DOCUMENT RESUME

ED 222 526 TM 820 674

AUTHOR Goodwin, Stephanie S.; Raths, James
TITLE Global and Specific Ratings: Are They Related?
PUB DATE Mar 82
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Feedback; *Global Approach; *Higher Education; *Instructional Improvement; Lecture Method; *Student Evaluation of Teacher Performance; Teacher Workshops; Videotape Recordings

ABSTRACT

The purpose of this study was to determine if making specific changes in teaching, changes that were validated by experts, would change the overall student ratings of instructors. Thirty graduate teaching assistants who participated in a workshop on improving lecturing skills significantly improved their scores on specific lecturing behaviors that were taught in the workshop when videotapes of their teaching were evaluated by experts. A group of 18 undergraduate students then assessed these same videotapes giving an overall rating of the instructor and a rating of the instructor's warmth. An analysis of the data generated the conclusion that improvement on the part of instructors in specific areas is not likely to affect their global ratings. Additionally, student perceptions of an instructor's warmth play a large part in determining the overall rating the students give the instructor, whereas, specific ratings do not seem to account for the overall global rating. The data appear to support the idea that, while specific skills can be changed through intervention, the overall impressions that students have of their instructors do not necessarily improve. (Author/PN)

Global and Specific Ratings:

Are They Related?

Stephanie S. Goodwin and James Raths

University of Illinois at Urbana-Champaign

Paper prepared for presentation at the American Educational Research Association Annual Meeting; New York, New York; March 19-23, 1982
Session 52.17.

Problem and Research Questions

In many institutions, student ratings of instructors are used for at least two purposes: (1) to give feedback to instructors for the purpose of improving their teaching, and (2) to serve as evidence in making promotion or tenure decisions. A recent meta-analysis of student ratings suggests that global ratings are more valid than specific ratings in predicting subsequent global ratings, ratings from peers, and even in some limited instances student achievement (Cohen, 1982). Indeed, it is often the practice to only include the global ratings in the promotion and tenure recommendation papers that are prepared in support of a faculty member.

Because of the seriousness of the student rating outcomes many institutions offer help for instructors who receive less than satisfactory evaluations. Such persons may avail themselves of a tutoring service of sorts whose principal aim is to improve instruction on campus. Measures of the effectiveness of this service are often taken from changes in student ratings. Thus, if an instructor receives higher ratings from students after receiving help, the interventions are judged to be successful. If the ratings do not change, then the interventions are deemed benign.

Of course, it is difficult to help someone improve his instruction on an "overall" basis. The usual practice of instructional tutors is to scan the student evaluations to find areas of weakness, at least as perceived by students, and to help the instructor in those areas. For example, an instructor with low student ratings on an overall basis may also receive low ratings in areas such as (1) preparing examinations, (2) discussion

skills, or (3) lecturing. Given this information, instructional experts can help the instructor improve in those specific areas.

The research questions to which this study was addressed were:

1. If an instructor changes in specific ratings of instruction (on areas targeted for improvement) will his global ratings change correspondingly?
2. What is the relationship between overall effectiveness ratings, warmth ratings, and specific ratings of instruction?

Methodology

In the fall of 1979, thirty-seven teaching assistants, almost all of whom were new instructors, were asked by their department chairs to participate in a program designed to improve their instructional efforts. All of the teaching assistants participated in a workshop on effective lecturing techniques. The teaching assistants also taught two ten-minute mini-lessons which were videotaped. One mini-lesson preceded the workshop and one followed it. These videotapes were rated by a panel of three experts on an instrument which assessed ten specific lecturing behaviors taught in the workshop. The instrument used by the experts is in Appendix A. The inter-rater reliability for the experts on this instrument was .94. As a result of the workshop the teaching assistants significantly improved their pre- to post-measures on a composite score of the ten lecturing behaviors (Sharp, 1981).

Our research made use of the artifacts generated by the previous research—namely the pre- and post-videotapes. Also we used the experts' composite score for each teaching episode which we called the specific rating for each lesson. A panel of eighteen undergraduates in the field of education was asked to rate the pre- and post-videotapes on six items:

1. Rate the instructor's overall teaching ability.
2. Rate your interest in the content.
3. Rate the instructor's lecturing ability.
4. Rate the instructor's warmth.
5. Rate the instructor's statement of objectives for the lesson.
6. Rate the instructor's ability to establish and maintain eye contact during the lesson.

Each item was rated on a five-point scale, with labels on a continuum from very poor to excellent. Each of the eighteen student raters viewed a set of thirty pre- and post-videotapes. Seven sets of tapes were not usable. The pre- and post-videotapes were viewed in a random order, and students never saw the pre- and post-tapes for the same teaching assistant back to back. The inter-rater reliability for the student ratings was computed, using a model described by Winer (1962). Inter-rater reliability for the Global Item 1, was .94.

Findings

The mean scores and standard deviations for the pre- and post-ratings on the six evaluation items and on the experts' specific ratings can be found in Table 1. The pre- and post-average mean on item 1, "rate the instructor's overall teaching ability," were the same.

Table 1

Mean Scores and Standard Deviations
for Pre- and Post-Measures

Variable	Pre-Rating X (N = 30)	Pre-Rating S.D. (N = 30)	Post-Rating X (N = 30)	Post-Rating S.D. (N = 30)
Student Ratings				
Item 1 Rate the instructor's overall teaching ability	3.33	.75	3.33	.69
Item 4 Rate the instructor's warmth	2.96	.71	3.12	.75
Expert Ratings				
Specific Lecturing Behaviors	20.66	4.67	24.89	4.00

The instructors were classified into a two-by-two table based on the student ratings they received on Item 1,"rate the instructor's overall teaching ability," and the experts' specific ratings of their lecturing skills. The results of this analysis are in Table 2.

Table 2

Relationship between Improvement in Lecturing Skills
and Improvement in Overall Teaching Ability

		Lecturing Skills As Assessed by Experts: Showed Improvement	Lecturing Skills As Assessed by Experts: Did Not Show Improvement	
Global Ability As Assessed by Students	Showed Improvement	13	1	14
	Did Not Show Improvement	13	3	16
		26	4	30

Chi-square - .10 (n.s.)

The hypothesis of "no relationship" was not rejected by this analysis.

A stepwise multiple regression was performed to determine the extent to which a linear combination of two or more predictor variables could account for the variance in the criterion variable (final global ratings). In step 1, "warmth" was determined to be the predictor variable which explained the greatest amount of variance in the global rating. R^2 for this variable was .57. In step 2, initial global ratings were chosen in conjunction with warmth. The R^2 using these two variables was .77. In step 3, statement of objectives was chosen in conjunction with warmth and initial global ratings. The R^2 using all three variables was .83. The final variable which significantly improved the prediction of final global ratings was maintaining eye contact. The multiple regression equation with all four predictor variables yielded an R^2 of .86. Specific ratings and student interest did not meet the criterion for significance ($p < .05$) and therefore did not enter into the multiple regression.

Conclusions

The following conclusions seem warranted by the findings of this study:

1. Improvement on the part of instructors in specific areas are not likely to affect their global overall ratings.
2. Students perceptions of an instructor's warmth play a large part in determining the overall rating the students give the instructor whereas, specific ratings do not seem to account for the overall global ratings.

Limitations

First, the findings and the interpretations need to be qualified by several factors which delimit the study. They include:

1. The study was in effect a simulation of what happens in real life. While the sample of instructors utilized in this study were naive to teaching and could reasonably be expected to benefit from the interventions that were given to them, they were not in fact representative of teachers seeking to improve overall ratings after some successive failures in that area.

2. The student raters were not representative of the student body at large - but were education majors. It may be that characteristics of persons entering education are quite different from others on campus.

3. The ratings were performed on videotapes and not "live." It may be that student ratings are artificially altered in unknown ways by rating instructors in this fashion.

4. This study only investigated an intervention in the area of lecturing. If the intervention dealt with discussion skills, or preparing more effective examinations, perhaps different patterns would have resulted.

5. Experts were used to document changes in lecture skills because it was anticipated that the students, less sophisticated in making evaluations, would equate lecturing skills with overall skills. (Actually, the correlation between these ratings by students was .99). It is the case that students did not perceive differences in lecturing skills as did the experts; it may be that the experts might have rendered greater "overall" ratings to match their greater specific ratings. Those data were not collected. However, it is usually the case that experts judge the efficacy of interventions such as those used and that students present the university officials with the ultimate criterion - their overall ratings - so in effect this design matches fairly well with how things work in the "real world."

In spite of these limitations, the data seem to support the idea that while specific skills can be changed through intervention, the overall impressions that students have of their instructors do not necessarily also improve.

Implications

1. People who are in the "improving teaching" business must be extremely cautious in making claims about the efficacy of their treatments. First, efficacy may well vary with the measures that are taken as evidence. If the measures reflect changes in specific teacher behaviors, it may be the case, indeed it may be likely, that the overall ratings of instructor will not change.

   This factor is specially important, given the fact that it is the overall rating that has the best predictive validity. Changing behaviors without changing overall ratings looks like a waste of time.

2. As suggested above, the selection of an appropriate criterion variable to assess programs of teacher improvement seems problematic. While it would appear that programs should be assessed on what they promise, and if they promise to only change specific behaviors, then that should be the litmus test for those interventions. On the other hand, knowing that specific ratings are not productive, and that changes in specific behaviors are not likely to change overall behaviors, then perhaps the key question is "how have the overall ratings changed?"

3. It appears that students perception of an instructor's warmth affects his overall rating. This information probably needs to be passed on to instructors.

4. The question has been raised of how instructional tutors can begin to change overall ratings. At this point we don't really know how to change those ratings.

References

Cohen, Peter A. "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multi-Section Validity Studies." Review of Educational Research 51 (1982): 281-309.

Sharp, Gregory. "Acquisition of Lecturing Skills by University Teaching Assistants: Some Effects of Interest, Topic Relevance, and Viewing a Model Videotape." American Educational Research Journal, 18 (Winter, 1981): 491-502.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw Hill, 1962.

APPENDIX A

#_____

LECTURE PRESENTATION SKILLS EVALUATION FORM

	Yes	Not Sure	No	WOW + -
1. Used an "attention getting" device at the beginning of the lesson.	____	____	____	____
2. Stated objectives or goals for the lesson.	____	____	____	____
3. Attempted to monitor student progress through use of questions.				
a. First 3 minutes	____	____	____	____
b. Second 3 minutes	____	____	____	____
c. Third 3 minutes	____	____	____	____
4. Defined terminology.	____	____	____	____
5. Provided for closure.	____	____	____	____
6. Established and maintained eye contact with the group.				
a. First 3 minutes	____	____	____	____
b. Second 3 minutes	____	____	____	____
c. Third 3 minutes	____	____	____	____
7. Spoke in a conversational manner.				
a. First 3 minutes	____	____	____	____
b. Second 3 minutes	____	____	____	____
c. Third 3 minutes	____	____	____	____
8. Provided vocal variety for emphasis.	____	____	____	____
9. Used gestures to reinforce or complement verbal statements.	____	____	____	____
10. Acknowledged student responses/ contributions.	____	____	____	____