Research in Higher Education, Vol. 36, No. 5, 1995

THE RELATIONSHIP BETWEEN INDIVIDUAL INSTRUCTIONAL CHARACTERISTICS AND THE OVERALL ASSESSMENT OF TEACHING EFFECTIVENESS ACROSS DIFFERENT INSTRUCTIONAL CONTEXTS

Joseph M. Ryan and Paul D. Harrison

There is an ongoing debate in the student evaluation of teaching literature about whether an overall rating or factorial dimensions of teaching effectiveness should be used in personnel decisions. Marsh and his colleagues have advocated the use of a weighted average approach to computing overall evaluations. A policy-capturing experiment was carried out where students in three different instructional contexts made overall evaluations of hypothetical instructors based on a manipulation of the teaching factors in Marsh's SEEQ. The results indicated (1) amount learned was consistently the most important factor affecting overall evaluations; (2) course difficulty was consistently the least important factor affecting overall evaluations; and (3) there was a strong similarity among the three groups in the relative importance of the various teaching factors in arriving at an overall evaluation. The implications of this research are discussed, as well as directions for future research.

The status of teaching in higher education is currently in a period of change in which teaching is being seen as increasingly more important relative to the research mission of higher education. The Carnegie Foundation for the Advancement of Teaching has called for a reevaluation of the role of teaching in today's university (Boyer and Rice, 1990). The Carnegie Foundation for the Advancement of Teaching and the Irvine group have both recommended a redirected focus for higher education, with more priority being given to teaching and to curriculum and course development.

The renewed emphasis being placed on teaching in the university makes it increasingly more important that effective and credible measures of teaching

Joseph M. Ryan, Arizona State University West, 4701 W. Thunderbird Road, Phoenix, AZ 85069.
Paul D. Harrison, College of Business Administration, University of South Carolina, Columbia, SC 29208.

0361-0365/95/1000-0577$07.50/0

effectiveness be developed and used. Student ratings of teaching are the predominant mechanism employed to evaluate teaching in the university setting (Feldman, 1976a, 1976b, 1977, 1988, 1989a). Some movement is developing in higher education toward using portfolios to evaluate university teaching (Seldin, 1991). As an evaluation method, the emphasis with portfolios is on evaluating teaching with multiple measures of teaching effectiveness collected over time. Even with the use of portfolios, however, student ratings of teaching would remain an important component of any teaching evaluation, and thus the importance of student evaluations will continue.

One of the debates over the use of student evaluations of teaching effectiveness deals with whether an overall rating or factorial dimensions of teaching effectiveness should be used in personnel decisions. Some researchers support the use of overall ratings of teacher effectiveness (Abrami, 1985, 1989) through the use of either global rating items on the evaluation instrument itself or, if necessary, through an unweighted average of the individual factors in a rating instrument. However, other researchers argue that teaching is multidimensional (Marsh, 1977, 1982, 1984; Marsh, Fleiner, and Thomas, 1975; Marsh and Overall, 1980; Marsh, Overall, and Kesler, 1979; Marsh and Hocevar, 1991), and that the individual dimensions should be considered separately in evaluating teaching effectiveness. These researchers assert that if an overall rating of teaching effectiveness is to be used at all, the overall rating should be a weighted average of the individual factors, with the weights being determined by logical and empirical analysis. The major question addressed in this study is: How are students' overall evaluations of teaching effectiveness influenced by various factors or dimensions of teaching?

LITERATURE REVIEW

Overall vs. Multidimensional Evaluations

A major issue frequently debated in teacher evaluation research and practice deals with the relative merits of using an overall evaluation versus a multidimensional profile of teaching effectiveness. This debate is informed by considering the purposes of student evaluations. In general, student evaluations are used to provide (1) diagnostic feedback to faculty about the effectiveness of their teaching, (2) a measure of teaching effectiveness used in personnel decisions (i.e., tenure/promotion and annual evaluation), (3) information for students to use in the selection of instructors and courses, and (4) an outcome or a process description for research on teaching (Marsh, 1984, 1991a; Abrami, d'Apollonia, and Cohen, 1990). For personnel decisions, there is considerable debate as to whether a single score is more useful and appropriate than a profile of scores reflecting multiple dimensions (see Abrami, 1989; Abrami and d'Apollonia, 1991; Cashir and Downey, 1992; Marsh, 1987, 1989, 1991a; Marsh and Hocevar, 1991).

Abrami and his colleagues (Abrami, 1985, 1989; Abrami and d'Apollonia, 1991) favor the use of several global items to evaluate teaching for personnel decisions. They are against the use of separate factor scores for personnel decisions for several reasons. First, they are not convinced that any of the carefully developed, well-validated rating forms represent these dimensions invariantly. They failed to find evidence of the replicability of teaching factors across rating forms (Abrami and d'Apollonia, 1990). Second, they are concerned about the content validity of specific items and some of the dimensions they compromise when ratings are used across a wide variety of courses, instructors, students, and settings, Third, Abrami and his colleagues feel that Cohen's (1981) review of multisection validity studies suggests that many rating dimensions have lower correlations with student learning than with overall instructor ratings. Fourth, less is known about the generalizability of specific factors than overall ratings. Finally, researchers have concerns about the ability of administrators or nonexperts to properly weigh the information provided by factor scores in arriving at a single decision about the quality of good teaching (Franklin and Theall, 1989).

On the other side of the issue, Frey (1973, 1974, 1978) argues strongly that only individual teaching dimensions should be considered, and he excluded global rating items from his endeavor instrument. His subsequent research on two higher-order dimensions (Frey and Flay, 1978) led him to conclude "that personnel decisions should not be made on a single global evaluation measure" (Frey and Flay, 1978, p. 25). Frey's main arguments were that (a) global items are too much influenced by variables that are not associated with effective teaching, (b) global ratings are unduly influenced by student evaluation of teaching effectiveness (SETE) components that are minimally related to student achievement, and (c) it is better to focus on components that are maximally related to a particular criterion than to rely on global items (Marsh, 1991b).

Marsh and his colleagues (Marsh, 1987, 1989, 1991a; Marsh and Hocevar, 1991; Marsh and Dunkin, 1992) have chosen a middle ground between these two positions, recommending the use of both specific dimensions and global ratings. Marsh feels that it is important to differentially weight the specific dimensions. These weights could be constructed on the basis of empirical research findings or ratings of the relative importance of specific components by the department head, a promotions committee, or the instructor (Marsh, 1991b). Marsh further notes that the use of weighted averages is "a compromise that seems consistent with recommendations by Abrami, Frey and myself" (Marsh, 1991b, p. 419).

Weighted-Average Approach to Computing Overall Evaluations

Cashin and Downey (1992) used a multiple regression approach to determine how much of the variance of the criterion variable, overall evaluation, was

accounted for by each of two global predictor variables. The overall evaluation was a weighted average of self-reported learning on a set of 10 general course objectives using the IDEA survey form developed at Kansas State University. The independent variables were a global evaluation of the teacher and course, respectively. Results indicated that each global item individually accounted for more than 50 percent of the variance in the weighted composite criterion measure. The authors interpret these results as supporting the position of Abrami and his colleagues (Abrami, 1985, 1989; Abrami and d'Apollonia, 1991) that global items account for much of the useful information that student ratings provide for making personnel decisions.

Marsh and Roche (1993) evaluated the effectiveness of students' evaluations of teaching effectiveness (SETEs) as a means for enhancing university teaching. Their research used a weighted average approach in arriving at an overall evaluation of teaching effectiveness. The individual dimensions in Marsh's SEEQ were weighted in relative importance by the teacher being evaluated. Thus, Marsh and Roche (1993) were able to construct a teacher-rated importance weighted average of the SEEQ dimensions in arriving at an overall evaluation.

RESEARCH METHODS

From the research reviewed, it is apparent that overall evaluations of teaching effectiveness are used in personnel decisions in universities. While there have been studies that have computed an overall evaluation through the use of a weighted average (Cashin and Downey, 1992; Marsh and Roche, 1993), there have not been any studies that have systematically examined how students weight various teaching factors in arriving at their overall evaluation of teaching effectiveness. This study's objective is to determine the relative importance students in different instructional contexts place on individual teaching factors in assigning a single value as an overall evaluation of teaching effectiveness.

Policy-Capturing Approach

A policy-capturing approach was used to determine the relative importance of various teaching factors to students' overall evaluations of teaching effectiveness. In this approach, a dependent variable or decision variable is defined. In this research, the dependent/decision variable is the students' overall ratings from 1 to 9 for a set of hypothetical instructor profiles. A set of independent variables or cue variables is also defined. In this research, the independent/cue variables are the nine teaching factors in the Student Evaluation of Educational Quality (SEEQ) developed by Marsh and his colleagues (Marsh, 1977, 1982, 1984; Marsh, Fleiner, and Thomas, 1975; Marsh and Overall, 1980; Marsh, Overall, and Kesler, 1979; Marsh and Hocevar, 1991). These are (1) learning,

(2) enthusiasm, (3) organization, (4) group interaction, (5) individual rapport, (6) breadth of coverage, (7) examination fairness, (8) assignments, and (9) course difficulty. The rationale for choosing the SEEQs will be described shortly. Values of 0 = low and 1 = high were assigned to the SEEQs/cue variables and presented to the students. Subjects then assign an overall evaluation to the dependent/decision variable based on the combination of values they were given for the independent/cue variables. Subjects repeat this process over several combinations of values for the cue variables.

The values students assign to the decision variable are regressed onto the values of the cue variables and the resulting regression coefficients indicate the relative influence each cue variable has on the value students assign to the decision variable (Marques, Lane, and Dorfman, 1979). In this research, students assigned an overall rating (1–9) for different hypothetical instructor profiles defined by the various combinations of "low" (0) and "high" (1) assigned to the nine SEEQs.

A complete factorial or fractional factorial design on the independent variables is preferred in policy-capturing research so that the cue variables are independent of each other. When such designs are used, the independent variables are uncorrelated and their regression coefficients can be compared directly with no confounding effects.

There are two important experimental dynamics created by the use of a policy-capturing approach in this research. First, ambiguity is eliminated in the minds of the students about characteristics of the course and instructor because students are told explicitly what the course is like through the assignment of the values high and low to the nine SEEQs. This differs from real classroom settings in which students may be unclear and/or may have differing opinions about course and instructor characteristics. Second, with the policy-capturing approach students' overall evaluations are not influenced by their beliefs about whether they or their instructors are responsible for the characteristics of the course. In real classroom settings, students may attribute the strengths or weaknesses of a course to themselves or to the instructor, and this attribution may influence their overall evaluation. For example, students who believe they have not learned much in a course may attribute this, fairly or unfairly, to the instructor and this attribution might influence the overall rating they give to the instructor.

The teaching factors or dimensions to be used in a policy-capturing approach should be based on a well-established student rating instrument of known validity. The factor structure in the Student Evaluation of Educational Quality (SEEQ) developed by Marsh and his colleagues is well established in the research literature cited earlier. The factor structure in the SEEQ has by far the most extensive and supportive evidence for the validity and usefulness of student evaluations (Howard, Conway, and Maxwell, 1985). The SEEQ has been

found to be reliable and stable (Marsh, 1982, 1983; Marsh and Hocevar, 1984), and relatively valid against a variety of indicators of effective teaching (Marsh, Overall, and Kesler, 1979). It has been validated using a multisection validity research design (Marsh, Fleiner, and Thomas, 1975; Marsh and Overall, 1980) and a multimethod-multitrait research design (Marsh, Overall, and Kesler, 1979).

Task and Experimental Design

Each subject received a set of materials that included (1) a description of each of the factors of teaching effectiveness identified in Marsh's SEEQ, (2) a set of instructions, and (3) a set of 32 hypothetical instructor profiles. The profiles contained the nine factors of teaching effectiveness identified in the SEEQ discussed earlier. A 2^9 1/16 fractional factorial design was used (32 hypothetical instructor profiles), with each of the teaching factors varied at two levels, high or low. Each teaching factor appeared 16 times as "low" and 16 times as "high."

The participants were told to carefully read the description of each of the nine teaching factors in Marsh's SEEQ. The accounting students were then given the following written instructions:

> You are to place yourself in the position of a student enrolled in an accounting course at this university. Each of the cases given below describe a hypothetical accounting instructor at this university. After you read through the information given in each case pertaining to the hypothetical instructor's performance, you are to give a global evaluation pertaining to the overall level of performance of this hypothetical instructor.
>
> Read through each case very carefully, for each case is different from every other case. Once you have finished a particular case, go on to the next one. Do *NOT* turn back to a previous case once you have started working on the next one.
>
> If you have any questions, contact the experimenter. Thank you in advance for your cooperation.

The other two groups of subjects were given the identical instructions, except the word *accounting* was replaced with *education* and *geology*, respectively.

The ratings were assigned on a nine-point scale from 1 (very poor) to 9 (very good). Figure 1 contains the descriptions of the nine teaching factors (based on the SEEQ) given to the subjects, and Figure 2 contains one of the hypothetical teacher profiles used in the study. A random order was used in presenting the descriptions of the teachings factors, and two random orders were used in presenting the 32 instructor profiles.

Subjects

The study was replicated in three different instructional settings in a southeastern state university. Sample one consisted of cost accounting (predomi-

1. Enthusiasm: The instructor was enthusiastic about teaching the course. The instructor's style of presentation held the students interest during class.

2. Individual Rapport: The instructor was friendly towards individual students. The instructor made students feel welcome in seeking help/advice in or outside of class. The instructor was adequately accessible to students during office hours or after class.

3. Learning: Students found the course intellectually challenging and stimulating. Students learned something which they considered valuable. Students have learned and understood the subject materials in this course.

4. Course Difficulty: This compares the difficulty of this course and its workload relative to other courses.

5. Organization: The instructor's explanations were clear. The course materials were well prepared and carefully explained. The proposed objectives were those actually taught so the students knew where the course was going.

6. Breadth: The instructor contrasted the implications of various theories. The instructor presented points of view other than his/her own when appropriate. The instructor adequately discussed current developments in the field.

7. Group Interaction: The students were encouraged to participate in class. Students were invited to share their ideas and knowledge. Students were encouraged to ask questions and were given meaningful answers.

8. Assignments: The required readings/text were valuable. The readings, homework, etc. contributed to appreciation and understanding of subject.

9. Examinations: The feedback on the examinations/graded materials was valuable. The examinations/graded materials tested course content as emphasized by the instructor.

FIG. 1. Description of each factor in the SEEQ.

nantly juniors) students and advanced cost accounting (predominantly seniors) students ($n = 82$). These were advanced students taking a course in their major. Sample two consisted of 53 education graduate students in a survey course in human growth and development. These were education majors in a required course. Sample three was comprised of 94 students in an introductory environmental geology course. These were nonmajors fulfilling a core science requirement that they could have fulfilled with a course in physics, chemistry, or biology. The decision-making exercise was filled out in class. Students were instructed to work at their own pace, and most students finished the exercise in

You are given the following information:

1. The instructor's enthusiasm in this course was . . . . High

2. The instructor's rapport with individual students in this class was . . . . . Low

3. The amount learned in this course was . . . . . Low

4. The difficulty of this course with this instructor was . . . . . High

5. The instructor's organization in this course was . . . Low

6. The breadth of the material covered by the instructor in this class was . . . . . High

7. The instructor's group interaction with the students in this course was . . . . . Low

8. The value of the assignments and the textbook used in this course by the instructor were . . . . . High

9. The fairness of the examinations given by this instructor was . . . . . Low

Overall, how would you rate the classroom performance of this hypothetical instructor **(Please circle one of the numbers below)**?

1	2	3	4	5	6	7	8	9
Very Poor				Average				Very Good

FIG. 2. Hypothetical instructor scenario.

20 to 30 minutes. The 32 hypothetical instructor profiles were presented in two random orders; no order effects were present.

Analysis Procedures

The data analysis proceeded in two phases. In the first phase, two analyses were performed on the data within each of the three instructional contexts. Multiple regression was performed for each student with the overall rating as the dependent measure and the nine SEEQs coded (0,1) as the independent variables for the 32 scenarios to which each student responded. Standardized regression coefficients were derived from these regressions. The purpose of this analysis was to (1) determine if the variation in the SEEQs accounted for most of the variation in the subjects' responses, and (2) determine the relative importance of each SEEQ in making an overall evaluation of teaching effectiveness. A second analysis took advantage of the research design in which each of the nine SEEQs appeared 16 times coded low and 16 times coded high. The mean overall ratings that each student gave under the 16 low and 16 high conditions for each SEEQ were calculated as was the difference between the means for the

low value and high value for each SEEQ. This analysis was performed to determine the relative impact that a high and low value on each SEEQ had on the overall evaluations made by the subjects.

Phase two of the data analyses compared results across the three instructional contexts (groups). The standardized regression coefficients for the three groups were compared using a one-way analysis of variance with post hoc Scheffe contrasts. Accounting students were contrasted to education students to compare advanced students in their majors in two professional schools. Accounting students and education students were then combined and contrasted to geology students who were in an introductory-level core course taken mainly by nonmajors. The same ANOVA procedures were then used to compare students in the three instructional settings on the differences in the mean overall ratings for the SEEQs coded as low versus high.

This research focuses on the data analyses which compare students from the three instructional settings. Results from the inferential procedures are reported as non-significant for probabilities $> .05$. Exact probabilities are reported for $p < .05$. Since nine dependent variables are used in each analysis, however, a test-wise error rate of .005 (e.g., $05/9 = .005$) should be applied to maintain an overall experiment-wise alpha of .05.

RESULTS

The average individual squared multiple correlation (R^2) obtained for the regression analyses for accounting, education, and geology data were .79, .81, and .77 respectively, indicating that the SEEQs accounted for much of the variation in the global ratings. The mean and standard deviation of the beta weights for the nine SEEQs are reported in Table 1 for the three different subject groups. These data are ordered from highest to lowest based on the accounting data.

There is a striking similarity among the orderings of the SEEQ beta weights across the three groups. Indeed, the Spearman rank order correlation of the SEEQ beta weights for accounting and education is .91, for accounting and geology is .97, and for education and geology is .90. In each context, the amount learned has the highest weight and in the accounting and education data it is noticeably larger than exam fairness, which is second. Course difficulty had the lowest relationship to the overall ratings in all three courses. The ANOVA results comparing the beta weights across context are summarized in Table 2. There are significant ($p < .05$) overall differences for amount learned, enthusiasm, assignments, and course difficulty. Contrasts comparing accounting and education students reveal no significant differences. Contrasts combining accounting and education students compared to geology students show significant differences for amount learned, enthusiasm (marginal), assignments, group

TABLE 1. Mean Beta Weights and Standard Deviation (SD) for the Nine Teaching Factors

	Accounting (n = 82)		Education (n = 53)		Geology (n = 94)	
Teaching Factor	Mean	S.D.	Mean	S.D.	Mean	S.D.
1. Amount Learned	.482	.214	.524	.178	.392	.195
2. Exam Fairness	.358	.172	.324	.138	.336	.202
3. Enthusiasm	.228	.134	.275	.123	.291	.176
4. Individual Rapport	.212	.128	.195	.155	.237	.135
5. Organization	.196	.155	.203	.127	.212	.136
6. Assignments	.172	.121	.198	.124	.125	.122
7. Group Interaction	.172	.112	.163	.122	.208	.140
8. Breadth of Materials	.154	.121	.184	.127	.140	.133
9. Course Difficulty	.094	.168	.050	.140	.021	.193

interaction, and course difficulty. The contrast of accounting and education students compared to geology students is significant at $p < .005$ for amount learned and assignments.

The mean values for overall ratings for the low and high SEEQ values and their differences for the nine SEEQs are shown in Table 3. The results shown in Table 3 are similar to the regression analysis results. There is a striking consistency across the three groups in the ordering of the mean overall ratings for the SEEQs coded low, high, and for their difference. In each course, the highest mean overall rating was given when amount learned was coded high and the

TABLE 2. Analysis of Variance Summary with Standardized Beta Weights Compared Across Contexts (Groups)

	All Groups		Acct. vs Educ.		(Acct. + Ed) vs. Geol.)	
Teaching Factor	F-Stat.	Prob.	F-Stat.	Prob.	F-Stat.	Prob.
1. Amount Learned	8.67	.0002	1.42	ns	16.95	.0001
2. Exam Fairness	0.64	ns	1.14	ns	0.05	ns
3. Enthusiasm	3.98	.02	3.11	ns	3.74	.0542
4. Individual Rapport	1.78	ns	0.51	ns	3.34	ns
5. Organization	0.30	ns	0.07	ns	0.46	ns
6. Assignments	6.72	.001	1.46	ns	12.93	.0004
7. Group Interaction	2.84	ns	0.15	ns	5.68	.0180
8. Breadth of Materials	1.95	ns	1.68	ns	2.74	ns
9. Course Difficulty	3.95	.02	2.05	ns	4.81	.0293

TABLE 3. Mean Values and Difference in Mean Values of Global Rating When Each Teaching Factor Is Given as High or Low

	Accounting			Education			Geology		
Teaching Factor	Low Value Mean	High Value Mean	(H-L) Diff. Mean	Low Value Mean	High Value Mean	(H-L) Diff. Mean	Low Value Mean	High Value Mean	(H-L) Diff. Mean
1. Amount Learned	3.73	5.56	1.83	3.35	5.33	1.98	3.97	5.35	1.38
2. Exam Fairness	3.97	5.32	1.35	3.76	4.93	1.17	4.06	5.26	1.20
3. Enthusiasm	4.21	5.08	0.87	3.85	4.85	1.00	4.15	5.17	1.02
4. Rapport	4.25	5.04	0.79	4.00	4.69	0.69	4.26	5.07	0.81
5. Organization	4.28	5.02	0.74	3.99	4.70	0.71	4.31	5.01	0.70
6. Assignments	4.32	4.97	0.65	3.99	4.70	0.71	4.45	4.88	0.43
7. Group Interaction	4.34	4.96	0.62	4.06	4.63	0.57	4.30	5.02	0.72
8. Breadth of Materials	4.37	4.91	0.54	4.03	4.66	0.63	4.43	4.89	0.46
9. Course Difficulty	4.47	4.82	0.35	4.26	4.43	0.17	4.64	4.68	0.04

lowest mean overall rating occurred when amount learned was coded low. At the other end of the scale, the mean overall rating seems relatively unaffected by whether course difficulty is coded low or high. The ANOVA results for the high-low mean global rating differences are reported in Table 4. There are significant overall group effects for amount learned, assignments, and course difficulty. The contrast testing the differences between accounting and education students shows no significant differences between these two groups. Contrasts

TABLE 4. Analysis of Variance Summary with Global Rating Differences for High and Low Values of Teacher Factors Compared Across Instructional Contexts (Groups)

	All Groups		Acct. vs Educ.		(Acct. + Educ.) vs. Geol.	
Teaching Factor	F-Stat.	Prob.	F-Stat.	Prob.	F-Stat.	Prob.
1. Amount Learned	9.32	.0001	1.01	ns	18.48	.0001
2. Exam Fairness	1.24	ns	1.78	ns	0.42	ns
3. Enthusiasm	1.60	ns	1.76	ns	1.10	ns
4. Individual Rapport	1.19	ns	1.57	ns	1.12	ns
5. Organization	0.11	ns	0.07	ns	0.12	ns
6. Assignments	8.22	.0004	0.62	ns	16.38	.0001
7. Group Interaction	2.07	ns	0.36	ns	4.03	.0458
8. Breadth of Materials	2.67	ns	0.84	ns	4.97	0.268
9. Course Difficulty	4.76	.0094	2.09	ns	6.24	.0132

combining accounting and education students compared to geology students show significant differences for amount learned, assignments, group interaction, breadth of material, and course difficulty. The contrasts of accounting and education students compared to geology students are significant at $p < .005$ for amount learned and assignments. The difference between the mean overall ratings for low compared to high for amount learned was 1.38 for the geology students and 1.83 and 1.98 for the accounting and education students respectively. These data suggest that "how much the students learned" influenced the overall ratings of the geology students less than accounting and education students. The difference between the mean overall ratings for low compared to high for assignments was 0.43 for the geology students and 0.65 and 0.71 for the accounting and education students respectively.

The description of assignments given to students in the study was:

> *Assignments*: The required reading/text was valuable. The readings, homework, etc. contributed to appreciation and understanding the subject.

These data suggest that the value of the assignments in terms of contributing to the appreciation or understanding of the subject influenced the overall ratings of the geology students less than accounting and education students.

DISCUSSION

This study was designed to address the following question: How are students' overall evaluations of teaching effectiveness influenced by various factors or dimensions of teaching? This research shows that factorial dimensions are salient (average individual $R^2 > .77$ in all three groups) and differentially influence the overall evaluations that students assign to instructors and courses. A relevant finding in this research is the consistent importance that "amount learned" played in shaping students' overall ratings. Amount learned received the highest beta weight in all three instructional contexts; scenarios in which amount learned was coded "high" received the highest overall ratings and the lowest overall ratings were assigned when amount learned was coded "low." The importance that students placed on amount learned was somewhat surprising to the researchers who observed this finding first based on the accounting students and then on the education students. The researchers speculated that students in these two instructional settings might be especially interested in how much they learned because the hypothetical courses they were rating covered important information and procedures in their majors that they would be expected to know in their professional work. To explore this speculation, students in the third course were added, which included nonmajors for whom the target course would not have direct professional benefits. The ANOVA results with the Scheffe contrasts showed that the accounting and education students were not different from each other on any variables whereas the accounting and

education students combined were significantly different from the geology students in the weights assigned to amount learned and value of assignments.

These results suggest that the instructional context influenced how much students are concerned with their learning. Students may be less concerned about learning in lower-level, core courses compared to upper-level major or graduate major courses.

Another relevant finding is the consistently small beta weight all three subject groups placed on course difficulty. This is consistent with observations made by several researchers suggesting that course difficulty should not enter into an overall evaluation (Abrami, 1985, 1989; Marsh 1991a).

A certain care is required in interpreting the differences in the beta weights across the three groups. Three elements are needed to fully describe the results. First, as was just mentioned, there are statistically significant differences in the beta weights across the three groups and the comparison of students in the two professional schools to the students in the general core course reflects these differences. Second, however, is the finding that the profiles of the beta weights are very similar across the three groups. Third, the differences in the magnitudes of the weights and similarities in the profiles of the weights are more or less critical depending on how the information is to be used. For example, the profile .1, .4, and .2, and the profile .2, .8, and .4 are identical as profiles but would result in very different composite scores if used as weights. Similar profiles that differ by a magnitude of scale could certainly be used to diagnose relative strengths and weaknesses, but using such profiles to calculate a composite could be problematic because two apparently similar profiles could yield significantly different composite totals.

Feldman (1989b) explored the differential importance of various instructional dimensions to student achievement. His analysis indicates that the most important factors in facilitating student achievement were clarity of explanations, preparation and organization of the course, stimulation of students' interests and motivation of students toward reaching high standards, class discussion and openness to the opinions of others, and the professor's availability and helpfulness. Feldman further indicates that, "Those specific instructional dimensions that are the most highly associated with student achievement tend to be the same ones that best discriminate among teachers with respect to the overall evaluations they receive from students" (Feldman, 1989b, p. 619).

The results of this study indicate that students' perceived learning is the most important factor affecting the overall evaluation students give instructors in an experimental setting. This is in contrast with the results of Cashin and Downey (1992), who found that each of two global items, one concerning the instructor, the other concerning the course, accounted for more than 50 percent of the variance in the weighted composite criterion measure. This suggests the need for further research to determine the direction of the causal relationships be-

tween individual instructional dimensions, student learning, and overall evaluations of the instructor.

As a final issue it is important to recognize the experimental nature of this study. Students responded to hypothetical instructors in imagined classrooms. This research approach is invaluable in studying the relationships among important constructs and research questions disambiguated from the sometimes unresolvable confounding variables that influence classroom research in situ. This study can be very useful in setting the directions for further theoretical research and for providing a focus for applied classroom research on faculty evaluation, but the results of this study cannot be generalized directly to classroom practice. The importance of further study in real classroom settings is evident by examining the work of Marsh (1983), who found that overall instructor ratings were more highly correlated with instructor enthusiasm and organization than with learning. This is clearly different from the results reported in the current context and requires further investigation.

Implications

The results of the research reported in this paper raise several questions about the use of student evaluations of teaching effectiveness in personnel decisions. Our results are encouraging in that the students in all three instructional contexts appeared to concentrate on educationally appropriate teaching factors (i.e., amount learned) while largely ignoring irrelevant factors (course difficulty) in arriving at their overall evaluations. Therefore, there is nothing in this study that would disqualify overall evaluations made by students as a reasonable source of information about a teacher's effectiveness. However, administrators must be cognizant of the variation in the relative strength of the various teaching factors across instructional contexts, and within an instructional context across different students.

We recommend that three types of student rating information be used in making personnel decisions: (1) individual teaching dimension ratings, (2) overall evaluations made by students, and (3) a composite weighted average overall evaluation. We believe that since students are already making an overall evaluation (at least with the SEEQ), the composite weighting scheme should be determined by faculty members. The use of a weighted average composite has several distinct features. First, the composite can be constructed differently for different instructional contexts. For example, the weighted average composite should probably be different for a Ph.D. seminar than for an introductory lecture course. Second, with a composite overall evaluation, the teaching factors will be weighted the same for all the students in an individual class. The only variation that will exist between the students in a classroom will be the individual rating differences on the individual teaching factors. Third, the faculty of an

individual department could determine the weighting scheme(s) for the different classes offered, thus giving the faculty some say about how the factors are weighted.

Two issues need to be resolved in using this approach: the relevant dimensions need to be identified and the "appropriate" weights for each dimension need to be obtained. The relevant dimensions to be used should be based on a student rating instrument of known validity, such as Marsh's SEEQ. The appropriate weights to be used for each dimension is more difficult to resolve.

Cashin and Downey (1992) used a weighted composite of the 10 IDEA course objectives in computing an overall evaluation. Essential objectives received double weights, important objectives received single weights, and minor important objectives received zero weights and were dropped from the calculation. For this approach to make sense, one must assume that all essential objectives are of equal importance, all important objectives are of equal lesser importance, and the minor objectives do not matter at all in making an overall evaluation of teaching effectiveness.

Marsh and Roche (1993) used a weighted average approach in arriving at an overall evaluation of teaching effectiveness, with the individual dimensions in Marsh's SEEQ being weighted in relative importance by the individual teacher being evaluated. This approach seems useful, but since most professors have a unique profile on a set of teaching dimensions (Marsh and Bailey, 1993), it will encourage a professor to heavily weight those individual dimensions on which he/she is strong, and assign minimal or zero weight to the other dimensions.

There are several other viable approaches for coming up with a weighted composite of the teaching dimensions. One would be to have the administrator of the academic unit assign the relative weights to the various teaching dimensions. This presumes that the administrator is knowledgeable in this area and has some special expertise to do this. Another method would be to have the faculty in an academic unit fill out a policy-capturing study similar to the one used in this study and to compute a weighted average composite based on weights arrived at by all the faculty in an individual academic unit. A third possibility would be to have the faculty assign 100 points across the individual teaching factors, with some constraints on minimum and maximum points for various factors. With any of these approaches, one needs to keep in mind that different courses at different levels (i.e., an introductory undergraduate course vs. a doctoral seminar) may require different weighting schemes.

The evidence we have presented here does not allow one to conclude that overall ratings made by students or a weighted composite overall rating, with the weights determined by the faculty, are superior. Future research is needed to compare these two overall evaluations. This research should include an external criterion of teaching effectiveness, such as student achievement. Including an external criterion will allow a judgment to be made as to whether an overall

teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education* 30(2): 137–194.

Feldman, K. A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education* 30(6): 583–645.

Franklin, U., and M. Theall (1989). Rating the readers: Knowledge, attitude, and practice of users of student ratings of instruction. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Frey, P. W. (1974, February). The ongoing debate: Student evaluation of teaching. *Change*, pp. 47–49.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education* 9(1): 69–91.

Frey, P. W., and B. R. Flay (1978). *A Cusp Catastrophe Model of Evaluation Person Perception with an Application to Student Ratings of Instruction.* Evanston, IL: Northwestern University.

Howard, G. S., C. G. Conway, and S. E. Maxwell (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology* 77(2): 187–196.

Marques, T. E., D. M. Lane, and P. W. Dorfman (1979). Toward the development of a system for instructional evaluation: Is there consensus regarding what constitutes effective teaching? *Journal of Educational Psychology* 71(6): 840–849.

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(4): 441–447.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology* 52(1): 77–95.

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(1): 150–166.

Marsh, H. W. (1984). Students' evaluations of university teaching dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology* 76(5): 707–754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11(3): 253–388.

Marsh, H. W. (1989). Responses to reviews of students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *Instructional Evaluation* 10: 5–9.

Marsh, H. W. (1991a). Multidimensional students' evaluations of teaching effectiveness: A test of higher-order structures. *Journal of Educational Psychology* 83(2): 285–296.

Marsh, H. W. (1991b). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia (1991). *Journal of Educational Psychology* 83(3): 416–421.

Marsh, H. W., H. Fleiner, and C. S. Thomas (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology* 67(6): 833–839.

Marsh, H. W., and M. Bailey (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education* 64(1): 1–18.

Marsh, H. W., and M. J. Dunkin (1992). Students' evaluations of university teaching: A multidimensional perspective. In John. C. Smart (ed.), *Higher Education: Handbook of Theory and Research*, vol. 8, pp. 143–233. New York: Agathon Press.

Marsh, H. W., and D. Hocevar (1984). The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal* 21(2): 341–366.

Marsh, H. W., and D. Hocevar (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching & Teacher Education* 7(1): 9–18.

Marsh, H. W., and J. U. Overall (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology* 72(4): 468–475.

Marsh, H. W., J. U. Overall, and S. P. Kesler (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal* 16(1): 57–70.

Marsh, H. W., and L. Roche (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal* 30(1): 217–251.

Seldin, P. (1991). *The Teaching Portfolio*. Bolton, MA: Anker Publishing Company, Inc.

Received July 26, 1993.