DOCUMENT RESUME

ED 138 165 HE 008 828

AUTHOR Overall, Jesse U.; Marsh, Herbert W.
TITLE The Relationship Between Students' Evaluations of Faculty and Instructional Improvement.
PUB DATE Jun 76
NOTE 19p.; Paper presented at the International Conference on Improving University Teaching (3rd, Newcastle-upon-Tyne, England, June 8-11, 1976)
AVAILABLE FROM Dr. J. U. Overall, Director, Programs in Administration, California State College--Domingues Hills, 1000 East Victoria, Domingues Hills, California 90747

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; *Affective Behavior; College Students; College Teachers; Experiments; Feedback; *Higher Education; *Instructional Improvement; Student Attitudes; *Student Evaluation of Teacher Performance; *Teacher Improvement; *Teaching Quality

ABSTRACT

A random half of the instructors of an introductory course in computer programming were given feedback from students' evaluations of instructional effectiveness conducted at the middle of the term. The impact of the feedback was assessed against three criteria: (1) change in students' evaluations between mid-quarter and end-of-quarter; (2) students' results on a standardized final examination; and (3) affective consequences (application of the subject matter, and plans to pursue the subject further). The results show a positive effect of student evaluations according to each of the three criteria. (Author/MSE)

THE RELATIONSHIP BETWEEN STUDENTS' EVALUATIONS OF FACULTY AND INSTRUCTIONAL IMPROVEMENT[1]

Jesse U. Overall
California State College

Herbert W. Marsh
University of Southern California

A random half of the instructors of an introductory course in computer programming were given feedback from students' evaluations of instructional effectiveness conducted at the middle of the term. The impact of the feedback was assessed with three different sets of criteria; Change in Students' Evaluations between Mid-Quarter and End-of-Quarter, Student Achievement (results on a standardized final examination), and Affective Consequences (application of subject matter and plans to pursue the subject further). The results provide strong support for the favorable impact of feedback from students' evaluations on each of the three sets of criteria.

[1]Paper presented at the Third International Conference on Improving University Teaching held in Newcastle-upon-Tyne, England, June 8-11, 1977.

THE RELATIONSHIP BETWEEN STUDENTS' EVALUATIONS OF FACULTY AND INSTRUCTIONAL IMPROVEMENT

Jesse U. Overall
California State College

Herbert W. Marsh
University of Southern California

Students' evaluations of instructional effectiveness have become increasingly prevalent at institutions of higher education. Students seek access to such evaluations for guidance in selecting courses and instructors; faculty, concerned with effective teaching, are interested in how their instruction is perceived by students; and, administrators, faced with steady-state or shrinking resources, seek effective and efficient means for making personnel decisions.

The comparative ease with which objective student evaluations of instruction may be collected, analyzed and reported is a fundamental reason for this popularity. One common and representative means of obtaining these evaluations is based on student completion of assessment questionnaires during 15 to 20 minutes of class time; the response sheets are then optically scanned at the rate of 500 per minute, and the raw data are collected on magnetic tape; finally, a computer reads the tape, summarizes the responses, and "orders' its printer to provide an easy-to-read, standardized summary of results for all those evaluated. Summaries are then quickly available for students, faculty and administrative review. Existing technology thus makes possible the efficient production of such information at even the largest of universities.

If institutions do not possess the necessary equipment, then the Center for Faculty Evaluation and Development (CFED) provides its Instructional Development and Assessment (IDEA) system, or Educational Testing Service (ETS) makes available the Student Instructional Report (SIR) for evaluation purposes. Schools that use the IDEA system or the SIR may request scoring, interpreting and summarizing services from ETS and CFED for a nominal charge.

While acknowledging the progress that has been made in efficiently accommodating the large amounts of individual data that must be collected and processed, members of the academic community have wondered whether--in our fascination with developing the technology for the process--we have become distracted from the real issues surrounding the product. Are we so certain that, even if students' assessments of instructional effectiveness are collected, there are any proven benefits that accrue from the subsequent dissemination of this information. The objective of this study was to provide insight into this question.

REVIEW OF THE LITERATURE

The issue of the usefulness of students' evaluations is one that has been raised increasingly in the past few years. In essence, it is concerned with the benefits that accrue from collecting students' evaluations of instructional effectiveness. Many possible benefits can be imagined; yet, certainly the central benefit is that of instructional improvement. Can it be demonstrated that there is some value to informing instructors of students' perceptions through periodic feedback? More specifically, do instructors who receive mid-term feedback subsequently receive higher ratings from their students at the end of the quarter than instructors who do not receive feedback from their students? Do students of instructors who

receive mid-quarter feedback learn more or develop a more positive attitude toward the subject matter than students of instructors who do not receive mid-quarter feedback?

Gage(1963) compared students' post-feedback ratings of their sixth grade teachers, adjusted for differences in initial student ability. Mean ratings for the feedback group instructors were higher than those for the no-feedback group instructors on 10 of 12 items; on one item, there was no difference and on the other item the no-feedback group averaged slightly higher. Four differences were statistically significant ($p<.05$) and favored the feedback group. These four items involved pupil-teacher interaction and the ability of the teacher to explain theory through everyday examples.

Tuckman and Oliver(1968) asked students of 286 vocational teachers at the high school or technical institute level to rate their teachers twice during a 12 week interval. Teachers assigned to the feedback group were given summaries of their students' initial assessments; those assigned to the no-feedback group were not. Results of the study showed that the presence of students' feedback produced statistically significant positive changes in ratings given to instructors selected to receive feedback, as compared to those who received no feedback.

Miller(1971) sought to determine empirically whether providing graduate teaching assistants (TAs) mid-semester feedback from students' ratings had any effect on subsequent ratings or on students' achievement. Students' achievement was measured by objective tests constructed from items submitted and discussed by the TAs. Results showed that average end-of-term ratings for TAs receiving mid-term feedback were not significantly different from average ratings of those who did not receive feedback.

Mean examination scores, however, were higher for students whose instructors received feedback during the term than for those whose instructors did not.

Centra(1972) assigned faculty members from five different types of colleges to a feedback, no-feedback or post-test only experimental condition. Only the feedback group instructors received results of mid-term and end-of-term evaluations from their students. Average end-of-term ratings for both feedback and no feedback groups were almost identical on each item, suggesting that instructors who received feedback were unable to benefit from it between the mid-term and end-of-term periods.

To further explore this phenomenon, instructors in the initial feedback group received mid-term feedback during an additional term. Their second semester end-of-term ratings were then compared to those of a randomly chosen group using the rating form for the first time. This time, those who received feedback did average higher ratings than those who did not. The results suggest that perhaps a longer period of time is necessary for the impact of feedback to be felt.

Marsh, Fleiner and Thomas(1975) examined evaluations of teaching assistants (TAs) furnished by undergraduate students in 18 different sections of an introductory course in computer programming. The TAs were evaluated by their students during mid-term and at the end of the quarter, on a 46-item evaluation instrument. Those TAs randomly assigned to the feedback condition received the results of their mid-quarter evaluations while the other TAs did not. No differences in performance on a common objective final examination were found between students of instructors who received feedback and students of instructors who did not. Positive differences in mean ratings on six of seven evaluation factors and three of four summary judgment items were obtained

favoring the feedback group over the no feedback group. Two of the factor score differences were statistically significant ($p<.05$): Instructor Approachability and Readings factors. Statistical significance ($p<.01$) was obtained for the summary item: "How does the quality of instruction at present compare to the quality of instruction at the beginning of the quarter?"

To summarize the existing studies on the usefulness of feedback, it can be said that while the utility of mid-term feedback was demonstrated in the two non-college studies, a lack of agreement existed at the college level. Miller(1971), using a cross-sectional design, found no statistically significant differences between average ratings for feedback and no feedback group instructors. However, students of instructors who received feedback did perform better on a common final examination, than did students of instructors who did not receive feedback. Centra(1972) also found no statistically significant differences over a one term period, but did discover significant differences over two terms. Marsh and Associates(1975), using a cross-sectional design, found statistically significant, positive differences on a number of items between feedback and no feedback group instructors.

METHODOLOGY

Subjects

Subjects were the 993 University of California, Los Angeles undergraduates who completed Engineering 10 (E 10), an introductory course in computer programming, during the fall, winter or spring quarter of academic year 1973-74. These students were typically social or behavioral science majors, and were taking the course to fulfill a departmental requirement. A different course was offered for students majoring in engineering or the physical sciences.

Students randomly enrolled in one of the sections of E 10 before the start of each 10-week quarter. Selection of sections was made solely on the basis of the time at which the section met; students had no prior information about who would teach the different sections. The course format, including the textbook, assignments, content, construction of a common final examination, and grading of the examination, was under the supervision of the course director. The course itself was taught solely by graduate teaching assistants (TAs). Each TA taught only a single section during any one quarter, but taught at least two of the three quarters during which the study was conducted.

Materials

A pretest, developed to predict final examination performance, consisted of background and demographic items (grade point average, Scholastic Aptitude Test score, background in mathematics, for example) and a short paper-and-pencil achievement test. Results of the pretest correlated 0.56 with the final exam. Preliminary analysis indicated that predicted final examination scores for the 30 sections involved did not differ to a statistically significant extent. This finding reinforced the assumption that the average ability levels in each section did not differ.

The student evaluation instrument, developed for this course by the Evaluation of Instruction Program (EIP), was designed to measure seven dimensions of teaching. Factor analysis of the end-of-quarter (EOQ) evaluations--a principal components solution followed by an oblique rotation (Nie and Associates, 1976)--showed that each of the seven factors was defined by the items designed to measure it. Facotr analysis of the mid-quarter (MQ) evaluations gave essentially the same factor solution. Factor scores for

each evaluation dimension were computed by taking an unweighted average of standardized responses to items that defined each factor. Factor scores for the MQ evaluations were computed in exactly the same manner. The evaluation instrument also included five items designed to provide information about the affective consequences of E 10 for each student. These consequences focused on the degree to which the student felt that he or she could apply the subject matter by the end of the course, and the extent to which the student planned to pursue the subject matter further after the course had been completed (Overall, 1977).

A common final examination was taken by all students during the 11th week of the course. The exam was basically of the objective type, with several practical programming exercises included. Each programming exercise was graded by only one instructor to insure consistency. The specific content of the exam, developed by the course director, was not made available to the TAs before the examination was administered to the students.

Procedure

Students enrolled in one section of the course before the start of each 10-week term. Selection of sections was made solely on the basis of the time at which the section met, as students had no information about who would teach the different sections. Three different instruments were administered during each term. The pretest was administered to all students during the first week, the mid-quarter evaluation was administered at the end of the fourth week, and the end-of-quarter evaluation was administered at the end of the tenth week. All three instruments were administered by an employee of EIP who had no connection with the actual course. Although students were required to put their registration number on each instrument, they were assured that their responses would remain anonymous.

Instructors were randomly assigned to the feedback (FBK) or no feedback (NFBK) conditions at the start of the first quarter. Results of the students' evaluations of instructors in the FBK condition were available by the start of the sixth week of classes; normative data indicating how the instructor's individual evaluation compared with the evaluations of instructors in the other sections were also available. FBK instructors were also given the results of the EOQ evaluations at the end of the term. FBK instructors continued to receive feedback for all three terms of the academic year. Results of the evaluations of students in the NFBK condition were not given to their instructors until after the end of the academic year during which the study was conducted. In some cases, instructors did not teach all three quarters, but this occurred equally often in the FBK and NFBK conditions. In these cases, any new instructors were randomly assigned to the two conditions. Every instructor, both those in the FBK and in the NFBK conditions, taught two of the three quarters.

Statistical Analysis

The impact of feedback from students' evaluations was assessed by comparing the responses of students who had an instructor either in the FBK or NFBK condition. Three different sets of comparisons were made: 1) changes between MQ and EOQ evaluations for the FBK and NFBK group instructors, 2) achievement on a common final exam by students of the FBK and NFBK group instructors, and 3) extent of affective consequences for students of both FBK and NFBK instructors. Analysis was complicated by the fact that a different number of responses was available for each of the three sets of criteria.

Changes in students' evaluations were assessed using analysis of covariance. Analysis of covariance is very similar to comparing change socres for the two conditions using a t-test, but provides a stronger statistical test of group differences because of its ability to control for pre-existing sources of potential bias (Anderson and Associates, 1975). The EOQ evaluations, controlled for the effect of any differences in ratings of effectiveness existing up to the MQ evaluations, were compared for the FBK and NFBK conditions. This analysis could only be performed on responses of students who had completed both the MQ and EOQ evaluation questionnaires (n = 674).

Final examination scores were available for all students who completed the course (n = 993). The statistical significance of average achievement differences between students of FBK and NFBK instructors was assessed by performing a t-test on the final exam score differences.

Affective consequences, perceptions of which were collected as part of the EOQ evaluations, were available for all students who evaluated instructors during the last week of the course (n = 796). The impact of feedback was again determined by doing a t-test on differences in students' perceptions with respect to instructors in the FBK and NFBK conditions.

RESULTS

Change in Students' Evaluations

If MQ evaluations given to instructors in the FBK condition were useful, then evaluations by their students at the end of the term should be improved to a statistically significant extent, compared to improvements in evaluations of the NFBK instructors. The EOQ evaluations, with the effect of MQ evaluations (covariate) removed through analysis of covariance, were generally more favorable for the group of instructors who received MQ feedback (see

Table One). Differences between the two conditions reached statistical significance for the two summary rating items, the perceived difference in instructional quality item, and four of the seven evaluation dimensions» Instructor Concern, Learning, Instructor-Student Interaction and Examinations. These results support the position that feedback from students' evaluations of instructional effectiveness is useful.

-----Insert Table One About Here-----

Existing research about the relationship of feedback to instructional improvement has not been consistent. Single term studies by Miller(1971) and Centra(1972) did not report any statistically significant differences between average ratings of instructors who received MQ feedback and average ratings of instructors who did not. Marsh, Fleiner and Thomas(1975), however, found statistically significant differences on two rating dimensions (Instructor Approachability and Readings) and for an identical perceived difference in instructional quality item ($t(285) = 3.85$, $p < .001$) favoring instructors who received mid-quarter feedback over a one quarter period. Thus, results of this study appear to be more supportive of the study by Marsh and Associates than of the studies by Miller and Centra.

Achievement

The second criterion used to assess the usefulness of feedback was student achievement on a common final examination. Students of instructors who received feedback performed better, to a statistically significant extent, on this test of course mastery, than students of instructors who did not receive feedback. This differential in average performance provides a strong argument for the position that students' evaluations are related to improved

student learning.

-----Insert Table Two About Here-----

Two other studies have dealt with the relationship of feedback to students' performance on a common final examination. Miller(1971) and Marsh, Fleiner and Thomas(1975) both found no statistically significant differences between the performance of students whose instructors did or did not receive feedback. Because of the contrast in findings with previous research, and because the results of this study reached an acceptable but not definitive level of significance, additional research is recommended.

Affective Consequences

Students of instructors in the FBK group--on the average--gave more favorable responses to each of five affective consequence items (see Table Three), the differences reaching statistical significance on three items. These differences in the affective consequences of E 10 also supported the usefulness of mid-quarter feedback from students' feedback.

-----Insert Table Three About Here-----

CONCLUSION

This study has presented evidence supporting the utility of students' evaluations of instructional effectiveness. First, students whose instructors received MQ feedback rated their instructors higher, as a group, by the end of the of the quarter than did students whose instructors did not receive feedback. These results provide a reasonable basis to assume that instructors found the feedback mechanism a useful source of information on which to base both their own assessments of instructional effectiveness by mid-quarter,

and plans for changes in their conduct of the class responsive to students' needs. This suggests the potent role that feedback may play in the learning environment.

Second, the positive relationship between receiving MQ feedback from students' evaluations and students' achievement on a common final exam suggests that feedback from such ratings results in a climate more conducive to cognitive growth. Students' learning is very much the concern of colleges and universities, and feedback from students' evaluations may help insure a classroom environment supportive of learning.

Third, two important indicators of a successful learning experience are the extent a student desires to pursue the subject matter further or apply what he or she has learned after the course has ended. In this study, students of instructors in the FBK condition showed a much stronger desire to pursue further coursework in the area of computer programming by the end of the course, and were more positive about joining the campus computer club to apply what they had learned, than were students of the NFBK instructors. Thus, this positive relationship between receiving feedback and more positive affective consequences for students is of great significance.

Based on the positive findings outlined above, it is clear that the allocation of resources to support periodic assessment can be justified. From the standpoint of instructional improvement alone, this research indicates that these evaluations have the potential to influence effectiveness in terms of students' cognitive and affective development. Furthermore, if the positive results shown in this study can be obtained through limited summary and comparative MQ feedback information, there is potential for even greater improvements through more elaborate and sophisticated means. Departments may have instructors who are strong in their research and community activities for the

university, but who are not perceived by students as effective instructors. These instructors, working through an instructional improvement unit on campus, might very well obtain assistance essential to this improvement through the administration of previously validated and tested evaluation instruments midway through their course.

BIBLIOGRAPHY

Anderson, S.B., Ball, S., Murphy, R.T., and Associates. Encyclopedia of Educational Evaluation. San Francisco: Jossey-Bass, Inc., 1973.

Centra, J.A. Two studies on the utility of student ratings for instructional improvement (Student Instructional Report No. 2). Princeton, N.J.: Educational Testing Service, 1972.

Gage, N.L. A method for "improving" teacher behavior. Journal of Teacher Education 14, 261-266, 1963.

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

Miller, M.T. Instructor attitudes toward, and their use of student ratings of teachers. Journal of Educational Psychology 62, 235-239, 1971.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Bent, D.H. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Company, 1975.

Overall, J.U. Students' Evaluations of Instructional Effectiveness: Validity and Utility. Unpublished doctoral dissertation. University of California, Los Angeles, 1977.

TABLE ONE

F-Values for FBK or NFBK Instructors' Differences in EOQ Students' Ratings

EVALUATION DIMENSION	F-Value[1]
Concern	F(1,669) = 20.93***
Organization	F(1,659) = 1.62
Learning	F(1,668) = 9.23**
Interaction	F(1,661) = 8.90**
Breadth	F(1,654) = 1.79
Difficulty	F(1,674) = 0.99
Examinations	F(1,657) = 8.46**
Overall Rating of Instructor	F(1,673) = 16.42***
Overall Rating of Course	F(1,671) = 3.68*
Perceived diff. in Quality	F(1,658) = 9.62**

*p<.05
**p<.01
***p<.001

[1]F-values are the result of an analysis of covariance testing for differences between the two conditions with the effect of MQ evaluations controlled.

TABLE TWO

Differences in Final Examination Scores of Students with Instructors in the Feedback or No Feedback Conditions

Final Examination	FBK Condition (n = 386)	NFBK Condition (n = 408)	t
Mean	51.28	49.74	
Standard Deviation	9.86	9.84	2.35*

*p<.05

TABLE THREE

Differences in Affective Consequences for Students with Instructors in the Feedback or No Feedback Conditions

AFFECTIVE CONSEQUENCE ITEM (Extent that...	FBK Mean (n = 386)	NFBK Mean (n = 408)	t
...you feel capable of writing and running a computer program to solve future problems.	5.76	5.43	2.08*
...you have gained enough understanding of what a computer is capable of to be useful to you in the future.	6.13	5.95	1.17
...you plan to become (or remain) a member of the campus computer club or find other sources of computer time in the future.	3.96	3.45	2.43*
...you plan to make practical application of the computer in the future.	5.00	4.65	1.73
...you plan to take more computer courses in the future.	4.23	3.47	3.50**

*p<.05
**p<.01

APPENDIX I

FACTOR ANALYSIS OF THE EVALUATION ITEMS

Factor analysis of the End-of-Quarter evaluation items, a principle components analysis[1] followed by an oblique rotation to a direct oblimin criteria[2], resulted in the following factor pattern matrix (loadings of less than .20 are indicated with dashes). The loading of each item on the factor it was designed to measure appears in a bold box. Every item loads higher on its own factor than any other. The evaluation factors are moderately intercorrelated, correlations ranging from .03 to .50 with a median of .25. The factors are generally positively related to each other with the exception of the Workload/Difficulty factor which has low negative to zero correlations with the other factors.

Evaluation Items	I	II	III	IV	V	VI	VII
I LEARNING							
Intellectual curiosity in subject stimulated	**.74**	-	-	-	-	-	-
Learned something valuable	**.72**	-	-	-	-	-	-
Present interest in course subject	**.69**	-	-	-	-	-	-
Developed understanding of practical implications	**.48**	-	-	-	-	-	-
Degree of course mastery	**.43**	-	-	-	-	-	.32
II CONCERN							
Instructors presentations made subject understandable	-	**.75**	-	-	-	-	-
Instructor concerned with student learning/understanding	-	**.55**	-	.30	-	-	-
Instructor enthusiastic about teaching	-	**.47**	-	.34	-	-	-
Instructor made course relevant	-	**.43**	-	-	.40	-	-
III ORGANIZATION							
Course material outlined and carefully explained	-	.20	**.67**	.25	-	-	-
Course objectives stated and agreed with those actually pursued	-	-	**.63**	.24	-	-	-
Nature/purpose of assignments clear	-	-	**.53**	.26	-	-	-
Presentation well prepared and integrated	-	.40	**.44**	-	-	-	-
Workload evenly spread over term	-	-	**.34**	-	-	-	-
IV STUDENT-TEACHING INTERACTION							
Students welcomed to seek help/advice	-	-	-	**.63**	-	-	-
Students encouraged to ask questions and were given answers	-	.32	-	**.59**	-	-	-
Students free to disagree and/or express own ideas	-	-	-	**.59**	-	-	-
V BREADTH OF COVERAGE							
Instructor contrasted implications of theories	-	-	-	-	**.79**	-	-
Instructor presented background/origin of ideas/concepts	-	-	-	-	**.65**	-	-
Instructor discussed different points of view	-	-	-	-	**.63**	-	-
VI EXAMINATIONS/GRADING							
Graded materials adequately measured your knowledge	-	-	-	-	-	**.81**	-
Graded materials measured content as emphasized in course	-	-	-	-	-	**.74**	-
Grading was fair and objective	-	-	-	-	-	**.56**	-
VII WORKLOAD/DIFFICULTY							
Workload/pace was difficult	-	-	-	-	-	-	**.65**
Course difficulty	-	-	-	-	-	-	**.63**
Hours/week outside of class	-	-	-	-	-	-	**.43**
OVERALL INSTRUCTOR RATING	-	.52	-	.26	-	-	-
OVERALL COURSE RATING	.51	-	.20	-	-	-	-

[1] Eigenvalues of the first seven principle components were 11.9, 3.0, 1.4, 1.3, 1.1, 1.1, 0.9.

[2] Factor analysis was preformed with the Statistical Package for Social Scientists (Nie, et. al., 1976).

FOOTNOTES

The authors would like to thank Dr. James Trent for encouragement and critical readings, Dr. Leon Levine, director of the UCLA Engineering course upon which the research was based, Brandon Beatty for assistance in data collection, and Chris Thomas for aid in the analysis.

[1]Reprint requests should be sent to Dr. J. U. Overall, IV, Director, Programs in Administration, California State College--Dominguez Hills, 1000 East Victoria, Dominguez Hills, California 90747.