DOCUMENT RESUME

ED 104 284 HE 006 411

AUTHOR McKeachie, W. J.; Lin, Y. G.
TITLE Use of Student Ratings in Evaluation of College Teaching. Final Report.
INSTITUTION Michigan Univ., Ann Arbor.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Mar 75
GRANT NE-G-00-3-0110
NOTE 41p.

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS College Students; Educational Improvement; *Effective Teaching; Faculty; *Higher Education; Rating Scales; Research Methodology; Research Projects; Statistical Data; *Student Attitudes; Teacher Behavior; *Teacher Evaluation; Teacher Improvement; Teacher Promotion; *Teacher Rating

ABSTRACT

Student ratings have been used for three major purposes--in each case by a different group: (1) to assist teachers to improve their teaching; (2) to aid administrative decisions with respect to promotions or salary increases of teachers; and (3) to provide descriptions of course and teachers for students choosing to enroll in courses or sections of courses. The three studies described in this report relate to the first two of these purposes: using student ratings to improve teaching; an experimental investigation of factors affecting university promotion decisions; and do discrepancies between student ratings, teacher expectations, and teacher ideals result in changes in teacher behavior? For each of the studies, the hypothesis, method, procedures and results are indicated. (MJM)

Use of Student Ratings in Evaluation of College Teaching

Final Report

March 1975

National Institute of Education
Grant No. NE-G-00-3-0110

FINAL REPORT

Grant No. NE-G-00-3-0110

USE OF STUDENT RATINGS IN EVALUATION OF COLLEGE TEACHING

W. J. McKeachie and Y. G. Lin

Researchers

Monica Daugherty
Mary M. Moffett
Cynthia Neigler
John Nork
Martina Walz

The University of Michigan
Ann Arbor, Michigan

March 1975

NATIONAL INSTITUTE OF EDUCATION

TABLE OF CONTENTS

LIST OF TABLES

BACKGROUND

Student ratings have been used for three major purposes—in each case by a different group:

a. to assist teachers in improving their teaching;

b. to aid administrative decisions with respect to promotions or salary increases of teachers; and

c. to provide descriptions of courses and teachers for students choosing to enroll in courses or sections of courses.

The basic premise of the studies proposed herein is that different sets of items are useful for each of these three purposes. This report describes studies relevant to the first two of these purposes.

STUDY I: USING STUDENT RATINGS TO IMPROVE TEACHING

So far as we know no studies have tested the usefulness of student ratings for all three purposes. There have, however, been studies relevant to one part of the problem—the use of student ratings to affect teacher behavior.

In a study of effects of feedback of pupil ratings on elementary school teachers' behavior, Gage, Runkel, and Chatterjee (1963) found that sixth grade teachers changed their behavior in the direction of pupil's descriptions of the ideal teacher. Tuchman and Oliver (1968) found at the high school level that feedback from student ratings changed teachers' behavior significantly as compared with teachers not receiving feedback. Miller (1971), however, most recently found no significant effect of feedback of midsemester college student ratings on the end-of-semester ratings of graduate student teaching assistants. He found, however, that mean final examination scores were higher for those students whose instructors received feedback of student ratings during the semester than those students whose instructors did not receive feedback in two of three courses. Miller concluded that "...the results suggest some limitations in the use of student ratings as a method of improving instruction." Centra's results (1972) support this conclusion.

A previous finding from our own research program helps explain the Miller results and points toward further research needed. In this study Pambookian (1972) also found little overall effect of feedback of student ratings, but detailed analysis of his data indicated that significant positive changes did occur for those instructors in the mid-third of the distribution of initial ratings. Pambookian suggests that those instructors in the top third had little need to improve while those in the bottom third may have become more anxious and defensive.

Miller provided feedback on a 15-item scale of which only 10 items dealt with instructor characteristics. We believe that more specific feedback on a large number of items is more likely to be helpful than feedback on a smaller number of more general items. We believe that the effect of feedback will be specific and is more likely to be apparent if separate measures on several dimensions are used rather than a single global measure.

HYPOTHESIS

Our hypothesis was that teachers given personal feedback of student ratings of instruction with encouragement and suggestions for improvement during a term will be superior on end-of-term measures of teaching effectiveness to teachers receiving printed feedback or those receiving no feedback.

MEASURES

Student Perception of Teaching and Learning

The primary measure was the Michigan Student Perception of Teaching and Learning form. The Michigan Student Perception of Teaching and Learning form has evolved over a period of 20 years in which items from the major student rating of teaching forms in use in the 1950's were factor analyzed to obtain items best representing the major dimensions used by students in rating teachers (Isaacson, et al., 1964). The form was then revised as the result of a series of validity studies (McKeachie, et al., 1971) and of a multiple discriminant analysis (McKeachie and Lin, in press). The version of the form used in this study consisted of 32 items. The form administered as a post-test differed from the pre-test form in containing two additional items requiring evaluation of the instructor's general teaching effectiveness and the value of the course as a whole. The form was administered at approximately the one-third point of the term and readministered with other outcome measures during the 11th and 12th weeks of the 14-week term. One-third of the sample did not receive the pre-test administration.

Psychological Thinking

While most other studies of feedback of student ratings have dealt with the effect of feedback on later student ratings, the ultimate criterion is not student ratings but student learning. An important second outcome measure therefore was student performance on items selected from the Introductory Psychology Criteria Test (Milhollana, 1966). This test was developed by a Committee of the Division on Teaching of the American Psychological Association and has been used in much of our previous research. It provides a measure of achievement of the higher level cognitive outcomes of the Bloom Taxonomy.

Attitudes

An Attitude toward Psychology Questionnaire consisting of eight Likert-type items drawn from Carrier's scale (1966) was administered as part of the test battery. Since previous use of the test had indicated that some of the test items appeared to involve an uncritical, naive, endorsement of psychology, the scoring key was changed to indicate simply agreement, disagreement, or undecided. Only those items were used which five introductory psychology teachers agreed to be appropriate in terms of the goals of introductory psychology courses.

A locally constructed 10-item Likert scale of Attitude toward Self was used to assess impact of the introductory classes on the student as a person. In one of the four introductory psychology courses involved in the experiment, a scale of Attitude toward Mental Illness was being used as a device for

assessing the effect of student participation at a mental hospital. Thus, it was also available as an additional criterion measure in that course.

Curiosity

To test student curiosity about psychology, students were informed that they could skip a section of the test involving descriptions of several experiments and a behavior modification case. This test was scored in terms of how many of the studies a student read, as well as the student's rating of interest in the study.

SAMPLE

The sample consisted of 37 graduate student teaching fellows and three faculty members recruited from the individuals teaching in the introductory psychology courses at The University of Michigan, Fall 1973. Of the 37 teaching fellows, 21 had had at least one year of previous teaching experience and all of the faculty members were full professors with 15 to 40 years of experience. Twenty two of the teachers were men; 18 women. Teachers were assigned randomly to the three groups within each course.

PROCEDURES

Members of the research staff administered the Michigan Student Perception of Teaching and Learning form during the last 15 to 20 minutes of a class period during the 5th to 7th weeks of the 14-week term. (See the Appendix for a copy of the form.) Wording of some items was changed to the present, rather than past, tense to make the form appropriate for administration early in the course rather than at the end of the course. Table 1 indicates the N in each of the classes in the sample. As the table indicates, there was fairly serious shrinkage in some sections, but a sufficient number of students gave post-test ratings to provide reliable ratings of instruction. In each of the three groups one section had less than 50% return on the post-test. One, of course, worries about the possible bias of the remaining sample, but some comfort may be taken in the fact that these classes were equally distributed between the three experimental groups.

Rating forms were scored by computer and the computer print-out presented the mean of the class and the mean of all classes in the same course for each of the seven dimensions for which the form is scored: Impact on students, rapport, teacher as person, group interaction, difficulty, structure, and feedback.

Following these factor scores were individual class and course means for each item listed under the heading of the appropriate factor.

The print-outs were mailed during the 5th to 8th weeks of the term to

TABLE 1

STUDENT SAMPLE

Section	Students Enrolled	Pre-Test N	Post-Test N
		Group I	
170-10	39	31	19
170-11	30	28	29
170-12	25	25	23
170-16	24	23	24
171- [illegible]	16	18	7
171- [illegible]	19	14	17
171-33	29	24	19
171-24	27	26	24
171-28	20	20	26
171-32	26	18	15
172- 7	21	19	22
172- [illegible]	18	20	18
172-13	10	15	16
		Group II	
170- 5	28	19	19
170- 6	30	22	18
170-20	17	14	8
170- 8	25	16	18
170-22	32	13	23
171- 6	21	24	11
171-23	27	18	19
171-25	21	19	19
171-14	14	17	12
172- 9	21	16	18
172- 4	24	16	15
172-12	13	13	13
		Group III	
170-13	25	--	10
170-[illegible]	16	--	28
170- [illegible]	30	--	26
170-17	23	--	35
171-16	26	--	22
171-19	26	--	21
171-18	28	--	20
171-38	26	--	24
171-30	15	--	14
171-41	25	--	14
172- 3	14	--	13
172- 8	18	--	16
Totals	845	488	695

those in the "Printed Feedback" group and were returned and reviewed personally during the 5th to 8th week of this term with each teaching fellow in the group receiving personal feedback. Teachers were given their choice of receiving feedback from their course supervisor or from Professor McKeachie. Professor McKeachie gave the personal feedback to 12 of the 14 teachers in this group and Professor Elton McNeil and Judith Reitman each gave the personal feedback to one graduate teaching assistant whom they were supervising.

At the beginning of the feedback sessions teachers were asked to fill out forms indicating their expectation of the student ratings on each dimension, their own self-perceptions, and where they would like to be. Typically, Professor McKeachie then asked them how the class was going and in response to their reactions, suggested how the student ratings confirmed (or rarely did not confirm) their perceptions. He then pointed out factors on which the teacher differed significantly from the mean of all classes. If there seemed to be any problems, he suggested some possible alternative methods of handling the problem. All of the mean ratings, however, were relatively favorable (see Table 2), so that the hope that he could help teachers cope with very negative feedback was not realized.

TABLE 2

DISTRIBUTION OF MEAN PRE-TEST STUDENT RATINGS ON "IMPACT OF COURSE"[1]

	Personal[2] Feedback	Printed Feedback
4.9 - 5.0		
4.7 - 4.89		
4.5 - 4.6	1	1
4.3 - 4.4	2	1
4.1 - 4.2	5	4
3.9 - 4.0	4	5
3.7 - 3.8	3	2
3.5 - 3.6	1	2
3.3 - 3.49	0	1
	16	16

[1] Impact of Course mean consisted of mean ratings on 16 items such as "The instructor stimulates my intellectual curiosity"

1 = almost never
2 = seldom
3 = occasionally
4 = often
5 = very often
6 = almost always

[2] The N in this table is larger than the N of teachers because two teachers in Group 1 and three teachers in Group 2 taught two sections.

During the 11th and 12th weeks of the 14-week term students were invited by postcard to attend evening sessions to take the tests measuring the dependent variables including the post-test of student perception of learning and teaching.

Upon completion of the tests each student received written feedback describing the experimental design and the nature of the tests used.

RESULTS

As Table 3 indicates, the student ratings at the end of the term support the primary hypothesis. Group I was rated as most effective.

The effect of feedback upon effectiveness as measured by student achievement was not as clear cut. As Table 4 indicates, the hypothesis received statistically significant support in terms of student achievement on the Criteria Test for Introductory Psychology for classes of Psychology 170 and for the measure of curiosity in Psychology 171, but the other criteria did not support the hypothesis.

We had expected the favorable effects of personal counselling to be most helpful to those teachers with poor ratings. Pambookian had shown printed feedback to be helpful to those in the middle third of his distribution. We had expected personal feedback to help reduce the potential negative effects of negative feedback by reducing anxiety, by increasing motivation to improve through increasing hope of success, and by suggesting alternative behaviors to those criticized. We thus separated the groups into thirds for further analyses. As Table 5 indicates, the results were in the direction predicted. A similar analysis of the other criterion measures was also in the predicted direction for the test of thinking, the test of attitude toward psychology, and the test of attitude toward self, but not for the test of curiosity. None of these differences, however, were significant (see Table 6).

CONCLUSION

As the introduction indicated, previous studies had been rather discouraging about the usefulness of student ratings in improving teaching. The results of the present study indicate that ratings alone are not very helpful but that a plan for using student ratings in counselling teachers can be helpful.

TABLE 3

EFFECTS OF FEEDBACK ON TEACHING EFFECTIVENESS:
FACTOR SCORES OF STUDENT PERCEPTIONS OF TEACHERS

	Mean Class Scores			F	p
	Personal Feedback	Printed Feedback	No Feedback		
Overall Item[1]					
General teaching effectiveness	3.6	3.1	3.0	4.55	.017
Overall value of course	3.4	3.0	3.0	5.77	.007
Dimension[2]					
Impact on students[3] (Items 1 to 16)	4.17	3.95	3.80	3.60	.037
Rapport (Items 17 to 20)	4.63	4.59	4.54	.17	.846
Teacher as person[4] (Items 21 and 22)	2.18	2.60	2.46	2.82	.072
Group interaction (Items 23 to 25)	4.16	4.34	4.10	.69	.506
Difficulty (Items 26 and 27)	2.23	2.36	2.23	.23	.797
Structure (Items 28 and 29)	3.55	3.13	3.03	1.31	.281
Feedback (Items 30 to 32)	4.14	4.09	3.79	.90	.413
Number of sections	14	13	13		

[1]Scale for overall items:
5 = Excellent; 4 = Very good; 3 = Good; 2 = Fair; 1 = Poor

[2]Scale for dimensions:
1 = Almost never or almost nothing
2 = Seldom or little
3 = Occasionally or moderate
4 = Often or much
5 = Very often
6 = Almost always or a great deal

[3]High factor score represents high impact.

[4]High score is over-personal.

TABLE 4

EFFECT OF FEEDBACK ON TEACHING EFFECTIVENESS,
AS MEASURED BY STUDENT PERFORMANCE

Measure	Mean Class Scores: Personal Feedback	Printed Feedback	No Feedback	F	p
Psychology 170					
Psychological thinking	19.09	16.44	17.52	5.941	.02
Attitude to Psychology	10.50	9.57	10.75	8.014	.01
Attitude toward self	4.69	4.16	3.83	2.567	.12
Curiosity	14.32	12.41	17.49	2.584	.12
Number of sections	4	5	4		
Psychology 171					
Psychological thinking	15.65	15.91	15.90	.043	.96
Attitude to Psychology	10.04	10.47	9.11	1.925	.18
Attitude toward self	5.07	4.14	4.37	3.241	.14
Curiosity	17.90	9.36	13.81	5.557	.02
Number of sections	6	4	6		
Psychology 172 & 192					
Psychological thinking	17.03	16.61	16.45	.105	.90
Attitude to Psychology	9.92	11.10	9.70	.962	.42
Attitude toward self	4.78	4.92	4.56	.218	.81
Curiosity	11.97	16.07	11.77	1.208	.35
Number of sections	4	4	3		

TABLE 5

POST-COURSE STUDENT RATINGS OF IMPACT
FOR TEACHERS DIFFERING ON INITIAL RATINGS

Initial Ratings	Personal Feedback M	S.D.	N	Printed Feedback M	S.D.	N
Low	3.9	.21	5	3.5	.44	4
Mid	4.2	.19	4	4.0	.08	4
High	4.5	.27	5	4.3	.31	5

t = 1.02 for differences of differences
Low vs. mid and high.

TABLE 6

EFFECTS ON FEEDBACK ON LOW, MIDDLE, AND HIGHLY RATED TEACHERS

	Personal Feedback		Printed Feedback		No Feedback		F	p
	Mean	S.D.	Mean	S.D.	Mean	S.D.		
Low Group (On Post-Test Impact on Students)								
Psychological thinking	50.1	1.4	49.3	1.1	49.9	.6	.60	.57
Attitude to Psychology	10.2	1.2	9.4	.6	10.0	1.7	.42	.67
Attitude to self	4.6	.7	3.7	.7	4.0	.6	2.24	.16
Curiosity	13.8	2.8	10.6	4.8	16.1	4.0	2.05	.16
Number of sections	5		4		4			
Middle Group								
Psychological thinking	50.5	.4	50.0	1.1	49.9	1.1	.40	.68
Attitude to Psychology	10.1	1.7	11.1	1.6	9.7	.8	1.26	.32
Attitude to self	5.3	.7	4.6	.5	4.2	.8	2.33	.10
Curiosity	13.6	7.3	13.9	3.8	14.0	3.5	.01	.99
Number of sections	4		5		5			
High Group								
Psychological thinking	50.3	1.4	49.7	.3	50.1	.6	.47	.64
Attitude to Psychology	10.2	.3	10.3	1.1	9.6	.3	.97	.41
Attitude to self	4.8	.4	4.8	.5	4.6	.7	.20	.82
Curiosity	17.9	6.0	13.0	3.5	13.4	2.4	1.75	.22
Number of sections	5		4		4			
Means and Standard Deviations of Post-Test Impact on Students								
Low Group	3.8	.2	3.5	.4	3.5	.2		
Middle Group	4.2	.2	4.0	1.0	3.8	.1		
High Group	4.5	.2	4.4	.2	4.1	.2		

STUDY II: AN EXPERIMENTAL INVESTIGATION OF FACTORS AFFECTING UNIVERSITY PROMOTION DECISIONS

INTRODUCTION

Current pressures for use of student ratings focus heavily on their use as evidence on teaching effectiveness for decisions about faculty promotions or salary increases. Although data on the validity of student ratings are important in determining whether student ratings should be used, we have no data on how they actually influence decisions. We need to begin to determine what kind of contribution student ratings can make to such decisions. One first step is to determine what information decision makers use from reports of student ratings. The exploratory study here reported provides data relevant to the following questions:

Does the type of information about teaching competence affect promotion decisions?

What are the relative weights of teaching and research on promotions?

The present report describes an initial attempt to investigate empirically the factors involved in the decisions made by university promotion committees concerning the promotion of assistant professors to the rank of associate professor. For purposes of simplicity, only two major factors were considered in the present study—teaching ability and research productivity. The study was designed not only to give information about the effect of these variables on promotion but also to determine whether the methodology employed would be capable of providing relatively precise estimates of the relative emphasis on teaching ability and research productivity in salary and promotion decisions.

One of the primary issues of interest in the study was whether or not the type of information provided in the evaluation of a promotion candidate's teaching ability would affect the promotion decision. The two types of evaluative information that were investigated were: (a) the department chairman's subjective report of the candidate's teaching ability, and (b) a summary of the student evaluations of the candidate's teaching ability for each of the courses taught by the candidate in the past two years. Although an increasing number of colleges and universities have been using student evaluations to aid in the assessment of the quality of teaching, very little is known about whether or not this information is utilized at the level of promotion decisions. Thus the problem has practical importance.

Because of the design of the study, two supplementary issues could also be investigated. These were whether or not the sex and academic department affect the weighting of research and teaching in promotion decisions. The sex issue is of interest in light of recent Affirmative Action programs, and

disputes about the existence, or extent of, sex discrimination in university promotion decisions. The academic department of the candidate was manipulated primarily to increase the generality of any results.

METHOD

Subjects

Twenty senior faculty members at The University of Michigan were asked to judge the promotability of cases presented to them as described below. All judges were either currently members of, or had previously served on, University of Michigan promotion committees. No one refused the request, but one judge did not return his ratings.

Materials

Six fictitious individuals were created and their case histories were prepared in the same format as that used for recommending actual candidates for promotion at The University of Michigan. The biographical and educational information for the six candidates was devised in such a way that all candidates had approximately equivalent backgrounds; i.e., all were approximately the same age (31-33 years); all had attended prestigious graduate schools; and all had been first appointed in September, 1969. The independent variables in the study were the levels of teaching and research competence associated with each fictitious promotion candidate and the type of information presented about teaching. For each candidate alternate versions of each case history were prepared with different levels of research productivity and/or teaching ability. The combination of teaching and research levels employed are presented in Table 7.

TABLE 7

COMBINATIONS OF TEACHING ABILITY AND RESEARCH PRODUCTIVITY USED IN THE EXPERIMENTAL MATERIALS

Research Productivity	Teaching Ability		
	Excellent	Medium	Poor
Excellent	(a)	(b)	(c)
Poor	(d)	(e)	(f)

The levels along the Teaching Ability continuum were created in two ways. In the Chairman's report condition, the promotion candidates were rated on their teaching ability in purely verbal form, as conveyed by means of the department chairman's opinions concerning the candidate's teaching ability. The types of phrases characterizing each level were: Excellent—"excellent," "superior," "truly outstanding," and "about average"; and Poor—"somewhat below average," "not particularly impressive," and "perhaps not outstanding." In the student-rating condition numerical averages of the student ratings of teaching ability were included in the evaluations as well as the verbal phrases characterizing each level. The student ratings were represented as being derived from a 5-point scale ranging from 1 (for excellent) to 5 (for poor). The average ratings for each level were: Excellent—1.50, medium—2.60, and poor—3.75. These values were chosen because they approximated best, medium, and worst ratings received by a large group of teachers over a number of years at The University of Michigan. Type of teaching evaluation was treated as a between-subject factor of the candidate's teaching (hereafter designated as chairman's-report condition), and 10 subjects receiving the student-rating information as well as the chairman's report (hereafter referred to as the student-rating condition).

The two levels of research productivity were created by varying the number of research publications listed in the candidate's vita from an average of 3.1, with a range from 2 to 4, for the low-productivity level, to an average of 13.3, with a range from 11 to 16, for the high-productivity level. Also varied across the two levels of research productivity were the types of descriptive comments included in the evaluation of the research. In the low-productivity cases the comments included the following phrases: "not impressive in quantity," "not one of the most productive," "not very active," and "perhaps not outstanding." The comments in the high-productivity cases included such phrases as: "large number of high-quality articles in prestigious journals," "consistently high quality," "international recognition," "solid scientific reputation," and "impressive in quantity and quality."

Each of the evaluators or subjects in the study received six promotion case histories to evaluate, one at each of the combinations of teaching ability, and research productivity described above. The particular promotion candidates assigned to each combination level was varied across evaluators such that each promotion candidate was evaluated at several levels of teaching ability and research productivity.

Two other aspects of the candidates were also manipulated. These were the sex of the candidate—one of the six candidates presented to each evaluator was a female and the other five males, and the academic department with which the candidate was associated—Psychology for three of the candidates and Physics for three of the candidates. The sex manipulation was achieved by pairing two candidates and describing one as a female and the other as a male in half of the case histories and then reversing this sex relationship for the remaining case histories. The academic department manipulation was achieved by

presenting an equal number of Physics candidates as Psychology candidates to each of the evaluators.

Special care was exercised throughout the construction of all case histories to insure that the portrayals of the promotion candidates were as realistic as could be managed within the constraints of the study. Since the realism of the case histories was considered to be essential for the validity of the study, all of the evaluators were asked to rate the realism of the case histories upon completion of their evaluation decisions.

Because all other interpretations would be tainted if the experimental materials were perceived by the subjects to be unrealistic, the data concerning the realism judgments need to be considered before evaluating the other results. Only eighteen subjects responded to this item in the questionnaire, but their responses were encouraging, as may be seen in Table 8.

TABLE 8

SUBJECTS' JUDGMENTS ABOUT THE REALISM OF THE MATERIALS

	"Very Realistic"	"Generally Realistic"	"Not Very Realistic"	"Not at All Realistic"
Number of subjects	3	14	0	1

Several subjects offered comments about the materials, and from these the most serious reservation seemed to be that there were no letters from experts outside the department evaluating the candidate's research. Other comments mentioned that there was too little variation across candidates, and that there was little or no service on department or university committees in any of the candidate descriptions. In general, however, the materials were apparently believable and the decisions made regarding them were taken seriously.

PROCEDURE

Each evaluator was asked to make both a decision regarding promotion (yes or no) and a decision regarding the amount of salary increase (from $0 to $1500 per year) for each of six fictitious promotion candidates. Upon completion of the promotion and salary decisions for all candidates, the evaluators answered a questionnaire containing items designed to assess the evaluators':

(a) rank-orderings of the candidates in terms of their desirability for promotion;

(b) opinions of the realism of the case histories;

(c) estimates of the relationship between teaching ability and research productivity in current faculty members at The University of Michigan; and

(d) opinions of the most desirable combination of teaching excellence and research excellence.

The evaluators' opinions concerning the most desirable combination of research and teaching competences for a promotion candidate were elicited in the following manner: First, each evaluator was shown a graph in which the axes represented arbitrary scales of excellence such as percentile ranks. The ordinate was used to indicate increases in research quality and productivity and the abscissa was used to indicate increases in teaching quality and effectiveness. Next, the evaluators were requested "...to draw a line on the graph... to enclose the region of research and teaching percentile values in which a candidate would be seriously considered for promotion...(that is)...outline a region that would include all the possible combinations of teaching ability or excellence and research ability or excellence in which a candidate possessing those percentiles would receive serious consideration for promotion."

All of the materials described above (i.e., the six case histories, the salary and promotion decision forms, and the final questionnaire) were mailed to the evaluators along with a cover letter of explanation and instruction. It is estimated that the evaluators took from 1 to 6 hours to complete the materials. The materials were returned by nineteen of the subjects from one to eight weeks after the initial mailing. The twentieth subject failed to return the materials and since attempts to contact him were unsuccessful, he was dropped from the study.

RESULTS

The first dependent variable to be considered is the percentage of "yes" promotion decisions at each level of teaching ability and research productivity. These percentages are listed in Table 9.

TABLE 9

PERCENTAGE "YES" DECISIONS AT EACH LEVEL
OF TEACHING ABILITY AND RESEARCH PRODUCTIVITY

Research Productivity	Teaching Ability			
	Excellent	Medium	Poor	Mean
Excellent	100% / 88.8%	80% / 77.7%	60% / 77.7%	80% / 71.4%
Poor	50% / 22.3%	0% / 0%	0% / 0%	16.7% / 7.4%
Mean	75% / 55.5%	40% / 38.8%	30% / 38.8%	48.3% / 44.4%

Note: The numbers above the diagonals are the means from the 10 subjects in the condition in which student ratings were reported, and the numbers below the diagonals are the means from the 9 subjects in the chairman's-report condition.

Discussion of the pattern of results in Table 9 will be deferred until after the results from the other dependent variables have been presented.

In addition to the decisions about whether each candidate should be promoted, the subjects also made independent judgments about the amount of salary increase appropriate for each candidate, and the rank order of the candidates in terms of their desirability for promotion. Both of these sets of data were subjected to analyses of variance, using the mean values of each cell to substitute in the case of missing data. The three factors in the analysis of the salary data were: (a) the type of teaching-evaluation information; (b) the level of teaching ability; and (c) the level of research productivity. Only the teaching-level and research-level factors could be analyzed with the rank-order data since the ranks for each group and, indeed, for every subject, were constrained to have the same mean. The results from the two analyses of variance are presented in Tables 10 and 11.

TABLE 10

MEAN SALARY INCREASE AT EACH LEVEL
OF TEACHING ABILITY AND RESEARCH PRODUCTIVITY

Research Productivity	Teaching Ability			
	Excellent	Medium	Poor	Mean
Excellent	1355.4 / 1318.8	1030.0 / 887.5	925.0 / 865.6	1103.0 / 1074.3
Poor	792.5 / 837.5	440.0 / 438.9	360.0 / 335.5	530.8 / 543.9
Mean	1073.7 / 1078.1	735.0 / 663.2	642.5 / 660.6	

Analysis of Variance Results
From the Salary and Rank-Order Data

Dependent Variable: Salary Increase

Factor	df	F	w^2
Type of teaching evaluation information	1.18	.06	0
Teaching ability	2.36	55.99*	.223
Research productivity	1.18	164.13*	.864
Interaction (Type of teaching evaluation information X teaching ability)	1.36	1.21	

* $p < .001$.

Note: For all other interactions $F < 1.0$.

Note: The index w^2 supplements the F values by providing an estimate of the proportion of the variance in the dependent variable accounted for by the independent variable (see Hays, 1963).

TABLE 11

MEAN RANKS OF PROMOTABILITY

Research Productivity	Teaching Ability			
	Excellent	Medium	Poor	Mean
Excellent	1.00 / 1.00	2.65 / 2.40	2.75 / 2.80	2.13 / 2.07
Poor	3.70 / 3.80	5.40 / 5.10	5.50 / 5.90	4.87 / 4.93
Mean	2.35 / 2.40	4.02 / 3.75	4.13 / 4.35	

Dependent Variable: Rank Order

Factor	df	F	w^2
Teaching ability	2.36	85.16*	.224
Research productivity	1.18	1058.40*	.680
Interaction (Type of teaching evaluation information X teaching ability)	1.36	2.80	

*p < .001.

Note: For all other interactions F < 1.0.

Note: The index w^2 supplements the F values by providing an estimate of the proportion of the variance in the dependent variable accounted for by the independent variable (see Hays, 1963).

Following the convention introduced in Table 9, the numbers above the diagonals are from the subjects in the student-rating condition, and the numbers below the diagonals are from the subjects in the chairman's-report condition.

The results presented thus far can be summarized quite briefly. First, in none of the analyses did the type of teaching-evaluation information make a significant difference in the promotion decisions. Second, all of the data indicate a considerably larger emphasis on research productivity than on teaching ability. This is obvious in the relative magnitudes of the F values from the analyses of variance, and particularly in the size of the w^2 estimates. It is also apparent by inspecting Tables 9, 10, and 11, and comparing the range of values across research levels (i.e., the row means) with the range of values across teaching levels (i.e., the column means), or by comparing the means of the excellent research-poor teaching combination with the means of the poor research-excellent teaching combination. All of these comparisons reveal that in making promotion decisions, in making salary-increase decisions, and in assigning rank-orders to promotion candidates, the subjects placed much more emphasis on research productivity than upon teaching ability.

Additional Result: Sex and Department

The results from the supplementary issues concerning the sex and academic department of the candidate are presented in Table 12. Although there were certain consistent trends in all three dependent measures, the absolute magnitudes of the differences in Table 12 are quite small, particularly in comparison to the differences obtained across the various teaching and research levels in Tables 9, 10, and 11. Because of a mistake in implementing the design there was a partial confounding of sex and the particular combination of teaching ability and research productivity. This prevented the use of statistical tests on the data, and only weighted averages could be used to obtain the summary results in Table 12.

TABLE 12

ANALYSES OF CANDIDATE'S SEX AND ACADEMIC DEPARTMENT AS FACTORS INFLUENCING DECISIONS

	Percent Promotions	Salary Increase	Rank Order
Sex			
Males	41.7%	$733.34	3.73
Females	37.5	670.38	4.07
Academic Department			
Psychology	48.3	838.98	3.45
Physics	46.0	782.13	3.55

The data from the question dealing with the most desirable combination of teaching excellence and research excellence in the final questionnaire administered to the subjects lend further support to the conclusion that a much heavier emphasis is placed on research than on teaching in evaluating candidates for promotion. It will be remembered that the subjects were instructed to outline a region on a two-dimensional graph that represented the most desirable combinations of teaching and research excellence for a promotion candidate. These regions were analyzed by computing the areas of the outlined regions above the positive diagonal (i.e., those regions in which the emphasis on research excellence, scaled on the ordinate, is more than the emphasis on teaching excellence, scaled on the abscissa), and the regions below the positive diagonal (i.e., where teaching excellence, scaled on the abscissa, is emphasized more than research excellence, scaled on the ordinate) and dividing the area above the diagonal by the area below the diagonal. The ratio resulting from these computations might then be considered a measure of relative research emphasis, in the abstract, since it is not dependent upon any particular case histories as were the other analyses reported earlier. The mean ratio for the 18 subjects completing this item in the questionnaire was 1.79 in favor of research. Nine of the subjects assigned equal weight to teaching excellence and research excellence; none placed greater emphasis on teaching than on research.

A final datum from the post-experiment questionnaire was the subject's estimates of the actual correlation between research productivity and teaching ability in current members of the faculty at The University of Michigan. The mean correlation from the 17 subjects reporting an estimate was +.56, with a range from +.25 to +.75.

DISCUSSION

Before discussing the results of the study, it is perhaps best to mention some of the limitations governing any interpretations. First, there is the very obvious limitation that the results cannot be generalized beyond a single university at a given point in time. Second, the case histories were perhaps too simplistic since they were constructed so as to vary along only two major dimensions. And third, there is no assurance that the different levels along the teaching-ability continuum are in any way equivalent to the different levels along the research-productivity continuum. That is, the difference between what we have termed excellent and poor teachers may not have represented the same difference as that between what we have termed excellent and poor researchers.

With these limitations in mind, we can nonetheless still be optimistic about the potential of this type of study. We were successful in obtaining thoughtful representative judgments from subjects of the target population with which we were concerned, university promotion committee members. And, the materials were judged to be "generally realistic." Thus our methodology seems worthwhile for further studies.

One of the major goals in the study was to determine whether the type of teaching-evaluation information provided in the promotion candidate's description affected the decision regarding that candidate's promotion, salary increase, or relative rank in a "desirability-for-promotion" scale. As the results of the analysis of variance on the salary data indicate, and as is evident in Tables 9, 10, and 11, the addition of the student-rating information apparently made little difference in any of the decisions. From the present results, we find little evidence that information from student ratings of teaching is utilized where decisions regarding promotions and salaries are made.

It is interesting to speculate about the reasons for the failure to utilize the additional, and possibly more objective, information about the candidate's teaching. The most obvious interpretation is that the quality of the teaching was considered to be a very small factor in evaluating the candidates, and hence, the specific type of information used to assess the teaching did not really matter.

A second possibility for the apparent failure to utilize the student-rating information is that the subjects may not have believed that student ratings were an accurate, or valid, way to assess teaching quality.

A third possibility is that the form in which the student ratings were presented was not persuasive. We suspect that a combination of statistical summary and direct quotations would be more persuasive than the numbers alone.

A fourth possibility is that teaching information is critical to decisions only when the candidate's research qualifications are not clearly excellent or poor. In such cases teaching may tip the balance and the quality of information provided about teaching may be critical. With our basic methodology apparently established as adequate, we hope to investigate this possibility next.

The secondary issues of interest in the present study, concerning the sex and academic department of the candidate, appeared not to be major factors in the decisions although a confounding with teaching research level precluded a formal statistical test.

One of the clearest findings in the experiment was the marked emphasis on research productivity compared to teaching ability. That there is such an emphasis on research is not surprising, but the current techniques allow reasonably precise estimates of the relative emphasis on research compared to teaching. Indeed, so many estimates are available (see Table 13) that one is faced with the problem of deciding which particular one is best. Fortunately, all of the estimates are reasonably similar, indicating approximately twice as much emphasis on research compared to teaching.

TABLE 13

QUANTITATIVE ESTIMATES OF THE AMOUNT OF RESEARCH EMPHASIS IN PROMOTION DECISIONS

	Estimate
Percent Promotions	
(Excellent Research-Poor Research) / (Excellent Teaching-Poor Teaching)	2.00
(Excellent Research-Poor Teaching) / (Poor Research-Excellent Teaching)	1.90
Salary Increase	
w^2 Research / w^2 Teaching	3.87
(Avg. Excellent Research-Avg. Poor Research) / (Avg. Excellent Teaching-Avg. Poor Teaching)	1.28
(Excellent Research-Poor Teaching) / (Poor Research-Excellent Teaching)	1.15
Rank Order	
w^2 Research / w^2 Teaching	3.03
(Avg. Excellent Research-Avg. Poor Research) / (Avg. Excellent Teaching-Avg. Poor Teaching)	1.50
(Excellent Research-Poor Teaching) / (Poor Research-Excellent Teaching)	1.36
Graphic Representation of Most Desirable Characteristics	
Research Emphasis / Teaching Emphasis	1.79

Perhaps the contribution of greatest import in the present study is the introduction of a viable methodology for the investigation of decision making in academic institutions. It is our hope that the research reported here will serve as an impetus to begin the systematic investigation of these decision processes.

STUDY III: DO DISCREPANCIES BETWEEN STUDENT RATINGS, TEACHER EXPECTATIONS, AND TEACHER IDEALS RESULT IN CHANGES IN TEACHER BEHAVIOR?

INTRODUCTION

Much attention has been given to student ratings in recent years. Much research has investigated whether or not student ratings of instruction are related to the effectiveness of teachers as determined by student achievement. Other studies have focused on the effect of student ratings on the instructor's behavior. The present study examined one aspect of this effect, specifically, do student ratings of teacher effectiveness have differential effects on instructors whose own perceptions of their teaching ability were similar to their students' perceptions as compared to those whose perceptions were discrepant from the student ratings?

Both Centra (1972) and Pambookian (1972) studied student ratings in terms of such discrepancy. Centra found that teachers who were shown that they had unrealistically high opinions of their teaching changed the most in a positive direction. As this discrepancy, where students rated teachers less favorably than the latter expected, increased there was an increased likelihood of teacher change. Similarly, when teachers rated themselves as average or poor and students' ratings concurred, teachers showed very little change despite their awareness of a need for improvement. Pambookian separated his teacher sample into three groups: (a) unfavorably discrepant teachers whose self-ratings were higher than their students' ratings; (b) minimally discrepant instructors whose own perceptions of their teaching ability were similar to their students' opinons; and (c) the favorably discrepant teachers whose students rated them as better than what the teachers themselves perceived of their abilities. As predicted, Pambookian found that across all dimensions on the student rating forms, the unfavorably discrepant group improved the most after seeing their students' ratings, followed by the minimally discrepant and then favorably discrepant groups.

In the present study an additional factor was considered, that of the teacher's opinion of how he would like to teach, termed here as his "ideal." Thus teachers were divided among eight classification groups, as enumerated below:

(A) Teacher expected and ideal ratings higher than students'.

(B) Teacher expected rating higher than student pre-test (student ratings given after the first five weeks of classes) and ideal at or below student pre-test.

(C) No discrepancy between ideal, expected, and student pre-test ratings.

(D) No discrepancy between teacher expected and student pre-test ratings, but teacher ideal rating higher.

(E) No discrepancy between teacher expected and student pre-test ratings, but teacher ideal lower.

(F) Teacher expected rating lower than student pre-test and ideal rating higher than student pre-test.

(G) Teacher expected rating lower than student pre-test and ideal close to the student pre-test.

(H) Teacher expected rating lower than student pre-test and ideal rating lower than student pre-test.

On the basis of our analysis of research on effect of feedback on performance, we hypothesized that feedback results in greatest improvement when: (a) the knowledge of results gives new information to the learner; (b) the learner is motivated to improve; and (c) the learner knows what actions are necessary to improve. In terms of the above-listed groups, we expected group A (corresponding roughly to Pambookian's unfavorably discrepant group) to show the most improvement as a result of the feedback and group H to show the most negative changes in performance.

METHOD

Sample

The data were collected as part of the larger study on the effect of the feedback of student ratings on a teacher's effectiveness. The sample consisted of 28 instructors of introductory psychology classes at The University of Michigan.

Measures

The measure used to assess teacher performance was the Michigan Student Perception of Teacher Form. It consists of 32 items distributed among seven dimensions: Impact (the intellectual effect of the teacher on the student), Rapport, Teacher-as-Person, Group Interaction, Difficulty, Structure, and Feedback.

In addition, the instructors completed a form indicating their own perceptions of their teaching ability. For each of the seven dimensions, the teacher estimated whether he expected his students would rate him to be in the top 10%, above average, average, below average, or in the bottom 10% of

the sample. Instructors also indicated their "ideal" of where they would most like to be on each scale.

Procedure

The students of the teachers in the sample completed the student rating form for the first time at approximately the one-third point in the semester. Before the instructors saw these results, they completed the teacher expectation and ideal evaluation forms.

Approximately two weeks before the end of the semester, the students reevaluated their teachers. The same rating form as in the pre-test was used.

Mean change scores were computed between the pre-test and post-test ratings for each instructor on each dimension. In addition, discrepancy scores were computed between (a) the teacher's expectations and the actual ratings, and (b) the teacher's ideal compared to the actual ratings. Instructors were then distributed among the previously mentioned groups A through H. Not all of these groups emerged on every dimension (see Table 14 for group N's), but group E was the only one which was never represented on any of the seven dimensions.

RESULTS AND DISCUSSION

There was some support for our hypothesis that discrepancy between student pre-test ratings and teacher expected and ideal ratings would result in changes in teacher performance. Table 14 presents the mean change scores for each group of teachers as well as a brief reiteration of the type of discrepancy that characterizes each group. Although no significant differences or patterns emerged regarding the change scores among the teacher groups for most dimensions, there were significant differences on the dimensions of Group Interaction and Feedback.

Groups A and G showed the most change on Group Interaction after the pre-test ratings in contrast to groups D, F, and H who only changed slightly, the latter two changing in the negative direction. These results support our hypothesis that the teachers whose expectations and ideal ratings were higher than the students' ratings (group A) would show the most improvement. Group G's change in behavior is not consistent with our expectation because neither their expectations nor their ideals where above the actual ratings.

The mean change scores on the Feedback dimension were more dramatic. Groups A and F showed the most improvement, while group H changed markedly in the negative direction. This clearly supports our hypotheses. Groups A and F were both motivated to improve whereas the teachers in group H received ratings which were higher than both their expectations and ideals thus providing no motivation to improve. Groups A and H correspond roughly to

TABLE 14

THE EFFECT OF DISCREPANCY BETWEEN STUDENT PRE-TEST RATING, TEACHER EXPECTED, AND IDEAL RATINGS ON CHANGE OF STUDENT RATINGS FROM PRE-TEST TO POST-TEST

Group	Discrepancy[1] Between Student Rating and		Dimension													
			Impact		Rapport		Teacher as Person		Group Interaction		Difficulty		Structure		Feedback	
	Teacher Expectation	Teacher Ideal	N	Mean Change	N	Mean Change	N	Mean Change	N	Mean Change	N	Mean Change	N	Mean Change	N	Mean Change
A	+	+	17	.04	18	-.08	14	.04	15	.35	19	.16	13	-.15	17	.67
B	+	0,-									2	.04	1	-.38		
C	0	0	1	-.13	4	-.51	5	.27			2	.22			3	-.34
D	0	+			2	-.08	1	-.06	2	.12			1	.37	3	.29
E	0	-														
F	-	+	9	.06	4	-.14	1	-.51	3	-.08	1	-.33	-	-.26	3	.58
G	-	0							1	.38						
H	-	-	1	-.22			7	-.02	7	-.03	4	-.12	3	-.14	2	-1.31
			F = .39		F = 1.33		F = 1.8		F = 3.60*		F = .33		F = .46		F = 3.60*	

[1] Key:
+ indicates the teacher's expectancy or ideal was higher than the actual rating.
0 indicates minimal discrepancy between teacher expectations or ideal and the actual rating.
- indicates the teacher's expectancy or ideal was higher than the actual rating.

*$p < .05$

Pambookian's unfavorably discrepant and favorably discrepant groups, respectively. Thus, their change scores on this dimension concur with Pambookian's findings.

REFERENCES

Carrier, N. A. Evaluating the Introductory Psychology Course. Reading, Mass: Addison-Wesley, 1966.

Centra, J. A. Evaluating college teaching: The rhetoric and the research. (Paper presented at American Association for Higher Education, Chicago, Illinois, March, 1972.)

Gage, N. L., Runkel, P. L, and Chatterjee, B. B. Changing teacher behavior through feedback from pupils: An application of equilibrium theory. In W. W. Charters and N. L. Gage (Eds.). Readings in the social psychology of education. Boston: Allyn and Bacon, 1963, pp. 173-181.

Hays, W. L. Statistics. New York: Holt, Rinehart, and Winston, 1963.

Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Y. G., Hofeller, M., Baerwaldt, J. W., and Zinn, K. L. Dimensions of student evaluations of teaching. Journal of Educational Psychology, 1964, 55, 344-551.

McKeachie, W. J., Lin, Y. G., and Mann, W. Student ratings of teacher effectiveness: Validity studies. American Educational Research Journal, 1971, 8, 435-445.

McKeachie, W. J. and Lin, Y. G. Multiple discriminant analysis of student ratings of college teachers. Unpublished paper. The University of Michigan, Psychology Department, 1973.

Milholland, J. E. Measuring cognitive abilities. In McKeachie, W. J., Isaacson, R. L, and Milholland, J. E. Research on the Characteristics of Effective College Teaching. (Final report: Co-operative Research Project No. SAE 850, United States Office of Education and The University of Michigan.) Washington, D.C., U.S. Government Printing Office, 1964.

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

Pambookian, H. S. The effect of feedback from students to college instructors on their teaching behavior. Journal of Educational Psychology, in press.

Tuckman, B. W. and Oliver, W. F. Effectiveness of feedback to teachers as function of source. Journal of Educational Psychology, 1968, 59, 297-301.

APPENDIX

UNIVERSITY OF MICHIGAN
STUDENT PERCEPTION OF TEACHING

PLEASE PUT YOUR STUDENT NUMBER, COURSE,NUMBER AND SECTION NUMBER ON THE ACCOMPANYING IBM FORM, AS WELL AS ON THIS FORM!

Student No._____

Date_____

Section_____

Instructor_____

Your Grade Point - U of M

3.4 - 4.0_____
2.9 - 3.3_____
2.4 - 2.8_____
Below 2.4_____

First Semester Freshman
High School Rank

Top 5%_____
Top 25%_____
Below 25%_____

PLEASE INDICATE ON THE ACCOMPANYING IBM FORM YOUR REACTION TO EACH OF THE FOLLOWING STATEMENTS.

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

WRITE IN AFTER THE QUESTION ANY COMMENTS THAT YOU WISH TO MAKE. GIVE EXAMPLES WHEREVER POSSIBLE.

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

IMPACT ON STUDENTS

1. The instructor stimulates my intellectual curiosity.
   Comments:

2. I am learning how to think more clearly about the area of this course.
   Comments:

3. I am learning how to read materials in this area more effectively.
   Comments:

4. The instructor is effective in conveying the larger human context within which this subject lies.
   Comments:

5. I am acquiring a good deal of knowledge about the subject.
   Comments:

6. The course is making a significant contribution to my self-understanding.
   Comments:

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

7. The course is increasing my interest in learning more about this area.
   Comments:

8. I am generally bored in this class.
   If yes, why?

9. The instructor is enthusiastic.
   Comments:

10. The instructor gives good examples of the concepts.
    Comments:

11. The definitions and concepts given in class are generally clear.
    Comments:

12. The instructor goes into too much detail.
    Comments:

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

13. Students are confused.
    Comments:

14. The instructor is able to tell when students are confused.
    Comments:

15. The instructor is helpful when students are confused.
    Comments:

16. The instructor seems knowledgeable in many areas besides psychology.
    Comments:

RAPPORT

17. The instructor is permissive.
    Comments:

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

18. The instructor is friendly.
Comments:

19. The instructor invites criticism of his/her acts.
Comments:

20. It is very easy to learn to trust the instructor.
Comments:

TEACHER AS PERSON

21. The class is more pleasant than productive.
Comments:

22. The instructor spends so much time being "one of the gang", that we don't learn as much as we could.
Comments:

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

GROUP INTERACTION

23. Students volunteer their own opinions.
Comments:

24. Students argue with one another. (not necessarily with hostility).
Comments:

25. Students feel free to argue with the instructor.
Comments:

DIFFICULTY

26. The instructor assigns very difficult reading.
Comments:

27. The instructor asks for more than students can get done in the time available.
Comments:

0 = not applicable
1 = almost never or almost nothing
2 = seldom or little
3 = occasionally or moderate
4 = often or much
5 = very often
6 = almost always or a great deal

STRUCTURE

28. The instructor plans class activities in detail.
Comments:

29. The instructor follows an outline closely.
Comments:

FEEDBACK

30. Instructor keeps students informed of their progress.
Comments.

31. The instructor tells students when they have done a particularly good job.
Comments:

32. Tests and papers are graded and returned promptly.
Comments: