THE JOURNAL OF EDUCATIONAL RESEARCH
(Volume 68, Number 1, September 1974)

Teacher Evaluations and Student Learning: A Reexamination

REBECCA BRYSON, San Diego State University

ABSTRACT

The study was conducted to reexamine the question of the relationship between teaching skill as evaluated by college students and teaching skill as evidenced by how much the instructor's students learn. Mean scores on each of fourteen evaluation items for instructors of twenty sections of college algebra were correlated (across sections) with mean post-course performance of their respective students on the Cooperative Intermediate Algebra Test. For all evaluation items the relationship with amount learned was positive. These results are in direct conflict with results recently reported by Rodin and Rodin (7). Reasons for the discrepancy are discussed.

THERE HAS been considerable recent controversy about whether or not student evaluation of instructors is related to amount learned by students. This has to a large degree been evoked by Rodin and Rodin's article (7) which demonstrated a substantial negative correlation ($r = -.72$) between mean score (on a global evaluation item) of each of a set of calculus teaching assistants (TA's) and mean number of units covered by each of the TA's students.

These results run counter to virtually all previously published studies (e.g., references 1, 2, 3, 4, 5, 6), which, in general[1] show low positive correlations between indicators of teaching effectiveness and amount learned by students. Three related sources of error which tend to occur in these studies have been cited as possible reasons for the obtained relationships being low. These are: (1) lack of comparability across evaluated sections in what is taught, (2) biased sampling of items in tests constructed to assess what is learned, and (3) teaching to examinations which are non-exhaustive samples of course content.

Paradoxically, the Rodins carefully standardized course content, utilized an unbiased measure of amount learned in that it exhausted content, and took precautions to prevent teachers from seeing the actual test problems prior to administering them, yet obtained a strong negative rather than increased positive relationship. However, these controls made the evaluated TA's serve a role which was quite different from that of instructors evaluated in previous studies. All sections met together three times per week for a common lecture presented by the professor in charge of the course, then with the TA's twice weekly, once for review and the other time for testing. The tests were paradigm problems for each unit covered, and it was possible to take variants of each problem if the problem was not solved perfectly the first time. In essence, the role of the TA's was to review material and administer tests rather than to introduce material.

The present study was conducted to determine whether or not the results obtained by Rodin and Rodin could be generalized to a new situation where content coverage was controlled: the test administered was a standardized examination with questions representative of course content, the teachers did not see the examination prior to its administration, and at the same time, the evaluated teachers had primary responsibility for teaching the course.

Additionally, it seemed plausible that the relationship between student learning and teacher evaluations is, to some extent, dependent upon which specific behavioral characteristics the student is asked to evaluate. To examine this possibility, a set of fourteen items (including a global evaluation item equivalent to the one used by the Rodins) was examined to determine what, if, any, components of evaluated teaching were related to knowledge gains on the part of students.

Method

Subjects included students ($N = 582$) enrolled in twenty sections of college algebra and fourteen instructors, six of whom taught two sections and eight of whom taught one section of the course. The instructors taught from a common syllabus and used a common textbook/workbook.

Twelve items dealing with instructor characteristics were selected for analysis from a routinely administered faculty and course evaluation form. Two additional items were included, one equivalent to the global Rodin and Rodin item, and another asking students, "Apart from your personal feelings about this instructor, has he/she been instrumental in increasing your knowledge of mathematics?"

Evaluation forms were administered in January 1973, during the last 2 weeks of the fall semester. These were completed by 89 percent of the enrolled students. Mean scores on each item were calculated for each instructor. During final

Copyright © 2001. All Rights Reserved.

examination week the Cooperative Intermediate Algebra Test (CIAT) was administered, and mean scores of students were calculated for each class. Correlations were obtained between examination performance of the separate sections and evaluation of instructors of the sections on each of the evaluation items.

Additionally, pretest scores on the CIAT were available for 39 percent of the students, those who had attempted to exempt the course. Because evaluations were completed anonymously it was impossible to match evaluations with pre- and posttest scores. However, students were asked on the supplementary evaluation form whether or not they had taken the pretest. It was therefore possible to analyze the relationship between mean gains scores for sections and mean evaluation on the two supplementary questions, using evaluations supplied by those Ss for whom gains scores were available.

Results

All relationships between teacher evaluation and amount learned (CIAT score) were positive (see Table 1). On the two items where it was possible to compare gains scores with evaluations, the results closely approximated those obtained using the final CIAT scores (see footnotes in Table 1). Furthermore, gains scores and final CIAT scores were highly correlated ($r = .89$), indicating a fairly high degree of comparability between the two means of assessing amount learned.

Table 1.—Correlations Between Evaluations of Teacher and Student Performance

Evaluation Items	
Establishes good rapport in a friendly and supportive way	.36
Presents lecture material in an expressive and clear manner	.37
Demonstrates dynamism and enthusiasm for his subject	.44*
Demonstrates confidence and mastery within his subject	.19
Gives clear and concise answers to questions	.34
Communicates a genuine desire to teach students	.51*
Seems to keep lecture material updated	.30
Makes a serious effort to keep students appraised of their progress as the course progresses	.17
Carefully listens to and evaluates student opinion	.39
Gives directions for papers or assignments in a clear, logical manner so that requirements are clearly understood	.10
Objectively evaluates grades on adequate samples of each student's ability and performance	.30
Would you take a course from this instructor again?	.46*
Mean of the above items	.48*
Rate your teacher on a scale ranging from A to F where A represents an excellent teacher, and F, a miserable one	.[a]55**
Apart from your personal feelings about the instructor, has he/she been instrumental in increasing your knowledge of mathematics?	.[b]68**

[a, b] Correlations with gains scores were .56 and .70 respectively.
* $p < .05$; ** $p < .01$.

Evaluation items which were significantly ($p < .05$) related to amount learned included the following: "Demonstrates dynamism and enthusiasm for his subject," "Communicates a genuine desire to teach students," "Would you take a course from this instructor again," "Grade the teacher on a scale of A to F," "Apart from your personal feelings about the instructor has he/she been instrumental in increasing your knowledge of mathematics," and total score on the evaluation form.

Discussion

These results imply that evaluation of an instructor is directly related to amount learned from that instructor. The clear discrepancy between these findings and those reported by the Rodins invites further comment. While it had initially been assumed that general evaluation was perhaps more dependent on pleasantness, amount of fraternization with students, or some other variable which could at the same time be inversely related to teaching proficiency, this was apparently not the case in the present study. General evaluation (on an A to F scale) was highly and positively related to both gains scores and final examination scores.

There are, however, at least three major differences in the manner in which the two studies were conducted which may effect the discrepant results: (1) differences in amount of responsibility the evaluated instructors had for their classes, (2) differences in the criterion of amount learned, and (3) differences in samples from which evaluations were collected.

Because the role of the TA's in the Rodins' study was a relatively minor one in terms of content communication, a nonsignificant relationship would be easy to accept. It is more difficult to reconcile the significant negative relationship. The explanation of such a relationship requires postulating either another implicit criterion of evaluation which is negatively related to amount learned, or a disdain for teaching assistants teaching the sort of information which maximizes test performance. The possibility that teaching assistants in the Rodins' study were evaluated on the basis of personality or likeableness falls into the first category. The role of the teaching assistant may have been that of peer rather than authority, evoking a likeableness criterion of judgment. This point is given some credence by the fact that class size (a probable indicator of popularity was, in the Rodin study, positively related to evaluation ($r = .31$) and negatively related to number of units completed ($r = -.64$).

Copyright © 2001. All Rights Reserved.

It would seem that performance on tests designed by the professor in charge of the course would be maximized by the presentation of information which was almost totally redundant with that presented by the professor in charge. At the same time, such drill may not be positively regarded by students in a relatively advanced mathematics course. At any rate, introduction of new material and novel insights and applications by the teaching assistants would probably have little payoff in terms of performance on these examinations.

In the present study instructors were told what should be covered, but were responsible for introducing the material in any manner they chose. They had complete responsibilty for their classes, including the discretion to use the final examination scores as they saw fit in assigning final grades.

Another important difference between the two studies is in the criterion of amount learned. In the present study, amount learned was assessed by performance on a common examination and by gains demonstrated on that examination over performance on the same test before entering the course. In the Rodins' study the objective criterion of amount learned was number of problems solved at mastery level during the semester. Two possible problems emerge at this point. Since class size was negatively related to class performance, it may be that the TA's responsible for large classes simply had difficulty in providing sufficient individual tutoring and motivation for students to take tests. The other problem is the possibility of an artifact which may be a general concern in viewing evaluations collected at the end of a course where students are to some extent responsible for pacing themselves. Students who complete enough units to earn an A in the course prior to the end of the semester have less reason to be present on the day evaluations are collected. In their reinterpretation of the Remmers studies, which demonstrated an average near zero correlation within classes between amount learned and evaluation, the Rodins point out that the range of these correlations is rather large and suggest that the relationship may be negative for instructors who pitch their lectures at a low level. The general suggestion is that students appreciate instructors who direct lectures to their own performance level. Extending this argument to the across instructor situation, it follows that the relatively poorer students should evaluate positively the teachers whose classes (on the average) did least well.[2]

Because 89 percent of the Ss used in the present study completed evaluations, it is unlikely that sample discrepancy had a major effect on the results. However, if the persons who were absent were those who knew least, the obtained correlations could perhaps be regarded as inflated over what would be obtained from the total sample.

Implications

This study taken in conjunction with most of the previously reported data suggests that evaluations of teacher do reflect how much the teacher has taught his students. While the results are diametrically opposed to those reported by the Rodins, the conflict may be a function of differences in the contexts in which the two studies were conducted. Because results may to some extent be dependent upon type of course, level at which the instructor directs his lectures, and individual differences among students in the relative importance ascribed to various components of perceived teaching effectiveness, it would seem wise to investigate the effects of these variables separately and in combination. Meanwhile, however, the bulk of accumulated evidence would suggest that in the absence of other valid criteria of teacher performance, the continued use of student evaluations is warranted.

FOOTNOTES

1. Gessner (1) obtained large positive correlations, but methodological difficulties with his study preclude its comparison with the others.
2. Data supplied by the Rodins show that self-reported grade in the previous course in the calculus sequence was slightly higher for those who filled in the evaluations than was actual grade (obtained from the registrar) for the total class. Whether this is due to biased self report or to true differences is unknown. The extent to which differences in performance existed in the *evaluated* course between those who completed evaluations and those who did not is the more important question and cannot be determined.

REFERENCES

1. Gessner, P. K., "Evaluation of Instruction," *Science*, 180:566–570, 1973.
2. McKeachie, W. J.; Lin, Y.; Mann, W., "Student Ratings of Teacher Effectiveness: Validity Studies," *American Educational Research Journal*, 8:435–445, 1971.
3. McKeachie, W. J.; Solomon, D., "Student Ratings of Instructors: A Validity Study," *Journal of Educational Research*, 51:379–382, 1958.
4. Meinkoth, M., "Teachers of Economic Principles: Effect on Student Achievement and Attitudes," *Journal of Experimental Education*, 40:66–72, 1971.
5. Morsh, J. E.; Burgess, G. C.; Smith, P. N., "Student Achievement as a Measure of Instructor Effectiveness," *Journal of Educational Psychology*, 47:79–88, 1956.
6. Remmers, H.; Martin, F.; Elliot, D., "Are Students' Ratings of Instructors Related to Their Grades?" *Purdue University Studies in Higher Education*, 66: 17–26, 1949.
7. Rodin, M.; Rodin, B., "Student Evaluations of Teachers," *Science*, 177:1164–1166, 1972.

Copyright © 2001. All Rights Reserved.