DOCUMENT RESUME

ED 061 071 SE 013 429

AUTHOR Waits, Bert K.; Elbrink, Larry C.
TITLE Student Evaluation of Mathematics Instruction.
INSTITUTION Ohio State Univ., Columbus.
PUB DATE [69]
NOTE 14p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Achievement; *Bias; College Mathematics; Correlation; *Course Evaluation; Mathematics Education; *Research; *Teacher Evaluation; *Test Validity

ABSTRACT

The purpose of this study was to determine whether students' responses to the Ullman student evaluation of teaching, the Student Evaluation of Teaching (SET) instrument, were biased by their achievement. The instrument was administered to all 325 students taking a second course in calculus with economic applications, near the end of the course; and separate evaluation ratings were obtained for the course, the teachers, and the examinations. An achievement rating (average midterm score) was also obtained for each student, and correlations between this rating and the SET ratings were investigated for each section of the class. Also, a one-way analysis of variance was run to investigate a possible relation between these ratings over the entire class. In most sections, the correlations between achievement and SET ratings were positive, but only 24 out of 48 were significant at the 5 percent level. The analysis of variance investigation revealed no further relationships. The authors conclude that the SET instrument under consideration may give unbiased evaluations for one teacher and biased evaluations for another, and, as such, is not to be recommended for general use. (MM)

STUDENT EVALUATION OF MATHEMATICS INSTRUCTION

Bert K. Waits
Assistant Professor of Mathematics
The Ohio State University, Columbus, Ohio 43210

Larry C. Elbrink
Visiting Assistant Professor of Mathematics
The Ohio State University, Columbus, Ohio 43210

Francis Dwyer of Pennsylvania State University, in his exhaustive survey of the literature on teacher evaluations, notes that evaluation ratings are subjective and have many inherent limitations. Many teachers feel that students are not capable of separating their personal feelings concerning a teacher from their evaluation of his teaching. Consequently, it is believed that teachers feel student evaluations are biased toward "liking" the teacher if they are doing well in the course and "disliking" the teacher if they are doing poor work in the course. Furthermore, such teachers often feel student evaluations are of no use (at least to them) and feel "threatened" if the results were to be available to others for interpretation. However, we cannot overlook the increased desire of students to evaluate what they experience, whatever that may be. Therefore, we need to know more about how to interpret SET (student evaluation of teaching) results. In this paper we examine one aspect of this interpretation - the relationship between student evaluation and student achievement.

The Problem Mathematics teachers in high school or college who teach required mathematics courses are particularly prone to feel threatened by student evaluations because many students view mathematics as difficult, unpleasant or uninteresting. If it can be shown (relative to a particular evaluation instrument) that students do give unbiased (with respect to course achievement) evaluations, then the classroom teacher could feel that valid student evaluations are beneficial and meaningful. It is the purpose of this paper to report on the results of a recent study concerning the usage of student evaluations of the mathematics teacher and of mathematics instruction.

The problem is stated in the following question. Do students give unbiased evaluations of their mathematics teachers? It is not our purpose to investigate the validity of student evaluations. (Students could give their teacher an unbiased, but also invalid, evaluation rating.)

The Investigation A SET instrument developed by Dr. Robert W. Ullman, Director of the Office of Evaluation at The Ohio State University, was used in the experiment. His instrument has been used at 28 different universities. The instrument consists of 48 questions divided into 3 categories - course, instructor and examinations. There is approximately a 50-50 split between positively phrased questions and negatively phrased questions. Students were given the choice of 4 responses - strongly agree, agree, disagree and strongly disagree. The instrument was given to all students enrolled in Mathematics 117 near the end of the course in the spring quarter 1969. Mathematics 117 is the second course in a calculus with economic applications sequence for non-math and non-physical science majors offered at The Ohio State University. There were 16 individual sections of the course taught by thirteen different teachers (advanced graduate students and instructors in the department). Each section instructor followed a common syllabus and all students took identical departmental examinations given in the evening.

The students were asked to specify their average midterm score in addition to the other information requested on the Ullman SET instrument. The forms were processed through a Digitek 100 optical scanning system (IBM card output). A computer program was written that assigned an evaluation rating (R) to each category (course, instructor and examinations) by the following formula: $R = \sum_{i=1}^{N} r_i$; N is the total number of questions in the category and $r_i = 3$ if the response to the i^{th} question is "strongly positive," or $r_i = 2$ if the response to the i^{th} question is "positive," or $r_i = 1$ if the response to the i^{th} question is "negative," or $r_i = 0$ is the response to the i^{th} question is "strongly negative." ("Strongly positive" equals a "strongly agree" response to a "positive" question while "strongly positive" equals a "strongly disagree" response to a "negative" question. Similarly for "positive," "negative," and "strongly negative.") The ranges of evaluation ratings were 0-60 for the course, 0-66 for the teacher and 0-18 for the examinations.

The following data were gathered for each student at the conclusion of the experiment - student number, section number (1 through 16), course evaluation rating, instructor evaluation rating, examination evaluation rating and average midterm score (range 0-100). Correlation statistics (Pearson) were developed for each section relating the achievement variable (average midterm score) with (1) the course evaluation rating variable, (2) the instructor evaluation rating variable and (3) the examinations evaluation rating variable.

The following were the correlation hypotheses of the experiment relative to each section.

H_1 (H_2,H_3): There is a significant positive correlation between student evaluation ratings of course (teacher, examinations) and student achievement.

These hypotheses were tested in the usual null formulation, denoted by H_1'.

H_1' (H_2',H_3'): The correlation coefficient between the variables of course (teacher, examination) rating and achievement is zero.

The graphs in Figure 1 contain correlation plots for two sample sections. In the first there is a significant correlation while in the second the correlation is not significant.

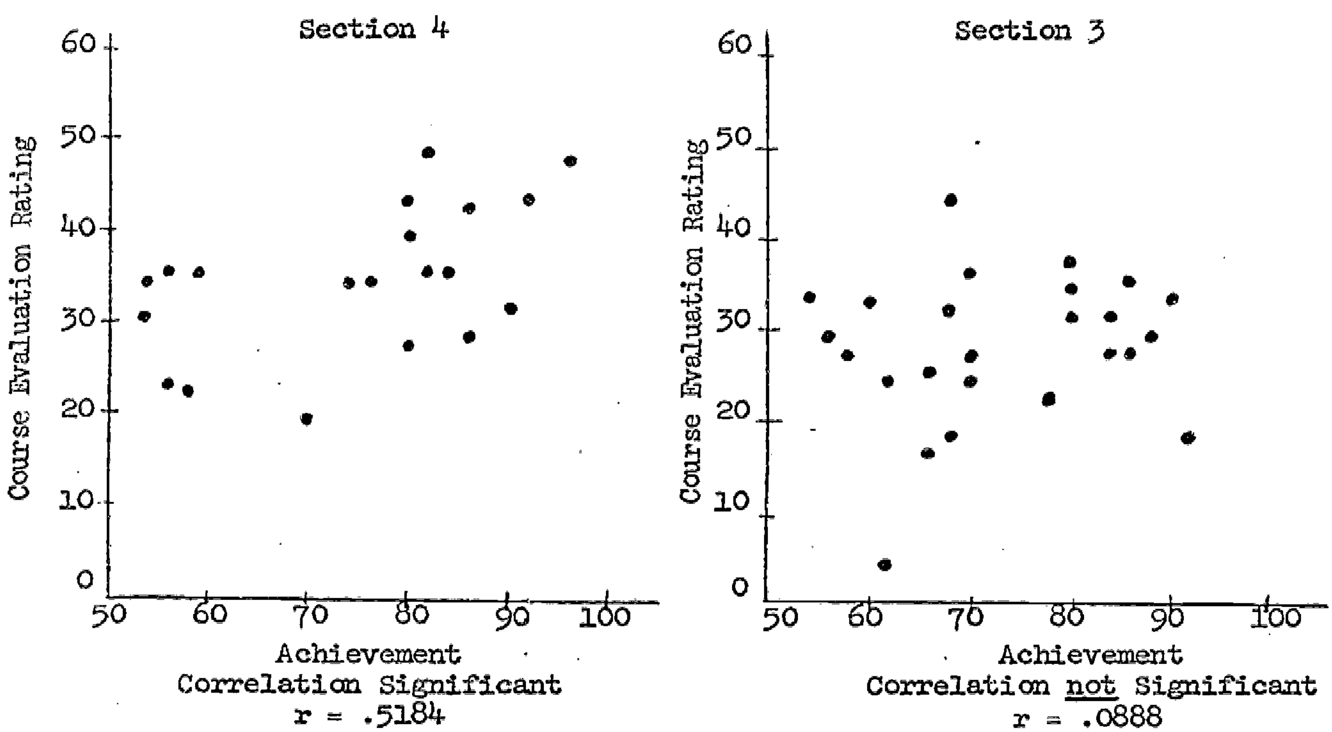

Figure 1. Course rating and achievement correlation plots.

In addition, the students were identified as belonging to one of five classification groups defined as follows:

Group 5 - average midterm score 90 or above (100 possible),

Group 4 - average midterm score between 79 and 90,

Group 3 - average midterm score between 69 and 80,

Group 2 - average midterm score between 59 and 70,

Group 1 - average midterm score less than 60.

A one-way analysis of variance model was applied with the course (instructor, examination) rating as the dependent variable and an effect for classification (defined above) as the independent variable. Relative to each section, the following were the null hypotheses tested in the analysis of variance investigation.

H_4 (H_5,H_6): There is no effect for achievement rank on the course (teacher, examination) evaluation ratings.

It is worth noting that if there is a significant non-linear relationship between the two variables, the analysis of variance model can detect differences in effects for achievement ranking that a correlation analysis can not. For example, suppose the correlation plot looked like the first graph in Figure 2. A correlation analysis would indicate no (linear) relationship where there is indeed an interesting (and significant) relationship. The F test should indicate that some significant relationship exists.

It should be noted that the F test analysis uses grouped data and consequently it is, in one sense, weaker than a correlation analysis. This means significant correlations will exist between the variables

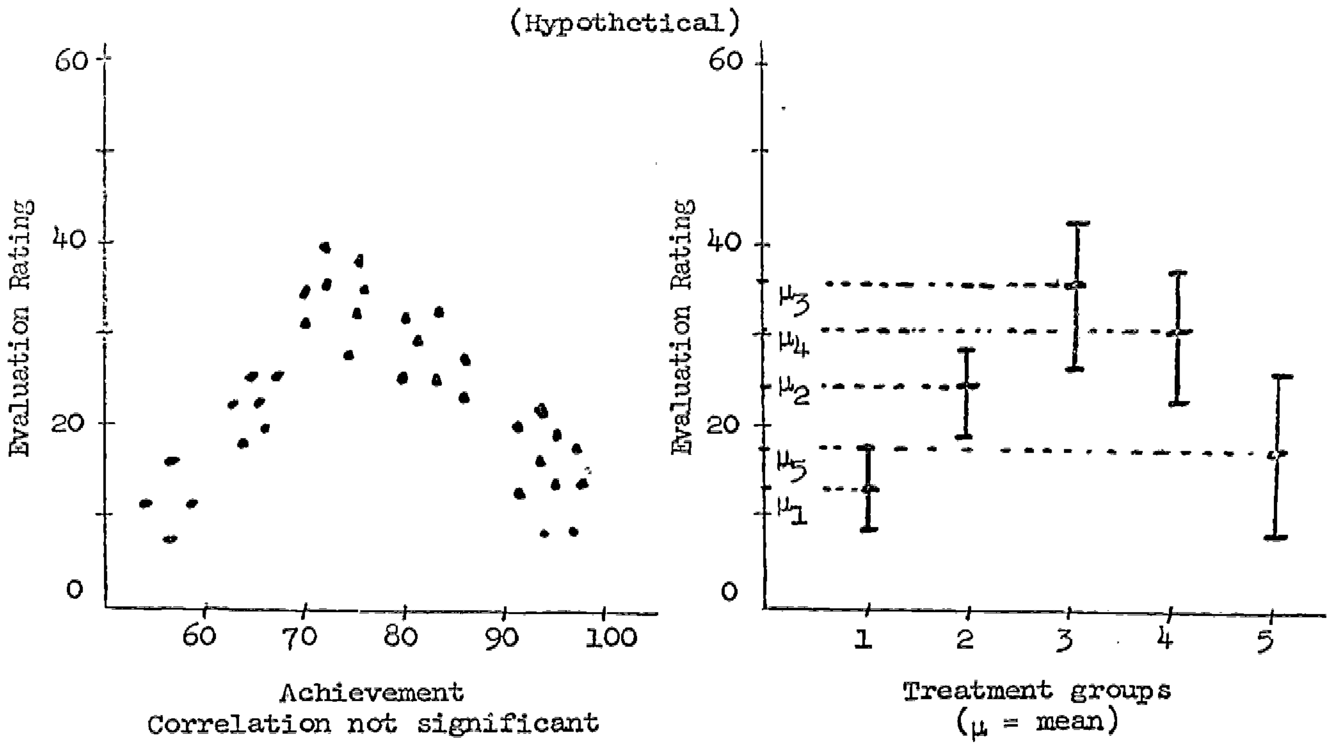

Figure 2. Hypothetical Example.

while the F test analysis indicates an insignificant relationship.

All hypotheses were tested at the .05 level of significance using two-tailed tests. The basic results of the study are summarized in the following tables. (r is the observed correlation coefficient and F is the calculated F statistic in the analysis of variance investigation.)

TABLE I. Course Rating Summary

Section Number	Sample N	r	Reject H_1'	F	Reject H_4
1.	18	.6556	Yes	7.5477	Yes
2.	21	.0881	No	0.8270	No
3.	25	.0888	No	0.7388	No
4.	20	.5184	Yes	1.6775	No
5.	17	.4368	No	1.8655	No
6.	25	.4685	Yes	3.5777	Yes
7.	15	.3718	No	1.8284	No
8.	15	.5324	Yes	2.3282	No
9.	22	-.0013	No	0.3160	No
10.	16	.3660	No	1.0853	No
11.	22	.4588	Yes	1.2698	No
12.	30	.3787	Yes	1.4766	No
13.	17	.0031	No	0.8956	No
14.	17	.1813	No	1.2148	No
15.	21	.5743	Yes	1.9105	No
16.	24	.1734	No	1.0117	No

TABLE II. Teacher Rating Summary

Section Number	Sample N	r	Reject H_2'	F	Reject H_5
1.	18	.5989	Yes	5.7424	Yes
2.	21	-.1225	No	0.8381	No
3.	25	-.0936	No	1.6780	No
4.	20	.3127	No	0.5119	No
5.	17	.3020	No	0.7033	No
6.	25	.3586	No	2.0142	No
7.	15	.5042	No	2.1399	No
8.	15	.3284	No	0.3465	No
9.	22	-.1013	No	1.7996	No
10.	16	-.0160	No	0.2982	No
11.	22	.0801	No	0.1997	No
12.	30	.2142	No	1.0632	No
13.	17	.3801	No	1.1310	No
14.	17	.3514	No	1.5377	No
15.	21	.4236	No	2.3657	No
16.	24	-.0848	No	2.2916	No

TABLE III. Examination Rating Summary

Section Number	Sample N	r	Reject H_3'	F	Reject H_6
1.	18	.6091	Yes	2.0520	No
2.	21	.3099	No	0.7124	No
3.	25	.2455	No	1.2285	No
4.	20	.3470	No	1.5865	No
5.	17	.6645	Yes	4.2614	Yes
6.	25	.4917	Yes	1.2300	No
7.	15	.0722	No	1.3297	No
8.	15	.3108	No	0.5188	No
9.	22	.2111	No	0.9740	No
10.	16	.5535	Yes	2.1796	No
11.	22	.5321	Yes	3.1690	No
12.	30	-.0235	No	0.1278	No
13.	17	.5452	Yes	2.0609	No
14.	17	.3314	No	1.7851	No
15.	21	.1991	No	0.3322	No
16.	24	.1408	No	1.2421	No

In Table IV means and standard deviation by section are noted for the variables in the correlation analysis. Table V contains the analysis of various data for the two sample sections shown in Figure 1. Complete data from the study is available from the authors by request.

TABLE IV. Means (μ) and Standard Deviations (σ)

	Achievement		Course Rating		Teacher Rating		Examination Rating	
	μ	σ	μ	σ	μ	σ	μ	σ
Section 1	74.3	14.8	31.5	11.3	45.3	10.6	9.9	3.9
Section 2	73.6	11.2	28.9	9.5	49.5	5.7	10.1	1.9
Section 3	73.0	11.3	28.2	7.1	49.7	5.8	10.2	2.2
Section 4	74.7	13.8	35.2	8.0	47.6	11.3	11.2	3.0
Section 5	66.7	17.4	23.0	8.5	32.0	8.6	9.9	2.9
Section 6	76.6	11.8	30.0	8.7	49.3	7.2	10.9	2.1
Section 7	59.4	13.4	24.4	7.6	42.0	9.4	8.7	2.4
Section 8	67.6	12.5	32.2	7.3	43.1	5.3	11.3	2.6
Section 9	79.1	8.7	31.4	6.7	49.3	6.7	11.0	2.2
Section 10	63.3	13.5	29.8	6.7	47.1	5.5	10.4	3.3
Section 11	71.0	15.3	29.1	8.4	46.4	8.4	9.5	3.1
Section 12	65.0	17.7	31.0	6.0	47.2	6.9	10.5	3.0
Section 13	68.4	12.8	26.2	7.9	38.1	12.7	8.8	2.4
Section 14	65.4	20.0	26.7	6.1	28.7	8.4	10.0	2.2
Section 15	78.0	13.4	33.7	9.5	52.3	6.6	11.8	2.6
Section 16	72.0	13.1	31.0	6.6	47.0	5.6	10.2	3.0

TABLE V. F Test Data

Section 4 (Course)
Number of Treatment Groups = 5

Treatment Group	Sample Size	Mean	Standard Deviation
1	6	30.8333	5.9805
2	0	0.	0.
3	3	30.0000	8.6602
4	8	38.1250	7.3180
5	3	41.3333	8.3267

Analysis of Variance Table

	Sum of Squares	DF	Mean Square	F-Ratio
Between Groups	376.8255	4	94.2064	1.6775
Within Groups	842.3698	15	50.1580	
Total	1219.1935	19		

Section 3 (Course)
Number of Treatment Groups = 5

Treatment Group	Sample Size	Mean	Standard Deviation
1	3	30.0000	2.6457
2	8	25.7500	10.1524
3	4	27.2500	6.1847
4	8	31.3750	3.7009
5	2	25.5000	10.6066

Analysis of Variance Table

	Sum of Squares	DF	Mean Square	F-Ratio
Between Groups	156.4150	4	39.1037	.7388
Within Groups	1058.6251	20	52.9313	
Total	1215.0401	24		

Conclusions and Recommendations It is apparent (after considering the results summarized in Tables I, II and III) that only in section 1 was the correlation between evaluation rating and student achievement significant (the higher evaluation ratings were associated with the "good" students while the lower ratings were assoicated with the "poor" students). In the fifteen remaining sections the correlation between evaluation rating and student achievement was statistically not significant. However, in eight of these sections the observed correlation coefficient was positive and the "no correlation" hypothesis was near the rejection level.

In almost one-half of the sections the correlations between course evaluation rating and student achievement and the correlation between examination evaluation rating and student achievement were significant. Again in both cases the better students gave the more favorable evaluation ratings. It is important to observe that the analysis of variance investigation indicated there was no significant new (not previously implied by the correlation analysis) relationship between the SET ratings and student achievement scores. What do all these statistics mean to the mathematics teacher?

We can conclude that some teachers (even those teaching required mathematics courses) might expect to receive unbiased evaluation ratings from their students. Also we can conclude that the ratings of some teachers will be biased, the favorable ratings coming from the good students and less favorable ratings coming from the poorer students. We assume that a low correlation is evidence of unbias. However, a low correlation could result from instrument insensitivity or from the small numbers (15-30) involved in each section. We also assume the evaluation ratings obtained from the Ullman instrument are valid. It has been our experience that this is usually the case. We found that the SET ratings would rank most of our teachers in the same order as we would. However, some staff members that we know to be excellent teachers from our frequent observations and other close contacts are sometimes rated unfavorably. We cannot recommend any evaluation instrument (including the Ullman instrument) nor can we recommend that administrators encourage student evaluations without a careful investigation of psychological and "local" considerations. A word of warning is in order. It is possible that inexperienced teachers might direct their teaching activities toward developing "favorable" evaluation ratings. If SET results are made "public" (used in salary and promotion considerations), then even experienced teachers might also direct their teaching activities towards developing "favorable" ratings! Such activities need not be in the students' best interest and possibly would result in ineffective teaching. We do recommend that an instructor who uses any evaluation instrument perform a correlation or regression analysis on the variables (evaluation rating and student achievement) and then interpret the results accordingly.

REFERENCES

1. Blackman, A. F., et al, "Students Rate Their Professors and Courses," Phi Delta Kappa, XLVIII (Fall, 1967), 266-269.

2. Brickman, W. W., "Student Evaluation of Professors," School and Society, XCIV (March 19, 1966), 143.

3. Brogan, H. O., "Evaluation of Teaching," Improving College and University Teaching, XVI (Summer, 1968), 191-192.

4. Bruan, R. C., "Student Rating of Teachers," Improving College and University Teaching, XVI (Summer, 1968), 200-202.

5. Dwyer, Francis M., A Review of Characteristics and Relationships of Selected Criteria for Evaluating Teacher Effectiveness, Unpublished Monograph, University Division of Instinctional Services, Pennsylvania State University, 1968.

6. Kent, L., "Student Evaluation of Teaching," Educational Record, XLVII (Summer, 1966), 376-406.

7. Musella, D. and Rusch, R., "Student Opinion on College Teaching," Improving College and University Teaching, XVI (Spring, 1968), 137-150.

8. Quarles, N. K., "Evaluation: Teacher Rating by Pupils," Agriculture Education Magazine, XXXVIII (June, 1966), 270-271.

9. Samalonis, B., "Ratings by Students," Improving College and University Teaching, XV (Winter, 1967), 11.

10. Slobin, D. Y. and Nichols, D. G., "Student Rating of Teaching," Improving College and University Teaching, XVII (Autumn, 1969), 244-248.