STUDENT ASSESSMENT OF TEACHING EFFECTIVENESS IN A MULTI-INSTRUCTOR COURSE FOR MULTIDISCIPLINARY HEALTH PROFESSIONAL STUDENTS

NANCY S. PALCHIK
ALPHONSE R. BURDI
GEORGE E. HESS
TED E. DIELMAN
University of Michigan Medical School

Student evaluations of teaching effectiveness in a large multi-instructor human anatomy course for students from four professional programs (nursing, dental hygiene, pharmacy, and physical education) were examined over a three-year period to assess the influence of professional program on student ratings of instruction. In spite of wide differences in mean achievement, students in the four professional groups were relatively consistent in their differential evaluations of the three course instructors and in their evaluations of learner motivational and course performance dimensions of instruction. Results supported previous findings regarding both (a) the internal consistency and interrater reliability of student evaluations of instruction in a course format increasingly being used in health professional education and (b) the need for multiple assessments of instruction over time before generalizations can be made about an individual instructor's teaching skill.

AUTHOR'S NOTE: An earlier version of portions of this article was presented at the 1986 annual meeting of the American Educational Research Association, San Francisco. The authors would like to thank Robert F. Dedrick, Amy T. Butchart, and Pamela C. Campanelli for their helpful comments and assistance with the data analysis, Deborah M. Ottaway for her help in the preparation of the manuscript, and the University of Michigan Center for Research on Learning and Teaching for use of the Instructor Designed Questionnaire System in constructing the evaluation instruments. Address reprint requests to Nancy S. Palchik, Department of Postgraduate Medicine and Health Professions Education, University of Michigan Medical School, G-1116 Towsley Center, Ann Arbor, Michigan 48109-0201.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 11 No. 1, March 1988 55-73

Student ratings of instruction are one of the primary measures increasingly used by those in institutions of higher education for evaluating the teaching accomplishments of their faculties. Student ratings are being used not only by individual faculty members for monitoring their teaching skill and improving their course offerings, but also by administrative and promotion and tenure committees for making more informed decisions about the teaching accomplishments of the faculty. But as institutions have increased the use of student ratings to evaluate instructor effectiveness, those using these ratings have come to focus more of their attention on the reliability and validity of the ratings. Of particular concern to faculty and administrative decision makers has been the extent to which student evaluations are potentially influenced by factors other than those related to the actual quality of the teaching itself (i.e., by student and course characteristics over which an instructor may have only minimal control). An additional concern has been the extent to which student ratings obtained over time are sufficiently sensitive to detect improvements or declines in an individual faculty member's teaching effectiveness.

Despite some skepticism, research on the reliability of student ratings has nonetheless found that these evaluations are generally moderately reliable, consistent within a given class, and relatively stable over time (Costin et al., 1971; Irby et al., 1977; Kulik and McKeachie, 1975; McGaghie, 1975; Dielman and Horvatich, 1985). Even within a multi-instructor course, the research of Irby et al. (1977) found that students are able to identify differences among teachers and that there is a relatively high correlation of ratings made immediately after an instructor has completed the lecture(s) and again at the end of the course. However, when one turns to the literature on the validity of student ratings, one finds that research has been hampered by the absence of universally accepted criteria of effective teaching (Dowell and Neal, 1982; Marsh and Overall, 1980). In general, studies of the validity of student ratings have compared these ratings to other measures used to assess instruction, namely, peer ratings, alumni ratings, and achievement measures (Centra, 1980; Doyle and Crichton, 1978; Marsh and Overall, 1980). In these studies, correlations

between student and peer ratings have been generally high. Correlations between student and alumni ratings have also found an essential agreement in overall assessments of teaching effectiveness.

Ultimately, of course, effective teaching should be reflected by gains in student learning. Although student learning has typically been measured by examination scores and final course grades, these indices may not accurately reflect the full influence of instructor skill on student performance. Attempts to correlate student ratings of instruction with examination scores and final course grades have generally yielded only low or moderate relationships (Canaday et al., 1978; Doyle, 1975; Kulik and McKeachie, 1975; Mendelson et al., 1978). These low-to-moderate correlations suggest that factors other than instruction may be affecting student achievement (e.g., student motivation, prior ability, or interest in the subject matter). Alternatively, these low-to-moderate correlations may suggest that students who achieve at different levels are nevertheless rating particular aspects of instruction in a similar manner. In this regard, several investigators have noted that the dimensional structure of teaching effectiveness and of most rating instruments has not been systematically taken into account (Cohen, 1981; Marsh and Overall, 1980; McGaghie, 1975; McKeachie et al., 1971; Rippey, 1975; Slotnick and Durkovic, 1975). A dimension focusing on "course difficulty," for example, might not relate to student achievement to the same extent as a dimension focusing on "instructor skill."

A major focus for research on student evaluations of instruction has concerned the potential influence of instrument-, course-, and student-related factors on the evaluations an instructor receives (Abrami et al., 1982; Centra, 1980; Cranton and Smith, 1986; Doyle and Whitely, 1974; Kulik and McKeachie, 1975; Marsh and Overall, 1980). These are variables over which an individual instructor may have little control. The identification of these factors and the nature of their potential influence on student ratings is, therefore, essential if comparisons among instructors are to yield meaningful results. Alternatively, the absence of an influence of student-, course-, and instrument-related factors on

the evaluations an instructor receives would tend to support the validity of the student ratings as measures of teaching effectiveness.

The present three-year study was designed to examine the influence of student professional programs on student ratings of instruction in a required multi-instructor lecture course. This type of course format, characterized by more than one instructor and by students from diverse undergraduate programs, has become increasingly common in higher education and specifically in health professions education. The purpose of the study was to identify the extent to which meaningful evaluations of instruction could be obtained in a large class of students from diverse professional programs and with wide differences in ability levels. Specifically, the investigators sought to determine the extent to which: (1) the students' professional programs influenced their ratings of instruction in the course; (2) the students' programs differentially affected their evaluations of particular instructors; (3) differences in achievement associated with different professional programs were also reflected in differences in the students' assessments of instruction; and (4) the evaluations of an individual instructor's teaching effectiveness changed or, alternatively, remained stable over time. In addition, the dimensional structure of the evaluation instrument was examined to determine if the students' professional program or achievement level selectively influenced their evaluations of particular dimensions of instruction.

METHODS

SUBJECTS

During each of the three study years, student ratings of instruction were collected on the last day of an undergraduate human anatomy course offered by the University of Michigan Medical School and required for professional students in the dental hygiene, pharmacy, physical education, and nursing

programs. The course was lecture-oriented and taught by three senior faculty members in the Department of Anatomy and Cell Biology. Of the 743 undergraduate students who completed the course over the three-year period, 339 were nursing students (46%), 172 were physical education students (23%), 115 were dental hygiene students (15%), and 117 were pharmacy students (16%).

INSTRUMENTS

Students were asked to evaluate teaching effectiveness on a 25-item, 5-point Likert-type scale ranging from "strongly agree" (5) to "strongly disagree" (1). The 25 items were selected from the Instructor Designed Questionnaire (IDQ), a cafeteria-type evaluation system developed by the University of Michigan Center for Research on Learning and Teaching. The IDQ contains a menu of 141 potential items that instructors use to design their own evaluation forms. The 25 items in the present study were selected from the IDQ by the course director and course faculty with the assistance of faculty in the Medical School's Office of Educational Resources and Research. Ten items were selected to focus specifically on the teaching skill of the individual instructors in the course, 10 items were student- or course-related, and 5 items were university-wide core items. Student achievement was measured by the percentage correct on two interim course examinations and a final examination. The final examination was administered approximately one week after the students completed their course evaluations.

PROCEDURES

Each student completed an evaluation instrument for each instructor in the course. Each of these instruments contained the 10 instructor-related items and 5 core items. The evaluation instrument for the course director also contained the 10 student- and course-related items. The evaluation forms were completed anonymously. However, students were asked to indicate their professional program on each instrument. The instructors were

given their mean ratings on each item after each study year.

The 743 students who completed the course over the three-year study returned 74% of their evaluation forms. The overall response rates by study year were 76% for year 1, 68% for year 2, and 79% for year 3. These response rates were influenced both by the number of students who attended class on the day of the evaluation and by the number of those students attending class who returned their evaluation forms. The overall response rates by student group were dental hygiene, 73%; nursing, 81%; pharmacy, 78%; physical education, 59%. Response rates were relatively consistent across the three instructors and across study years, with the exception of dental hygiene students, who had a lower response rate during year 2 (51%). Response rates for each student group by study year can be derived from the numbers of students completing the course and the numbers of students completing evaluation forms given in Tables 2 and 4, respectively.

Student responses to the 25 items were factor analyzed by the principal axis procedure using squared multiple correlations as the initial estimates of communalities. Kaiser's unity rule was applied as the criterion for determining the number of factors to retain (Guertin and Bailey, 1970). The initial solution was rotated to the Varimax criterion. Index scores for each of the factors were calculated by averaging the ratings of the items that had their highest loadings on each of the factors. Items were included in only one index. If a student omitted more than two items for a given factor, an index score was not computed for the student on that factor. Cronbach alpha coefficients (Cronbach, 1951) were computed to determine the internal consistency of scores on the items that loaded on each of the resulting factors. The course director's evaluation form was used in these preliminary analyses, as this form contained student responses to all 25 assessment items.

Intraclass correlations were used as estimates of interrater reliability for the "instructor skill" index for year one. The intraclass correlation was calculated as $(MSB - MSW)/[MSB + (C - 1)(MSW)]$, where MSB was the mean square for the "instructor skill" factor, MSW was the mean square for error, and C was the

number of raters. Intraclass correlations were computed for the total class and for each of the four groups of students considered separately. They were used to obtain overall estimates of the extent of agreement among the student ratings of each instructor. These estimates may be an underestimate to the extent that student ratings are correlated across instructors, as this source of variance was included in the error mean square because of the inability to match student ratings across instructors. Intraclass correlations could not be calculated for the indices that pertained only to the course.

A three-way analysis of variance (instructor by student professional group by year) was conducted to test the influences of student professional program, course instructor, and study year on student evaluations of instructor skill. Because student ratings were collected anonymously, a repeated measures analysis was not possible for the three-way analysis of variance. This generally has the effect of increasing the standard error, as the correlation among student ratings of instructors would have been subtracted from the standard error in a repeated measures analysis. Two-way analyses of variance (student professional group by year) were conducted to test the main effects and interactions of professional program and course year on student achievement and on student ratings of motivational and course performance dimensions of instruction. Scheffé post hoc comparisons were used to evaluate the differences between means when the F-test in the analysis of variance indicated overall significance.

RESULTS

DIMENSIONS OF THE ASSESSMENT INSTRUMENT

The factor analysis resulted in three factors with eigenvalues greater than 1.0, accounting for 50% of the total variance. The fourth eigenvalue was .74. The 11 items that loaded most highly on Factor I (factor loadings greater than.53) included all 10 of the

TABLE 1
Items Used to Assess Instructor Skill, Student Motivation, and Course-Performance Dimensions of Instruction

Instructor Skill
Overall, the instructor is an excellent teacher.
The instructor gives clear explanations.
The instructor makes good use of examples and illustrations.
The instructor stresses important points in lectures or discussions.
The instructor is enthusiastic.
The instructor puts material across in an interesting way.
The instructor seems to enjoy teaching.
The instructor appears to have a thorough knowledge of the subject.
The instructor is not confused by unexpected questions.
The instructor teaches near the class level.
The instructor seems well prepared for each class.
Motivation
Overall, this is an excellent course.
The instructor motivates me to do my best work.
I had a strong desire to take this course.
I learned a good deal of factual material in this course.
I gained a good understanding of concepts/principles in this field.
I deepened my interest in the subject matter of this course.
I developed enthusiasm about the course material.
Course Performance
I feel that I am performing up to my potential in this course.
The amount of work required is appropriate for the credit received.
The amount of material covered in the course is reasonable.
Exams are reasonable in length and difficult.
The grading system was clearly explained.
The grading system was a fair assessment of my ability in this course.

items initially selected from the IDQ to assess instructor skill and one instructor-related university core item ("Overall, the instructor was an excellent teacher"). Factor I has been given the label "instructor skill" for convenience of discussion (see Table 1). These 11 items loading on Factor I were completed for each of the three instructors individually.

Seven items loaded most highly on Factor II (factor loadings greater than .47). These items tended to focus on the learner's motivation, interest, understanding, and enthusiasm for course material. Factor II has been labeled "motivation" (Table 1). Six

items had their highest factor loadings on Factor III, which was labeled "course performance." Five of these items had factor loadings greater than .47. One item ("The grading system was clearly explained") had a factor loading of only .22. The items that loaded on Factor III tended to be concerned with the reasonableness of course requirements and expectations for student performance (Table 1). One item that dealt with outside reading assignments did not load onto any of the factors and was excluded from further analyses. To determine the internal consistency of scores, Cronbach alpha coefficients were computed. The Cronbach alpha coefficients computed for each of the three index scores derived from the factors were .91, .87, and .76, respectively. These coefficients indicate that the items within each of the three indices were relatively homogeneous.

The interrater reliabilities (as measured by the intraclass correlations) for the "instructor skill" index were: .38 for the total student group; .60 for dental hygiene students; .38 for nursing students; .20 for pharmacy students; and .30 for physical education students. The intraclass coefficient reaches a maximum of one when student ratings within each instructor are identical and ratings differ only between instructors. Thus these interrater reliabilities indicate a moderate agreement among student ratings of the course instructors when students in the entire class were considered, and also when students in each of the four programs were considered separately.

STUDENT ACHIEVEMENT

The four groups of professional students performed at substantially different levels in the course (see Table 2). The two-way analysis of variance to test the effects of student professional group and course year on cumulative performance in the course resulted in a significant main effect of professional group ($p < .01$). The mean performance of pharmacy students was significantly higher and the mean performance of physical education students was significantly lower than the mean performance of students in each of the other professional groups ($p < .01$). The

performance means of the dental hygiene and nursing students were not significantly different. There was no significant effect of course year on student achievement and no significant student professional group by study year interaction.

INSTRUCTOR SKILL

As shown in Table 3, the three-way interaction of instructor by student professional group by study year was not statistically significant. However, the three two-way interactions and all three main effects reached at least the .001 level of significance. The number of student raters, mean ratings, and standard deviations for the "instructor skill" index are shown in Table 4 by study year for each of the three instructors as rated by each of the student groups. These same results are displayed graphically in Figure 1 for each of the three study years considered separately and for all three study years combined.

Mean ratings of "instructor skill" combined over all three years for the three instructors by student program are shown in Figure 1d. As can be seen in the figure, the four groups of students were similar in their ratings of the three instructors. Although the interaction of instructor and student program was statistically significant, this interaction was due to relatively localized differences among the student groups. Instructor 1 received the highest mean ratings from all four groups of students. Differences in the mean ratings of instructor 1 and instructor 2 were statistically significant for all four student groups ($p < .01$), and differences in the mean ratings of instructor 1 and instructor 3 were significant for all but pharmacy students ($p < .01$). Although nursing and dental hygiene students rated instructor 3 more highly than they rated instructor 2 ($p < .01$), there were no significant differences in the overall mean ratings that these two instructors received from either pharmacy or physical education students. When mean ratings for each of the three instructors were compared across professional groups, the students' professional program showed no significant influence on mean ratings of either instructor 1 or instructor 2. Pharmacy students gave instructor 3 a higher mean rating than this instructor

TABLE 2
Student Achievement (Percentage Correct) by Year and Program

	Year 1			Year 2			Year 3		
Student Program	N	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.
Dental Hygiene	47	77.7	7.8	37	78.2	9.9	31	79.8	7.4
Nursing	122	75.8	9.6	118	80.0	8.8	99	75.9	10.0
Pharmacy	35	86.0	8.2	39	86.0	6.8	43	85.3	7.1
Physical Ed.	61	68.0	12.8	50	69.3	12.7	61	71.4	13.2

TABLE 3
Summary of Analysis of Variance Results: Instructor Skill

Source	Degrees of Freedom	Mean Square	F
Instructor (I)	2	55.54	168.54***
Student (S)	3	2.65	8.04***
Year (Y)	2	5.25	15.92***
IS	6	1.36	4.11***
IY	4	18.31	55.57***
SY	6	1.61	4.88***
ISY	12	0.46	1.38
Error	1619	0.33	

***$p < .001$.

received from physical education students ($p < .05$). However, none of the other comparisons of mean ratings for instructor 3 were statistically significant.

Figure 2 shows the interaction of instructor and study year. This figure was derived from the means contained in Table 5. As can be seen in the figure, changes in the mean ratings of the three course instructors over the three study years were not the same for each instructor. During each consecutive study year, the mean evaluations of instructor 3 improved significantly ($p < .01$). The mean evaluations of instructor 2 essentially remained the same during the first two study years, but significantly declined during the third study year ($p < .01$). The mean evaluations of instructor 1 were consistently high and did not change significantly over the three-year period. A comparison of the ratings of the three instructors by study year in Figure 2 shows significant differences

TABLE 4
Instructor Skill Ratings by Year, Student Program, and Instructor

	Year 1			Year 2			Year 3		
	N	Mean Rating	s.d.	N	Mean Rating	s.d.	N	Mean Rating	s.d.
Instructor 1									
Dental Hygiene	38	4.78	.25	18	4.32	.49	27	4.80	.23
Nursing	100	4.56	.41	91	4.50	.37	84	4.61	.45
Pharmacy	26	4.31	.43	28	4.23	.58	37	4.71	.37
Physical Ed.	37	4.45	.55	27	4.34	.50	38	4.43	.45
Instructor 2									
Dental Hygiene	37	3.91	.55	20	3.81	.70	25	3.25	.60
Nursing	98	3.93	.60	92	4.26	.63	86	3.66	.63
Pharmacy	25	3.96	.63	28	3.86	.70	38	3.85	.62
Physical Ed.	36	3.94	.65	28	3.77	.71	36	3.40	.86
Instructor 3									
Dental Hygiene	41	3.60	.62	19	3.57	.70	26	4.43	.47
Nursing	101	3.63	.74	89	3.99	.57	85	4.48	.50
Pharmacy	26	3.69	.67	29	4.13	.49	37	4.69	.36
Physical Ed.	38	3.46	.95	28	3.91	.59	36	4.36	.55

among the ratings of the three instructors during the first study year ($p < .01$), with instructor 1 receiving the highest ratings and instructor 3 receiving the lowest ratings. During the third study year, however, instructor 2 received the lowest ratings ($p < .01$), while the ratings of instructors 1 and 3 were not significantly different. A similar pattern is shown in Figure 1 for each of the four student groups in their relative ratings of the three course instructors over time. During the first study year (Figure 1a), each of the four groups of students rated instructor 1 more highly than instructor 2, and instructor 2 more highly than instructor 3. During the third study year (Figure 1c), instructor 2 was rated lowest by all four student groups and instructor 3 was rated almost as highly as instructor 1.

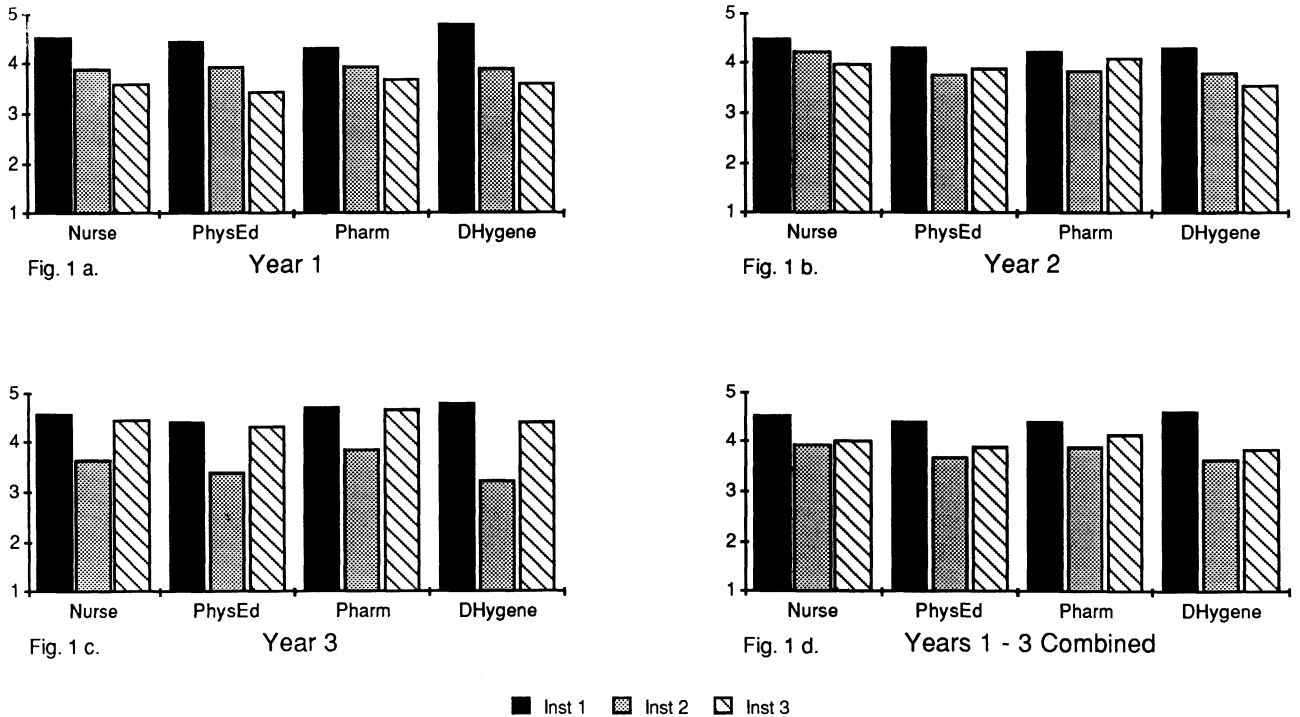

Figure 1: Mean Ratings of Instructor Skill by Study Year, Instructor, and Student Program

TABLE 5
Instructor Skill Ratings by Year and Instructor

	Year 1			Year 2			Year 3		
	N	Mean Rating	s.d.	N	Mean Rating	s.d.	N	Mean Rating	s.d.
Instructor 1	201	4.55	0.44	164	4.41	0.45	186	4.62	0.42
Instructor 2	196	3.93	0.60	168	4.06	0.69	185	3.59	0.70
Instructor 3	206	3.60	0.75	165	3.95	0.59	184	4.49	0.49

STUDENT MOTIVATION AND COURSE PERFORMANCE

The means, standard deviations, and number of ratings for the "student motivation" index by student professional group and study year are shown in Table 6. The two-way (student professional program by year) analysis of variance of scores on this index indicated a significant interaction of student program by year ($F = 2.57$, $p < .05$) and significant main effects ($F = 10.6$ for student program, $F = 5.81$ for study year, $p < .01$ in each case). The main effect of student program was primarily due to the ratings on this index by nursing students being somewhat higher than the other groups during years one and two. The main effect of year was due to somewhat different overall means by study year. The interaction effect was due to differing patterns of changes by study year, depending on the student group. The Scheffé tests for differences among the individual means were not significant, however.

The means, standard deviations, and number of ratings for the "course performance" index are also presented by student professional program and study year in Table 7. The two-way analysis of variance (student professional program by year) resulted in a significant program by year interaction ($F = 3.26$, $p < .01$) and a significant main effect of student program ($F = 6.27$, $p < .01$). As with student ratings on the "student motivation" index, none of the Scheffé comparisons between individual means on the "course" performance index were statistically significant.

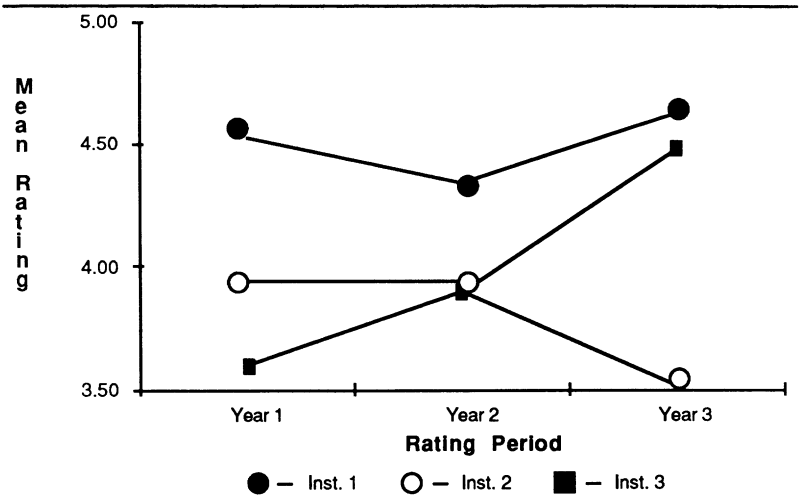

Figure 2: Mean Student Ratings of Instructor Skill for Each Course Instructor by Study Year

DISCUSSION

The results of the present study suggest that the nature of students' professional program does not appear to have a major influence on student ratings of instruction in a large multi-instructor lecture course. In spite of wide differences in the average achievement of students in the four professional groups studied, students from the different professional programs were relatively consistent in their evaluations of instruction in the course and in their assessments of the relative effectiveness of the three course instructors. Although there were some differences among the four groups of students on each of the three dimensions of instruction studied, these differences were not consistent across either student groups or course years. The differences also did not appear to relate in any systematic way to the achievement levels of the four professional groups.

Although the student characteristics examined in the present study did not appear to systematically affect their ratings of instruction, there were enough differences among the four

TABLE 6
Student Motivation Ratings by Year and Student Program

	Year 1			Year 2			Year 3		
Student Program	N	Mean Rating	s.d.	N	Mean Rating	s.d.	N	Mean Rating	s.d.
Dental Hygiene	36	4.00	.48	18	3.78	.57	27	4.25	.61
Nursing	103	4.28	.51	92	4.34	.45	86	4.28	.61
Pharmacy	26	3.85	.74	28	3.86	.60	36	4.27	.42
Physical Ed.	38	4.05	.80	27	4.12	.64	39	4.11	.56

professional groups for instructors to be cautioned to follow Centra's advice to inspect the responses of subgroups of students for identifiable patterns that might suggest that a segment of the class is being slighted or having difficulty with a particular portion of the course (Centra, 1980). This is particularly important when student ratings are being used for purposes of course improvement.

The present study provides evidence that supports previous findings concerning the internal consistency and interrater reliability of student ratings of instruction (e.g., Dielman and Horvatich, 1985). It further suggests that, even in a large and diverse multi-instructor course, meaningful evaluations of instruction can be obtained. Despite differences in professional program and mean achievement level, students in the present study were able to identify differences among instructors with relative consistency. These results are particularly important since many colleges and universities are increasing their efforts to evaluate teaching effectiveness (in addition to research effectiveness). At the same time that they are increasing their efforts to evaluate teaching effectiveness, they are also tending to use larger classes with larger numbers of faculty for each curricular offering. However, changes in the performance of individual instructors over the course of the three-year study also underscore the importance of obtaining multiple assessments of teaching effectiveness over time. Multiple assessments of teaching effectiveness are particularly important if the ratings of individual instructors are to be used for administrative decision-making. Although

TABLE 7
Course Performance Ratings by Year and Student Program

	Year 1			Year 2			Year 3		
Student Program	N	Mean Rating	s.d.	N	Mean Rating	s.d.	N	Mean Rating	s.d.
Dental Hygiene	38	3.81	.62	16	3.43	.67	27	3.93	.54
Nursing	99	3.63	.61	90	3.86	.62	86	3.57	.75
Pharmacy	27	3.82	.71	29	3.79	.52	38	3.95	.68
Physical Ed.	38	3.61	.72	27	3.30	1.00	37	3.41	.66

instructor 1 consistently received the highest mean ratings over the three-year study, the ratings of instructor 3 showed significant improvement from the first course-year, when this instructor received the poorest ratings, to the third course-year, when the mean rating of instructor skill was as high for instructor 3 as for instructor 1. If teaching effectiveness had been assessed only during a single course-year, this change in the performance of instructor 3 would not have been detected.

One methodological limitation of this study is that the interrater reliability coefficients and the analysis of the ratings of instructor skill ideally should have been based on a repeated measures analysis of variance, as each student within each year of the study rated all three instructors. This was not possible because of the lack of identifying data for the students. Had repeated measures analyses been possible, the interrater reliabilities and the main and interaction effects involving the instructor factor would probably have been higher, as the standard errors would have been reduced to the extent that ratings of the three instructors were correlated. Another methodological limitation that could have resulted in a different interpretation of the results is the relatively small number in some of the subgroups in the analysis of differences by student program across the three years of the study. Although significant main effects of student program and significant interactions of student program and study year were found with respect to both the motivation and course performance indices, the Scheffé comparisons were not significant. These results were tangential to the goals of the study,

however, in that significant results would have indicated that students in the different programs rated their motivation and course performance differently.

One possible interpretation of the comparative stability of student ratings in the present study is that the assessment of instruction was not sensitive enough or did not ask questions that might have detected more substantial differences among the student groups. Although such lines of inquiry will require further research, the sensitivity of the assessment instrument was supported by the consistency of the differences reported in ratings of the three instructors that were obtained across professional groups and over time. Comparisons of the student ratings obtained in the present study with university-wide norms on these same items also supported the sensitivity of the instruments to detect differences in teaching effectiveness. An additional question that must await the collection of additional data has to do with the extent to which the results obtained in the present study may have been influenced by the relatively positive evaluations that this course and its instructors received. The challenge remans to determine whether the comparatively favorable ratings of this course overshadowed differences that night have emerged in a course that was less highly rated by students.

REFERENCES

ABRAMI, P. C., R. P. PERRY, and L. LEVENTHAL (1982) "The relationship between student personality characteristics, teacher ratings, and student achievement." J. of Educ. Psych. 74: 111-125.

CANADAY, S. D., M. A. MENDELSON, and J. H. HARDEN (1978) "The effect of timing on the validity of student ratings." J. of Medical Education 53: 958-964.

CENTRA, J. A. (1980) Determining Faculty Effectiveness: Assessing Teaching, Research and Service for Personnel Decisions and Improvement. San Francisco: Jossey-Bass.

COHEN, P. A. (1981) "Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies." Rev. of Educ. Research 51: 28 1-309.

COSTIN, F., W. T. GREENOUGH, and R. J. MENGES (1971) "Student ratings of college teaching: reliability, validity, and usefulness." Rev. of Educ. Research 41: 511-535.

CRANTON, P. A. and R. A. SMITH (1986) "A new look at the effects of course characteristics on student ratings of instruction." Amer. Educ. Research J. 23: 117-128.

CRONBACH, L. J. (1951) "Coefficient alpha and the internal structure of tests." Psychometrika 16: 297-334.

DIELMAN, T. E. and P. K. HORVATICH (1985) "Interrater reliability and internal consistency of student and staff ratings of medical instruction." Presented at the Annual Meeting of the American Educational Research Association, Chicago (ERIC Doc. No. ED 254 554, March, 1985).

DOWELL, D. A. and J. A. NEAL (1982) "A selective review of the vaidity of student ratings of teaching." J. of Higher Education 53: 51-62.

DOYLE, K. O. (1975) Student Evaluation of Instruction. Lexington, MA: D. C. Heath.

DOYLE, K. O. and L. I. CRICHTON (1978) "Student, peer, and self evaluations of college instructors." J. of Educ. Psychology 70: 815-826.

DOYLE, K. O. and S. E. WHITELY (1974) "Student ratings as criteria for effective teaching." Amer. Educ. Research J. 11: 259-274.

GUERTIN, W. H. and J. P. BAILEY (1970) Introduction to Modern Factor Analysis. Ann Arbor, MI: Edwards Brothers.

IRBY, D. M., N. P. SHANNON, M. SCHER, P. PECKHAM, G. KO, and E. DAVIS (1977) "The use of student ratings in multiinstructor courses." J. of Medical Education 52: 668-673.

KULIK, J.A., and W. J. McKEACHIE (1975) "The evaluation of teachers in higher education," in F. N. Kerlinger (Ed.) Review of Research in Education, Vol. 3. Itasca, IL: Peacock.

MARSH, H. W. and J. U. OVERALL (1980) "Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria." J. of Educ. Psych. 72: 468-475.

McGAGHIE, W. C. (1975) "Student and faculty ratings of instruction: another look." J. of Medical Education 50: 387-389.

McKEACHIE, W. J., Y. LIN, and W. MANN (1971) "Student ratings of teacher effectiveness: validity studies." Amer. Educ. Research J. 8: 435-445.

MENDELSON, M. A., S. D. CANADAY, and J. H. HARDEN (1978) "The relationship between student ratings of course effectiveness and student achievement." Medical Education 12: 199-204.

RIPPEY, R. M. (1975) "Student evaluation of professors: are they of value?" J. of Medical Education 50: 951-958.

SLOTNICK, H. B. and R. G. DURKOVIC (1975) "Dimensions of medical students' perceptions of instruction." J. of Medical Education 50: 662-666.