in expected grades with differences in student ratings for pairs of offerings of the same course taught by the same instructor on two different occasions. I reasoned that differences in expected grades in this situation probably represent differences in student performance because grading standards are likely to remain constant, and differences in prior subject interest were small and relatively uncorrelated with differences in student ratings. I found even in this context that students in the more favorably evaluated course tended to have higher expected grades, which argued against the grading leniency hypothesis. It should be noted, however, that although this study was in a setting where differences due to grading leniency were minimized, there was no basis for contending that the grading leniency effect does not operate in other situations.

Marsh, Fleiner, and Thomas (1975) and Marsh and Overall (1980) examined class-average pretest scores, expected grades, student achievement, and student ratings in multisection validity studies described earlier. Students selected classes in a quasi-random fashion, and pretest scores on an achievement test, motivational variables, and student background variables were also collected at the start of the study. Although this set of pretest variables was able to predict course performance with reasonable accuracy for individual students, section-average differences were small and generally nonsignificant. Also, in each study, students knew how their individual performance compared with other students within the same section, but not how the average performance of their section compared with that of other sections. Primarily as a consequence of this feature of the study, class-average expected grades, which were collected along with student ratings shortly before the final examination, did not differ significantly from section to section. Hence, the correlation between examination performance and student ratings could only be interpreted as support for the validity hypothesis, and is unlikely to be due to either preexisting variables or a grading leniency effect. It is ironic that when researchers propose class-average grades (expected or actual) as a potential bias to student ratings, a positive correlation between ratings and grades is nearly always interpreted as a grading leniency effect, whereas a positive correlation between grades, as reflected in examination performance, and ratings in multisection validity studies is nearly always interpreted as an indication of validity; both interpretations are usually viable in both situations. Again, it must be cautioned that support for the validity hypothesis found here does not deny the appropriateness of other interpretations in other situations.

Peterson and Cooper (1980) compared students' evaluations of the same instructors by students who received grades and those who did not. The study was conducted at two colleges where students were free to cross-enroll, but where students from one college were assigned grades but those from the other were not. Class-average ratings were determined separately for students in each class who received grades and those who did not, and there was substantial agreement with evaluations by the two groups of students. Hence, even though class-average grades of those students who received grades were correlated with their class-average evaluations and showed the expected grade effect, their class-average evaluations were in substantial agreement with those of students who did not receive grades. This suggests that expected grade effect was not due to grading leniency because grading leniency was unlikely to affect ratings by students who did not receive grades.

Some researchers have argued that the expected grade effect can be better examined by randomly assigning students to different groups that are given systematically different grade expectations. For example, Holmes (1972) gave randomly assigned groups of students systematically lower grades than they expected and deserved, and found that these students evaluated the teaching effectiveness as poorer than did control groups. Although this type of research is frequently cited as evidence for the grading leniency effect, this conclusion is unwarranted. First, Holmes's manipulation accounted for no more than 8% of the variance in any of the evaluation items and much less variance across the entire set of ratings, and reached statistical significance for only 5 of 19 items (Did instructor have sufficient evidence to evaluate achievement? Did you get less than expected from the course? Clarity of exam questions? Intellectual stimulation? Instructor preparations?). Hence the size of the effect was small, was limited to a small portion of the items, and tended to be larger for items that were related to the specific experimental manipulation. Second, the results based on "rigged" grades that violate reasonably accurate grade expectations may not generalize to other settings and seem to represent a different variable than that examined in naturalistic settings (see Murray, 1980; and Abrami, Dickens, Perry, & Leventhal, 1980, for further discussion).

Abrami et al. (1980) reviewed other studies that attempted to experimentally manipulate grading standards and conducted two Dr. Fox experiments (see discussion

below) in which grading standards were experimentally manipulated. Groups of students viewed a videotaped lecture, rated teacher effectiveness, and took an objective exam. Students returned 2 weeks later and were given their examination results and a grade based on their actual performance but scaled according to different grading standards (i.e., with an "average" grade earning either a B, C+, or C). The subjects then viewed a similar videotaped lecture by the same instructor, again evaluated teacher effectiveness, and took a test on the content of the second lecture. The manipulation of grading standards had no effect on performance on the second achievement test and weak inconsistent effects on student ratings. There were also other manipulations (e.g., instructor expressiveness, content, and incentive), but the effect of grading standards accounted for no more than 2% of the variance in student ratings for any of the conditions, and failed to reach statistical significance in some. Not even the direction of the effect was consistent across conditions, and stricter grading standards occasionally resulted in higher ratings. These findings fail to support the contention that grading leniency produces an effect that is of practical significance, though the external validity of this interpretation may also be questioned.

In this summary of research about the expected grade effect, a modest but not unimportant correlation between class-average expected grades and student ratings has consistently been reported. There are, however, several alternative interpretations of this finding, which were labeled the "grading leniency hypothesis," the "validity hypothesis," and the "student characteristics hypothesis." Evidence from a variety of different types of research clearly supports the validity hypothesis and the student characteristics hypothesis, but does not rule out the possibility that a grading leniency effect operates simultaneously. Support for the grading leniency effect was found with experimental studies, but the effect was weak and may not be generalizable to nonexperimental settings in which student ratings are actually used. Consequently, although it is possible that a grading leniency effect may produce some bias in student ratings, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of student ratings.

Summary of the Search for Potential Biases

The search for potential biases to student ratings has itself been so biased that I labeled it a witch hunt. Methodological problems listed at the start of this section are common. Furthermore, research in this area is seldom guided by any theoretical definition of bias, and the definitions that are implicit in most studies are often inadequate. Research described above examines relations between a selected set of background characteristics and student ratings. Characteristics were selected that are most frequently examined, and/or have been found to be substantial. Although a similar review of other potential biases which have been considered elsewhere is beyond the scope of this article, I have tried to summarize my impression of "typical" relations between other background characteristics and student ratings in Table 7. These summaries are based on my subjective interpretation of many studies and reviews of the literature, and should only be taken as rough approximations; there is clearly a need for meta-analyses and systematic reviews such as those by Feldman described earlier to provide more accurate estimates of the size of effects that have been reported, and the conditions under which they were found. For most of the relations, the effects tend to be small, the results are often inconsistent, and the attribution to a bias is unwarranted if *bias* is defined as an effect that is specific to students' evaluations and does not also influence other indicators of teaching effectiveness. Perhaps the best summary of this area is McKeachie's (1979) conclusion that a wide variety of variables that could potentially influence student ratings apparently have little effect. Similar conclusions have been drawn by Centra (1979), Menges (1973), Marsh (1980b), Murray (1980), Aleamoni (1981), and others.

There are, of course, an almost infinite number of variables that could also be related to student ratings and could perhaps

Table 7
Overview of Relations Found Between Students' Evaluations of Teaching Effectiveness and Specific Background Characteristics

Background Characteristic	Summary of "Typical" Findings
Prior subject interest	Classes with higher prior subject interest are rated more favorably, though it is not always clear if interest existed before the start of course or was generated by the instructor.
Expected/actual grades	Classes expecting (or actually receiving) higher grades give somewhat higher ratings, though this can be interpreted to mean either that higher grades represent grading leniency or that superior learning occurs.
Reason for taking a course	Elective courses and those with a higher percentage taking a course for general interest tend to be rated slightly higher.
Workload/difficulty	Harder, more difficult courses that require more effort and time are rated somewhat more favorably.
Class size	Mixed findings but most find that smaller classes are rated more favorably, though some report curvilinear relations and a few find the effect limited primarily to items related to class discussion and individual rapport.
Level of course/year in school	Graduate level courses rated somewhat more favorably; weak, inconsistent findings suggesting that upper-division courses are rated higher than lower-division courses.
Instructor rank	Mixed findings, but little or no effect.
Sex of instructor &/or student	Mixed findings, but little or no effect.
Academic discipline	Weak tendency for higher ratings in humanities and lower ratings in sciences, but too few studies to be clear.
Purpose of ratings	Somewhat higher ratings if known to be used for tenure/promotion decisions.
Administration	Somewhat higher ratings if surveys not anonymous and/or instructor present when the survey is completed.
Student personality	Mixed findings, but apparently little effect, particularly for class-average responses, since different "personality types" may appear in somewhat similar numbers in different classes.

Note. For most of these characteristics, particularly the ones that have been more frequently studied, some studies have found results opposite to those reported here, whereas others have found no relation at all. The size, and in some cases even the direction, of the relation varies considerably depending on the particular component of students' evaluations being considered. Few studies have found any of these characteristics to be correlated more than .30 with class-average student ratings, and most reported relations are much smaller.

be suggested as potential biases. However, any such claim must be seriously scrutinized in a series of studies that are relatively free from the common methodological shortcomings, are based on an explicit and defensible definition of bias, and employ the type of logic used to examine the variables described above. Single studies of the predictive validity of psychological measures have largely been replaced by a series of construct validity studies, and a similar approach should also be taken in the study of potential biases. Simplistic arguments that a significant correlation between student ratings and some variable x demonstrates a bias can no longer be tolerated and are an injustice to the field. It is unfortunate that the cautious attitude toward interpreting correlations between student ratings and indicators of effective teaching as evidence of validity has not been adopted in the interpretation of external variables as a source of potential bias.

Dr. Fox Studies

The Dr. Fox Paradigm

The Dr. Fox effect is defined as the overriding influence of instructor expressiveness on students' evaluations of college/university teaching. The results of Dr. Fox studies have been interpreted to mean that an enthusiastic lecturer can entice or seduce favorable evaluations, even though the lecture may be devoid of meaningful content. In the original Dr. Fox study by Naftulin, Ware, and Donnelly (1973), a professional actor lectured to educators and graduate students in an enthusiastic and expressive manner,

and teaching effectiveness was evaluated. Despite the fact that the lecture content was specifically designed to have little educational value, the ratings were favorable. The authors and critics agree that the study was fraught with methodological weaknesses, including the lack of any control group, a poor rating instrument, the brevity of the lecture compared with an actual course, the unfamiliar topic coupled with the lack of a textbook with which to compare the lecture, and so on (see Abrami, Leventhal, & Perry, 1982; Frey, 1979; Ware & Williams, 1975). Frey (1979) comments that "this study represents the kind of research that teachers make fun of during the first week of an introductory course in behavioral research methods. Almost every feature of the study is problematic" (p. 1). Nevertheless, reminiscent of the Rodin and Rodin (1972) study described earlier, the results of this study were seized upon by critics of student ratings as support for the invalidity of this procedure for evaluating teaching effectiveness.

To overcome some of the problems, Ware and Williams (1975, 1977; Williams & Ware, 1976, 1977) developed the standard Dr. Fox paradigm. In their experiments, a series of six lectures, all presented by the same professional actor, were videotaped. Each lecture represented one of three levels of course content (the number of substantive teaching points covered) and one of two levels of lecture expressiveness (the expressiveness with which the actor delivered the lecture). Students viewed one of the six lectures, evaluated teaching effectiveness on a typical multi-item rating form, and completed an achievement test based on all the teaching points in the high content lecture. Ware and Williams (1979, 1980) reviewed their studies and similar studies by other researchers and concluded that differences in expressiveness consistently explained much more variance in student ratings than did differences in content.

Reanalyses and Meta-Analyses

Marsh and Ware (1982) reanalyzed data from the Ware and Williams studies. A factor analysis of the rating instrument identified five evaluation factors that varied in the way they were affected by the experimental manipulations. In the condition most like the university classroom, in which students were told before viewing the lecture that they would be tested on the materials and that they would be rewarded in accordance with the number of exam questions they answered correctly, the Dr. Fox effect was not supported. The instructor expressiveness manipulation only affected ratings of Instructor Enthusiasm, the factor most logically related to the manipulation, and content coverage significantly affected ratings of Instructor Knowledge and Organization/Clarity, the factors most logically related to that manipulation. When students were not given incentives to perform well, instructor expressiveness had more impact on all five student rating factors than when external incentives were present, though the effect on Instructor Enthusiasm was still largest. However, without external incentives, expressiveness also had a larger impact on student achievement scores than did the content manipulation (i.e., presentation style had more to do with how well students performed on the examination than did the number of questions that had been covered in the lecture). This finding demonstrated that, particularly when external incentives are weak, expressiveness can have an important impact on both student ratings and achievement scores. Across all the conditions, the effect of instructor expressiveness on ratings of Instructor Enthusiasm was larger than its effect on other student rating factors. Hence, as observed in the examination of potential biases to student ratings, this reanalysis indicates the importance of considering the multidimensionality of student ratings. An effect which has been interpreted as a bias to students' evaluations seems more appropriately interpreted as support for their validity with respect to one component of effective teaching.

Abrami, Leventhal, and Perry (1982) conducted a review and a meta-analysis of all known Dr. Fox studies. On the basis of their meta-analysis, they concluded that expressiveness manipulations had a substantial impact on overall student ratings and a small effect on achievement, whereas content manipulations had a substantial effect on achievement and a small effect on ratings.

Consistent with the Marsh and Ware reanalysis, they also found that in the few studies that analyzed separate rating factors, the rating factors that were most logically related to the expressiveness manipulation were most affected by it. Finally, they concluded that although the expressiveness manipulation did interact with the content manipulation and a host of other variables examined in the Dr. Fox studies, none of these interactions accounted for more than 5% of the variance in student ratings.

In addition to content and expressiveness effects, Dr. Fox studies have considered the effects of a variety of other variables: grading standards (Abrami et al., 1980); instructor reputation (Perry, Abrami, Leventhal, & Check, 1979); student personality characteristics (Abrami, Perry, & Leventhal, 1982); purpose of evaluation (Meier & Feldhusen, 1979); and student incentive (Williams & Ware, 1976; Perry, Abrami, & Leventhal, 1979; Abrami et al., 1980). The incompleteness with which these analyses are generally reported makes it difficult to draw conclusions, but apparently only instructor reputation had any substantial effect on student ratings; when students are led, through experimentally manipulated feedback, to believe that an instructor is an effective teacher, they rate him or her more favorably on the basis of one short videotaped lecture and presumably the manipulated feedback. Researchers have also prepared videotaped lectures manipulating variables other than content and expressiveness. For example, Land and Combs (1981; see earlier discussion) videotaped 10 lectures that varied only in teacher speech clarity, operationally defined as the number of false starts or halts in speech, redundantly spoken words, and tangles in words. As teacher clarity improved there was a substantial linear improvement in both student ratings of teaching effectiveness and student performance on a standardized achievement test.

Interpretations, Implications, and Problems

How should the results of the Dr. Fox type studies be evaluated? Consistent with an emphasis on the construct validity of multifaceted ratings in this article, a particularly powerful test of the validity of student ratings would be to show that each rating factor is strongly influenced by manipulations most logically associated with it and less influenced by other manipulations. This is the approach used in the Marsh and Ware reanalysis of the Dr. Fox data described above, and it offers strong support for the validity of ratings with respect to expressiveness and, perhaps, limited support for their validity with respect to content.

Multiple ratings factors have typically not been considered in Dr. Fox studies (but see Ware & Williams, 1977; and discussion of this study by Marsh & Ware, 1982). Instead, researchers have relied on total scores even though they collect ratings that do represent multiple rating dimensions (i.e., the same form as was shown to have five factors in the Marsh & Ware reanalysis, and/or items from the 1971 Hildebrand, Wilson, & Dienst study described earlier). However, this makes no sense when researchers also emphasize the differential effects of the experimental manipulations on the total score rating and the achievement outcome (e.g., Perry, Abrami, & Leventhal, 1979). According to this approach, student ratings may be invalid because they are "oversensitive" to expressiveness and "undersensitive" to content when compared with achievement scores (but see Abrami et al., 1982). It is hardly surprising that the number of examination questions answered in a lecture (only 4 of 26 exam questions are covered in the low content lecture, whereas all 26 are covered in the high content lecture) has a substantial impact on examination performance immediately after the lecture, and less impact on student ratings; more relevant is the finding that content also affects student ratings. Nor is it surprising that manipulations of instructor expressiveness have a large impact on the total rating score when some of the items specifically ask students to judge the characteristic that is being manipulated; more relevant is the finding that some rating factors are relatively unaffected by expressiveness and that achievement scores are affected by expressiveness. Student ratings are multifaceted, the different rating factors do vary in the way they are affected by different ma-

nipulations, and any specific criterion can be more accurately predicted by differentially weighting the student rating dimensions. Because most of the Dr. Fox studies are based on student rating instruments that do measure separate components, reanalyses of these studies, as was done in the Marsh and Ware study, should prove valuable.

There are still serious problems with the design of Dr. Fox studies, which limit their potential usefulness. First, further research is needed to investigate whether the size and nature of experimental manipulations are representative of those that actually occur in the field and to ensure that the effects are not limited to some idiosyncratic aspect of the Dr. Fox paradigm. For example, the Dr. Fox tapes were produced to be of similar length even though the amount of content varied systematically. The length was equated by adding irrelevant content, and this may simulate a lack of teacher clarity, which negatively affects both ratings and achievement as described earlier. Also, in the Abrami et al. (1982) meta-analysis, videotaped lectures developed at the University of Manitoba produced effect sizes twice as large for student ratings and achievement as those produced by tapes Ware and Williams developed. Second, the Dr. Fox type of study should be extended to include other teacher process variables such as instructor clarity manipulations. Third, additional indicators of effective teaching should be examined as well as achievement scores and student ratings administered immediately after viewing a videotaped lecture in a content area that is relatively unknown to students. This might include long-term achievement, performance on higher-order objectives rather than tasks requiring primarily knowledge-level objectives, noncognitive objectives such as increased interest or desire to pursue the subject further (see Marsh & Overall, 1980), lectures in content areas already familiar to the viewers, or the inclusion of textual information that also fully covers all the teaching points in the high-content lecture.

Finally, I would like to suggest a counter-explanation for some of the Dr. Fox findings and to propose research to test this hypothesis. Student ratings, like all psychological impressions, are relativistic and based on some frame of reference. For students in university classes the frame of reference is determined by their expectations for that class and by their experience in other courses. What is the frame of reference in Dr. Fox studies? Some instructor characteristics such as expressiveness and speech clarity can be judged in isolation because a frame of reference has probably been established through prior experience, and these characteristics do influence student ratings. For other characteristics such as content coverage, external frames of reference are not so well defined. For example, covering four teaching points during a 20-minute lecture may seem like reasonable content coverage to students, or even to instructors or to curriculum specialists. However, if students were asked to compare high and low content lectures on the amount of content covered after viewing both, to indicate which test items had been covered in the lecture, to evaluate content coverage relative to textual materials representing the content that was supposed to have been covered, or even to evaluate content coverage after completing an examination in which they were told that all the questions were supposed to be covered, then they would have a much better basis for evaluating the content coverage and I predict that their responses would more accurately reflect the content manipulation. Some support for this suggestion comes from a recent study by Leventhal, Turcotte, Abrami, and Perry (1983), in which students viewed one lecture that was either "good" (high in content and expressiveness) or "poor" (low in content and expressiveness), and a second lecture by the same lecturer that was also either good or poor. The sum across all ratings of the second lecture varied inversely with the quality of the first. This is a "contrast" effect, which is typical in frame of reference studies (e.g., Parducci, 1968); after viewing a poor lecture a second lecture seems better, after viewing a good lecture a second lecture seems poorer. Here, the authors also examined different rating factors and found that the effects of manipulations of instructor characteristics varied substantially according to the rating component (though the evaluation of Group Interaction and Individual Rapport on the basis of video-

taped lectures seems dubious). Unfortunately, the effects of content and expressiveness were intentionally confounded in this design, which was not intended to represent a standard Dr. Fox study.

Utility of Student Ratings

Improvement of Instruction

The introduction of a broad institution-based, carefully planned program of students' evaluations of teaching effectiveness is likely to lead to the improvement of teaching. Faculty will have to give serious consideration to their own teaching in order to evaluate the merits of the program. The institution of a program that is supported by the administration will serve notice that teaching effectiveness is being taken more seriously by the administrative hierarchy. The results of student ratings, as one indicator of effective teaching, will provide a basis for informed administrative decisions and thereby increase the likelihood that quality teaching will be recognized and rewarded, and that good teachers will be given tenure. The social reinforcement of getting favorable ratings will provide added incentive for the improvement of teaching, even at the tenured faculty level. Finally, faculty report that the feedback from student evaluations is useful to their own efforts for the improvement of their teaching. None of these observations, however, provides an empirical demonstration of improvement of teaching effectiveness resulting from students' evaluations.

Feedback Studies

In most studies of the effects of feedback from students' evaluations, classes are randomly assigned to experimental or control groups; students' evaluations are collected near the middle of the term; at least the ratings from one or more groups are returned to instructors as quickly as possible; and the various groups are compared at the end of the term on a second administration of student ratings as well as other variables. (There is considerable research on a wide variety of other techniques designed to improve teaching effectiveness that use student ratings as an outcome measure; see Levinson-Rose & Menges, 1981). SEEQ has been used in two such studies using multiple sections of the same course. In the first study results from an abbreviated form of the survey were simply returned to faculty, and the impact of the feedback was positive, but very modest (Marsh, Fleiner, & Thomas, 1975). In the second study (Overall & Marsh, 1979) researchers actually met with instructors in the feedback group to discuss the evaluations and possible strategies for improvement. In this study students in the feedback group subsequently performed better on a standardized final examination, rated teaching effectiveness more favorably at the end of the course, and experienced more favorable affective outcomes (i.e., feelings of course mastery, and plans to pursue and/or apply the subject). These two studies suggest that feedback, coupled with a candid discussion with an external consultant, can be an effective intervention for the improvement of teaching effectiveness (also see McKeachie et al., 1980).

Reviews of feedback studies have reached different conclusions (e.g., Abrami, Leventhal, & Perry, 1979; Levinson-Rose & Menges, 1981; McKeachie, 1979; Rotem & Glassman, 1979). Cohen (1980), in order to clarify this controversy, conducted a meta-analysis of all known feedback studies. Across these studies, instructors who received midterm feedback were subsequently rated about one-third of a standard deviation higher than nonfeedback instructors on the total rating (an overall rating item or the average of multiple items), and even larger differences were observed for ratings of Instructor Skill, Attitude Toward Subject, and Feedback to Students. Studies that augmented feedback with consultation produced substantially larger differences, but other methodological variations had no effect. The results of this meta-analysis support my findings described above and demonstrate that feedback from students' evaluations, particularly when augmented by consultation, can lead to improvement in teaching effectiveness.

Several issues still remain in the feedback research. First, the studies demonstrate that feedback without consultation is only modestly effective, and none of the studies reporting significant feedback effects with consultation provide an adequate control for

the effect of consultation without feedback (i.e., a placebo effect due to consultation, or a real effect due to consultation that does not depend on feedback from student ratings). Second, the criterion of effective teaching used to evaluate the studies was limited primarily to student ratings; only the Overall and Marsh study demonstrated a significant effect of feedback on achievement. Most other studies were not based on multiple sections of the same course, and so it was not possible to test the effect of feedback on achievement scores. Third, nearly all of the studies were based on midterm feedback from midterm ratings. This limitation, perhaps, weakens the likely effects in that many instructional characteristics cannot be easily altered in the second half of the course, and also the generality of this approach to the effects of end-of-term ratings in one term to subsequent teaching in other terms has not been examined. Furthermore, Marsh and Overall (1980) demonstrated in their multisection validity study that although midterm and end-of-term ratings were substantially correlated, midterm ratings were concluded to be less valid than end-of-term ratings because they were less correlated with measures of student learning. Fourth, most of the research is based on instructors who volunteer to participate, and this further limits the generality of the effect because volunteers are likely to be more motivated to use the feedback to improve their instruction. (This was not the case with the two studies based on SEEQ). Finally, reward structure is an important variable that has not been examined in this feedback research. Even though faculty may be intrinsically motivated to improve their teaching effectiveness, potentially valuable feedback will be much less useful if there is no extrinsic motivation for faculty to improve. To the extent that salary, promotion, and prestige are based almost exclusively on research productivity, the usefulness of student ratings as feedback for the improvement of teaching may be limited.

Usefulness in Tenure/Promotion Decisions

During the last 25 years a variety of surveys have been conducted to determine the importance of students' evaluations and other indicators of teaching effectiveness in evaluating total faculty performance in North American universities (for reviews see Centra, 1979; Leventhal et al., 1981; Seldin, 1975). Each survey found that classroom teaching was considered to be the most important criterion of total effectiveness, though research effectiveness may be more important at prestigious, research-oriented universities. In the earlier surveys "systematically collected student ratings" was one of the least commonly mentioned methods of evaluating teaching, and authors of those studies lamented that there seemed to be no serious attempt to measure teaching effectiveness. More recently, however, survey respondents indicate that department chairperson reports, followed by colleague evaluations and student ratings, are the most common criteria used to evaluate teaching effectiveness and that student ratings *should* be the most important (Centra, 1979). These findings demonstrate that the importance and usefulness of student ratings as a measure of teaching effectiveness have increased dramatically during the last 25 years.

Leventhal et al. (1981), and Salthouse, McKeachie, and Lin (1978) composed fictitious summaries of faculty performance that systematically varied reports of teaching and research effectiveness, and also varied the type of information given about teaching (department chairperson's report or department chairperson's report supplemented by summaries of student ratings). Both studies found reports of research effectiveness to be more important in evaluating total faculty performance at research universities, although Leventhal et al. found teaching and research to be of similar importance across a broader range of institutions. Although teaching effectiveness as assessed by the department chairperson's report did make a significant difference in ratings of overall faculty performance, neither study found that supplementing the chairperson's report with student ratings made any significant difference. However, neither study considered student ratings alone or even suggested that the two sources of evidence about teaching effectiveness were independent. Information from the ratings and chairperson's report was always consistent so that one was redundant, and it would be reason-

able for subjects in these studies to assume that the chairperson's report was at least partially based on students' evaluations. These studies demonstrate the importance of reports of teaching effectiveness but do not appear to test the impact of student ratings.

Usefulness in Student Course Selection

Little empirical research has been conducted on the use of ratings by prospective students in the selection of courses. UCLA students reported that the *Professor/Course Evaluation Survey* was the second most frequently read of the many student publications, following the daily campus newspaper (Marsh, 1983). Leventhal et al. (1975) found that students say that information about teaching effectiveness influences their course selection. Students who select a class on the basis of information about teaching effectiveness are more satisfied with the quality of teaching than are students who indicate other reasons (Centra & Creech, 1976; Leventhal et al., 1975). In an experimental field study, Coleman and McKeachie (1981) presented summaries of ratings of four comparable political science courses to randomly selected groups of students during preregistration meetings. One of the courses had received substantially higher ratings, and it was chosen more frequently by students in the experimental group than by those in the control group. Based on this limited information, it seems that student ratings are used by students in the selection of instructors and courses.

Summary of Studies of the Utility of Student Ratings

Studies of the usefulness of student ratings are infrequent and often anecdotal. This is unfortunate, because this is an area of research that can have an important and constructive impact on the practice of using students' evaluations of teaching effectiveness. Both in this section and in earlier discussion, important, unresolved issues were identified that are in need of further research. For example, for administrative decisions students' evaluations can be summarized by a single score representing an optimally weighted average of specific components or by the separate presentation of each of the multiple components, but there is no research to indicate which is most effective. If different components of students' evaluations are to be combined to form a total score, how should the different components be weighted? Again there is no systematic research to inform policy makers. Debates about whether students' evaluations have too much or too little impact on administrative decisions are seldom based on any systematic evidence about the amount of impact they actually do have. Researchers often indicate that students' evaluations are used as one basis for personnel decisions, but there is a dearth of research on the policy practices that are actually employed in the use of student ratings. A plethora of policy questions exist (e.g., how to select courses to be evaluated, the manner in which rating instruments are administered, who is to be given access to the results, how ratings from different courses are considered, whether special circumstances exist where ratings for a particular course can be excluded either a priori or post hoc, whether faculty have the right to offer their own interpretation of ratings, etc.) that are largely unexplored despite the apparently wide use of student ratings (see Feldman, 1979). Anecdotal reports often suggest that faculty members find student ratings useful, but there has been little systematic attempt to determine what form of feedback to faculty is most useful (though feedback studies do support the use of services by an external consultant) and how instructors actually use the results they do receive. Some researchers have cited anecdotal evidence for negative effects of student ratings (e.g., lowering standards), but these are also rarely documented with systematic research. Although students' evaluations are sometimes used by students in their selection of courses, there is little guidance about the type of information students want and whether this is the same as is needed for other uses of students' evaluations. These, and a wide range of related questions about how students' evaluations are actually used and how their usefulness can be enhanced, provide a rich field for further research.

Overview, Summary, and Implications

Research described in this article demonstrates that student ratings are clearly multidimensional, quite reliable, reasonably valid, relatively uncontaminated by many variables often seen as sources of potential bias, and are seen to be useful by students, faculty, and administrators. However, the same findings also demonstrate that student ratings may have some halo effect, have at least some unreliability, have only modest agreement with some criteria of effective teaching, are probably affected by some potential sources of bias, and are viewed with some skepticism by faculty as a basis for personnel decisions. It should be noted that this level of uncertainty probably also exists in every area of applied psychology and for all personnel evaluation systems. Nevertheless, the reported results clearly demonstrate that a considerable amount of useful information can be obtained from student ratings, useful for feedback to faculty, useful for personnel decisions, useful to students in the selection of courses, and useful for the study of teaching. Probably, students' evaluations of teaching effectiveness are the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research.

Despite the generally supportive research findings, student ratings should be used cautiously, and there should be other forms of systematic input about teaching effectiveness, particularly when they are used for tenure/promotion decisions. However, although there is good evidence to support the use of students' evaluations as one indicator of effective teaching, there are few other indicators of teaching effectiveness whose use is systematically supported by research findings. Based on the research reviewed here, other alternatives that may be valid include the ratings of previous students and instructor self-evaluations, but each of these has problems of its own. Alumni surveys typically have very low response rates and are still basically student ratings. Faculty self-evaluations may be valid for some purposes, but probably not when tenure/promotion decisions are to be based on them. (Faculty should, however, be encouraged to have a systematic voice in the interpretation of their student ratings.) Consequently, although extensive lists of alternative indicators of effective teaching are proposed (e.g., Centra, 1979), few are supported by systematic research, and none are as clearly supported as students' evaluations of teaching.

Why then, if student ratings are reasonably well supported by research findings, are they so controversial and so widely criticized? Several suggestions are obvious. University faculty have little or no formal training in teaching, yet find themselves in a position where their salary or even their job may depend on their classroom teaching skills. Any procedure used to evaluate teaching effectiveness would prove to be threatening and highly criticized. The threat is exacerbated by the realization that there are no clearly defined criteria of effective teaching, particularly when there continues to be considerable debate about the validity of student ratings. (Interestingly, measures of research productivity, the other major determinant of instructor effectiveness, are not nearly so highly criticized, despite the fact that the actual information used to represent them in tenure decisions is often quite subjective and there are serious problems with the interpretation of the objective measures of research productivity that are used.) As demonstrated in this overview, much of the debate is based on ill-founded fears about student ratings, but the fears still persist. Indeed, the popularity of two of the more widely used paradigms in student evaluation research, the multisection validity study and the Dr. Fox study, apparently stems from an initial notoriety produced by claims to have demonstrated that student ratings are invalid. This occurred even though the two original studies (the Rodin & Rodin, 1972, study, and the Naftulin, Ware, & Donnelly, 1973, study) were so fraught with methodological weaknesses as to be uninterpretable. Perhaps this should not be so surprising in the academic profession, in which faculty members are better trained to find counter-explanations for a wide variety of phenomena than to teach. Indeed, the state of affairs has resulted in a worthwhile and healthy scrutiny of student ratings and has generated a

considerable base of research from which to form opinions about their worth. However, the bulk of research supports their continued use as well as advocates further scrutiny.

References

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology, 72,* 107–118.

Abrami, P. C., Leventhal, L., & Dickens, W. J. (1981). Multidimensionality of student ratings of instruction. *Instructional Evaluation, 6(1),* 12–17.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1979). Can feedback from student ratings help to improve teaching? *Proceedings of the 5th International Conference on Improving University Teaching.* London, 1979.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research, 52,* 446–464.

Abrami, P. C., Perry, R. P., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74,* 111–125.

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

Aubrect, J. D. (1981). *Reliability, validity and generalizability of student ratings of instruction* (IDFA Paper No. 6). Kansas State University: Center for Faculty Evaluations and Development. (ERIC Document Reproduction Service No. ED 213 296)

Bausell, R. B., & Bausell, C. R. (1979). Student ratings and various instructional variables from a within-instructor perspective. *Research in Higher Education, 11,* 167–177.

Blackburn, R. T. The meaning of work in academia. (1974). In J. I. Doi (Ed.), *Assessing faculty effort.* (A special issue of *New Directions for Institutional Research.*) San Francisco: Jossey-Bass.

Blackburn, R. T., & Clark, M. J. (1975). An assessment of faculty performance: Some correlations between administrators, colleagues, student, and self-ratings. *Sociology of Education, 48,* 242–256.

Braskamp, L. A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. *American Educational Research Journal, 16,* 295–306.

Brown, D. L. (1976). Faculty ratings and student grades: A university-wide multiple regression analysis. *Journal of Educational Psychology, 68,* 573–578.

Burton, D. (1975). Student ratings—an information source for decision-making. In R. G. Cope (Ed.), *Information for decisions in postsecondary education: Proceedings of the 15th Annual Forum* (pp. 83–86). St. Louis, MO: Association for Institutional Research.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Centra, J. A. (1973). *Two studies on the utility of student ratings for instructional improvement: I. The effectiveness of student feedback in modifying college instruction. II. Self-ratings of college teachers: A comparison with student ratings.* (SIR Report No. 2). Princeton, NJ: Educational Testing Service.

Centra, J. A. (1975). Colleagues as raters of classroom instruction. *Journal of Higher Education, 46,* 327–337.

Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal, 14,* 17–24.

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

Centra, J. A. (1981, May). *Research report: Research productivity and teaching effectiveness.* Princeton, NJ: Educational Testing Service.

Centra, J. A., & Creech, F. R. (1976). *The relationship between student, teachers, and course characteristics and student ratings of teacher effectiveness* (Project Report 76-1). Princeton, NJ: Educational Testing Service.

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.

Coleman, J., & McKeachie, W. J. (1981). Effects of instructor/course evaluations on student course selection. *Journal of Educational Psychology, 73,* 224–226.

Cooley, W. W., & Lohnes, P. R. (1976). *Evaluation research in education.* New York: Irvington.

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

Cranton, P. A., & Hillgartner, W. (1981). The relationships between student ratings and instructor behavior: Implications for improving teaching. *Canadian Journal of Higher Education, 11,* 73–81.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443–507). Washington, D.C.: American Council on Education.

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. O. (1975). *Student evaluation of instruction.* Lexington, MA: D. C. Heath.

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

Dunkin, M. J. (in press). Research on teaching in higher education. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: MacMillan.

Faia, M. (1976). Teaching and research: Rapport or misalliance. *Research in Higher Education, 4*, 235–246.

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. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. *Research in Higher Education, 18*, 3–124.

Firth, M. (1979). Impact of work experience on the validity of student evaluations of teaching effectiveness. *Journal of Educational Psychology, 71*, 726–730.

French-Lazovich, G. (1981). Peer review: Documentary evidence in the evaluation of teaching. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 73–89). Beverly Hills, CA: Sage.

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

Frey, P. W. (1979). The Dr. Fox Effect and its implications. *Instructional Evaluation, 3*, 1–5.

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

Gage, N. L. (1963). *Handbook of research on teaching.* Chicago: Rand McNally.

Gage, N. L. (1972). *Teacher effectiveness and teacher education.* Palo Alto, CA: Pacific Books, 1972.

Gilmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. *Journal of Educational Measurement, 15*, 1–13.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Guthrie, E. R. (1954). *The evaluation of teaching: A progress report.* Seattle: University of Washington Press.

Hildebrand, M., Wilson, R. C., & Dienst, E. R. (1971). *Evaluating university teaching.* Berkeley: Center for Research and Development in Higher Education, University of California, Berkeley.

Hines, C. V., Cruickshank, D. R., & Kennedy, J. J. (1982, March). *Measures of teacher clarity and their relationships to student achievement and satisfaction.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructors. *Journal of Educational Psychology, 63*, 130–133.

Howard, G. S., & Bray, J. H. (1979). Use of norm groups to adjust student ratings of instruction: A warning. *Journal of Educational Psychology, 71*, 58–63.

Howard, G. S., & Maxwell, S. F. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72*, 810–820.

Hoyt, D. P., Owens, R. F., & Grouling, T. (1973). *Interpreting student feedback on instruction and courses.* Manhattan: Kansas State University.

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

Jauch, L. R. (1976). Relationships of research and teaching. Implications for faculty evaluation. *Research in Higher Education, 5*, 1–13.

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In Kerlinger (Ed.), *Review of research in education* (Vol. 3, pp. 210–240). Itasca, IL: Peacock.

Land, M. L. (1979). Low-inference variables of teacher clarity: Effects on student concept learning. *Journal of Educational Psychology, 71*, 795–799.

Land, M. L. (in press). Vagueness and clarity in the classroom. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies.* Oxford: Pergamon Press.

Land, M. L., & Combs, A. (1981). *Teacher clarity, student instructional ratings, and student performance.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Land, M. L., & Smith, L. R. (1981). College student ratings and teacher behavior: An experimental study. *Journal of Social Studies Research, 5*, 19–22.

Larson, J. R. (1979). The limited utility of factor analytic techniques for the study of implicit theories in student ratings of teacher behavior. *American Educational Research Journal, 16*, 201–211.

Leventhal, L., Abrami, P. C., Perry, R. P., & Breen, L. J. (1975). Section selection in multi-section courses: Implications for the validation and use of student rating forms. *Educational and Psychological Measurement, 35*, 885–895.

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.

Leventhal, L., Turcotte, S. J. C., Abrami, P. C., & Perry, R. P. (1983). Primacy/recency effects in student ratings of instruction: A reinterpretation of gain-loss effects. *Journal of Educational Psychology, 75*, 692–704.

Levinson-Rose, J., & Menges, R. J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research, 51*, 403–434.

Linsky, A. S., & Straus, M. A. (1975). Student evaluations, research productivity, and eminence of col-

lege faculty. *Journal of Higher Education, 46*, 89–102.

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. (1979). *Annotated bibliography of research on the relationship between quality of teaching and quality of research in higher education.* Los Angeles: Office of Institutional Studies, University of Southern California.

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 Journal of 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. (in press). Students as evaluators of teaching. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies.* Oxford: Pergamon Press.

Marsh, H. W., & Cooper, T. L. (1981). Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research, 16*, 82–104.

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., & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal, 21*, 341–366.

Marsh, H. W., & Overall, J. U. (1979a). Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses." *Research in Higher Education, 10*, 139–147.

Marsh, H. W., & Overall, J. U. (1979b). *Validity of students' evaluations of teaching: A comparison with instructor self-evaluations by teaching assistants, undergraduate faculty, and graduate faculty.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 177 205)

Marsh, H. W., & Overall, J. U. (1980). Validity of students' 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. (1979a). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal, 16*, 57–70.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979b). 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.

Marsh, H. W., Overall, J. U., & Thomas, C. S. (1976). *The relationship between students' evaluations of instruction and expected grade.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 126 140)

Marsh, H. W., Touron, J., & Wheeler, B. (in press). Students' evaluations of university instructors: The applicability of American instruments in a Spanish setting. *Teaching and Teacher Education: An International Journal of Research and Studies.*

Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox Effect. *Journal of Educational Psychology, 74*, 126–134.

Maslow, A. H., & Zimmerman, W. (1956). College teaching ability, scholarly activity, and personality. *Journal of Educational Psychology, 47*, 185–189.

McKeachie, W. J. (1963). Research on teaching at the college and university level. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 1118–1172). Chicago: Rand McNally.

McKeachie, W. J. (1973). Correlates of student ratings. In A. L. Sockloff (Ed.), *Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students* (pp. 213–218). Measurement and Research Center, Temple University.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe*, 384–397.

McKeachie, W. J., Lin, Y-G, Daugherty, M., Moffett, M. M., Neigler, C., Nork, J., Walz, M., & Baldwin, R. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50*, 168–174.

Meier, R. S., & Feldhusen, J. F. (1979). Another look at Dr. Fox: Effect of stated purpose for evaluation,

lecturer expressiveness, and density of lecture content on student ratings. *Journal of Educational Psychology, 71*, 339–345.

Menges, R. J. (1973). The new reporters: Students rate instruction. In C. R. Pace (Ed.), *Evaluating learning and teaching.* San Francisco: Jossey-Bass.

Morsh, J. E., Burgess, G. G., & Smith, P. N. (1956). Student achievement as a measure of instructional effectiveness. *Journal of Educational Psychology, 47*, 79–88.

Murray, H. G. (1976). *How do good teachers teach? An observational study of the classroom teaching behaviors of social science professors receiving low, medium and high teacher ratings.* Paper presented at the Canadian Psychological Association meeting.

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

Murray, H. G. (1983). Low inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 71*, 856–865.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630–635.

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

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology, 72*, 181–185.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to 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.

Parducci, A. (1968). The relativism of absolute judgment. *Scientific American, 219*, 84–90.

Perry, R. P., Abrami, P. C., & Leventhal, L. (1979). Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. *Journal of Educational Psychology, 71*, 107–116.

Perry, R. P., Abrami, P. C., Leventhal, L., & Check, J. (1979). Instructor reputation: An expectancy relationship involving student ratings and achievement. *Journal of Educational Psychology, 71*, 776–787.

Peterson, C., and Cooper, S. (1980). Teacher evaluation by graded and ungraded students. *Journal of Educational Psychology, 72*, 682–685.

Pohlman, J. T. (1975). A multivariate analysis of selected class characteristics and student ratings of instruction. *Multivariate Behavioral Research, 10*, 81–91.

Price, J. R., & Magoon, A. J. (1971). Predictors of college student ratings of instructors (Summary). *Proceedings of the 79th annual convention of the American Psychological Association, 7*, 523–524.

Remmers, H. H. (1963). Teaching methods in research on teaching. In N. L. Gage (Ed.), *Handbook on teaching.* Chicago: Rand McNally.

Rodin, M., & Rodin, B. (1972). Student evaluations of teachers. *Science, 177*, 1164–1166.

Rosenshine, B. (1971). *Teaching behaviors and student achievement.* London: National Foundation for Educational Research.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. W. Travers (Ed.), *Second handbook of research on teaching.* Chicago: Rand McNally.

Rotem, A., & Glassman, N. S. (1979). On the effectiveness of students' evaluative feedback to university instructors. *Review of Educational Research, 49*, 497–511.

Salthouse, T. A., McKeachie, W. J., & Lin, Y. G. (1978). An experimental investigation of factors affecting university promotion decisions. *Journal of Higher Education, 49*, 177–183.

Scott, C. S. (1977). Student ratings and instructor-defined extenuating circumstances. *Journal of Educational Psychology, 69*, 744–747.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 244–271). Beverly Hills, CA: Sage.

Seldin, P. (1975). *How colleges evaluate professors: Current policies and practices in evaluating classroom teaching performance in liberal arts colleges.* Cronton-on-Hudson, New York: Blythe-Pennington, 1975.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretation. *Review of Educational Research, 46*, 407–411.

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17*, 419–433.

Stumpf, S. A., Freedman, R. D., & Aguanno, J. C. (1979). A path analysis of factors often found to be related to student ratings of teaching effectiveness. *Research in Higher Education, 11*, 111–123.

Ward, M. D., Clark, D. C., & Harrison, G. V. (1981, April). *The observer effect in classroom visitation.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Ware, J. E., & Williams, R. G. (1975). The Dr. Fox effect: A study of lecturer expressiveness and ratings of instruction. *Journal of Medical Education, 5*, 149–156.

Ware, J. E., & Williams, R. G. (1977). Discriminant analysis of student ratings as a means of identifying lecturers who differ in enthusiasm or information giving. *Educational and Psychological Measurement, 37*, 627–639.

Ware, J. E., & Williams, R. G. (1979). Seeing through the Dr. Fox effect: A response to Frey. *Instructional Evaluation, 3*, 6–10.

Ware, J. E., & Williams, R. G. (1980). A reanalysis of the Doctor Fox experiments. *Instructional Evalu-*

ation, 4, 15–18.
Warrington, W. G. (1973). Student evaluation of instruction at Michigan State University. In A. L. Sockloff (Ed.), *Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students* (pp. 164–182). Philadelphia: Measurement and Research Center, Temple University.
Webb, W. B., & Nolan, C. Y. (1955). Student, supervisor, and self-ratings of instructional proficiency. *Journal of Educational Psychology, 46,* 42–46.
Whitely, S. F., & Doyle, K. O. (1976). Implicit theories in student ratings. *American Educational Research Journal, 13,* 241–253.
Williams, R. G., & Ware, J. E. (1976). Validity of student ratings of instruction under different incentive conditions: A further study of the Dr. Fox effect. *Journal of Educational Psychology, 68,* 48–56.
Williams, R. G., & Ware, J. E. (1977). An extended visit with Dr. Fox: Validity of student ratings of instruction after repeated exposures to a lecturer. *American Educational Research Journal, 14,* 449–457.

Received December 19, 1983
Revision received June 8, 1984 ■

Editor for *Psychological Bulletin* Named: Search for New Editor Continues

David Zeaman, editor of *Psychological Bulletin*, died on July 19, 1984. Betty J. House, Zeaman's colleague at the University of Connecticut, and one of the journal's associate editors, will complete David Zeaman's term and serve as editor through 1986. Effective immediately, authors should submit manuscripts to:

Betty J. House, Editor
Psychological Bulletin
Department of Psychology U-20, Rm #107
25 Cross Campus Road
Storrs, Connecticut, 06268

APA's Publications and Communications Board is continuing its recently opened search for a new editor. Candidates for the journal editorship must be members of APA and should be available to start receiving manuscripts in early 1986 to prepare for issues published in 1987. The term of editorship is from 1987 through 1992. To nominate candidates, prepare a statement of one page or less in support of each nomination. Submit nominations no later than February 1, 1985 to the chair of the search committee:

Barbara Strudler Wallston
Box 512 Peabody
Vanderbilt University
Nashville, Tennessee 37203

The other members of the search committee are Elizabeth Loftus, Wilbert McKeachie, Paul Mussen, Lyman Porter, and Lee Sechrest.