DOCUMENT RESUME

ED 065 509 TM 001 029

AUTHOR Centra, John A.
TITLE Evaluating College Teaching: The Rhetoric and the Research. Research Memorandum.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO RM-72-3
PUB DATE Mar 72
NOTE 12p.; Paper presented at Conference of American Association for Higher Education, Chicago, Illinois, March 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *College Teachers; Educational Research; *Evaluation Methods; Faculty Promotion; Instructional Improvement; Rating Scales; *Self Evaluation; *Student Opinion; *Teacher Evaluation

ABSTRACT

Methods and reasons for evaluating teaching are discussed, and an experimental study of the effectiveness of students' ratings of teachers is described. The two main reasons for evaluating teaching as given in this paper are (1) to help make decisions about whom to promote, and (2) to improve instruction. In the experimental study, five diverse colleges participated. A total of some 470 faculty members were randomly assigned within each institution to one of three groups--feedback within a week (treatment group); no feedback, with summary of results given at end of the semester (control group); and posttest, which used rating form only at the end of the semester to determine whether simply using the form caused teachers to change, even without feedback. A 23-item form eliciting instructional procedures or behavior that an instructor could presumably change was used in the study. Results showed that instructors who received student feedback did not noticeably modify their teaching practices. A second aspect of the study was to determine to what extent instructors describe or rate their teaching differently from the students' ratings. Items from the student form were reworded slightly for instructor responses. It was found that there was a significant difference between instructor and student responses to most items, with instructors rating their teaching in more positive terms. The use of student pre- and post-test scores as a means of evaluating the effectiveness of teaching are seen as beneficial to the teacher, but their use as the sole criterion for determining teaching effectiveness is not advocated. Suggestions are made as to other evaluation techniques. (DB)

RM-72-3

RESEARCH MEMORANDUM

EVALUATING COLLEGE TEACHING: THE RHETORIC AND THE RESEARCH

John A. Centra

Paper presented at the Conference of the American Association for Higher Education, Chicago, March 1972.

Educational Testing Service
Princeton, New Jersey
March 1972

Evaluating College Teaching: The Rhetoric and the Research[1]

John A. Centra

Educational Testing Service

There are two major reasons for evaluating teaching: one is to help make decisions about whom to promote, and another is to improve instruction. The two need not, of course, be mutually exclusive. The title of this session poses the question "Can teaching be evaluated?", but a more proper question, it would seem, is whether evaluation can lead either to valid decisions on promotion or to the improvement of instruction.

There is more rhetoric than hard evidence offered in answer to the question in spite of a half-century of research on teaching. This is not to say that the research efforts have been totally fruitless; more likely the findings have not, for a variety of reasons, been put into practice.

It would not be difficult to criticize current faculty promotion practices. At larger institutions where the educational goals include advancing knowledge as well as passing it on, there is little question that faculty members are rewarded primarily for their efforts with the former even though lip service is given to the latter. The ability to publish, however, is at best only modestly related to teaching effectiveness according to the findings of various studies.[2] At smaller institutions where teaching is the

---

[1]Paper presented at the Conference of the American Association for Higher Education, March 1972.

[2]See, for example, D. P. Hoyt. Instructional effectiveness inter-relationships with publication record and monetary record. Office of Educational Research, Research Report #10, Kansas State University, May 1970.

primary function, recent surveys tell us that the judgments of one or more administrators are most frequently relied on to assess teaching effectiveness. Those judgments are often based on hearsay or on very slight evidence.

Systematic efforts to evaluate teaching have most typically had as their goal the improvement of teaching, and one of the most common methods employed has been the use of student ratings. Who else, it is argued, is in a better position to tell an instructor how to improve his course than the student-consumer? Indeed it has practically become part of the rhetoric or folklore on teaching to expect that student ratings will enable the instructor to improve his teaching.

Underlying the use of students' ratings for instructional improvement are the assumptions that, first, the instructor values student judgment enough to change his procedures when called for, and second, that the instructor learns something about his teaching from students that he does not already know. Both assumptions are open to question.

In an attempt to investigate these assumptions, I recently undertook an experimental study supported by a grant from the Esso Education Foundation. Five diverse colleges which did not have a formal program of student ratings of teaching participated in the study. They included two state colleges, one of which had a predominantly black enrollment, a selective liberal arts college, a multipurpose college, and an urban community college. A total of some 470 faculty representing over three-quarters of those teaching cooperated in the study; in fact, the cooperation of the institutions and their faculties could not have been much better.

Within each institution, teachers were randomly assigned to one of three groups:

1. The feedback group, which administered a student rating form at midsemester and received a summary of the results within a week, along with some comparison data to aid in interpretation. In research terms this is the "treatment" group, with the treatment being essentially what is done at most colleges which use student ratings for instructional improvement;
2. The no-feedback group, which used the rating form at midsemester but did not receive a summary of results until the end of the semester. This would be the so-called "control" group;
3. The posttest group, which used the rating form only at the end of the semester in order to determine whether the midsemester ratings had a sensitizing effect on teachers in the no-feedback group; i.e., whether simply using the form caused teachers to change, even without getting feedback.

In addition to using the rating form at midsemester, the feedback and no-feedback groups also administered the form at the end of the semester. Both midsemester and end of semester ratings were collected during fall semester of 1971. A single semester instead of two successive semesters was used for the study to enable the same students to provide both sets of ratings. Moreover, a suggestion sometimes made is that instructors should obtain feedback from students at midsemester so that students who provide the information might benefit from their own suggestions.

A 23-item form eliciting instructional procedures or behavior that an instructor presumably could change was used in the study. Included were items that I had asked faculty members in an earlier study to identify as providing

information they would like to have from students.[3] Among the areas included were those dealing with the organization of the course, the clarity of objectives and presentations, and the instructor's helpfulness or availability to students. Several of the items may be found, with slight variations, in a number of current student rating instruments.

If student feedback improved instruction, we would expect the end-of-semester ratings of the feedback group to be better than either the no-feedback or the posttest group. They were not. In fact, the three groups were nearly identical in their scores for each of the items, indicating that the group of instructors who received student feedback did not noticeably modify their teaching practices.

Was this also true for instructors in all disciplines, from both sexes, and with varying amounts of teaching experience? From the preliminary analyses now completed, the answer appears to be yes. One would certainly expect that instructors in their first or second year of teaching would change, since they are less likely to have established rigid teaching habits; but student feedback did not result in changes even for this less experienced group.

There was yet another possibility: that teachers who received the poorest student reports at midsemester had changed but that these changes were not reflected in the average scores for the entire group. In view of the well-known tendency for students' ratings of teaching to be highly skewed in a positive direction--that is, for students to rate instruction rather leniently--this too appeared to be a viable hypothesis. So we looked at instructors with the poorest ratings at midsemester to see if the end-of-semester ratings for those

---

[3] Centra, J. The student instructional report: Its development and uses. SIR Report Number 1. Princeton, N. J.: Educational Testing Service, 1972.

who had received student feedback had improved more than those of instructors from whom feedback had been withheld. They had not. Ratings did, however, improve for both the feedback and no-feedback groups, but this improvement could be explained by what statisticians refer to as regression effects--that is, the tendency for low or high scores to move in the direction of the average on a retest.

Do these findings mean that students' ratings of instruction are of little value in changing instruction? Perhaps. But there are also several alternative explanations to consider. It may be that a half-semester is too short a period of time for changes to take place, or that if they did take place, the second set of student ratings were not sensitive to them. I'm hoping to follow up a sample of instructors at the end of the current semester to see if teaching changes are reflected after the longer time period.

I should point out here that when studies like this one were conducted with elementary and secondary school teachers, slight changes did occur.[4] Why student ratings produced teaching changes at the lower educational levels but not at the college level is not at all clear. It does not seem likely that college professors would value student opinion any less; on the contrary, one would think that the opinions of college students would have greater impact than the opinions of high school students or sixth graders.

If the findings of this five-college study hold up under continued analyses of the data, then the assumption that college teachers do change after receiving

---

[4] See, for example, B. W. Tuckman and W. F. Oliver, Effectiveness of feedback to teachers as a function of source. Journal of Educational Psychology, 1968, 59(4), 297-301.

Also, N. L. Gage, P. J. Rankel, and B. B. Chatterjee, Changing teaching 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 & Bacon, 1963. Pp. 173-181.

feedback from their students must be seriously questioned. What this implies is that other methods must be relied on to improve instruction or that students' ratings may need to be given some additional "clout." At some colleges that clout is provided by using students' ratings as one of the inputs into salary and promotion decisions; at others the ratings are made public in various ways, such as in student-produced publications (of varied quality, I might add). The effects of using systematic students' ratings for decisions on faculty promotions are, to my knowledge, not yet completely known. While they may result in more justifiable decisions than are now generally being made, the possibility exists that such emphasis will also reduce risk taking and creativity in the classroom. We undoubtedly need more evidence on this question.

Another possible way in which student ratings may have more impact is by providing a better interpretation of the feedback than is typically given to each instructor. This could include written or graphic material, or even personal counseling; in any event, the emphasis would be on helping the instructor better understand his results and what he might consider doing about them.

Teacher Self-Evaluation

A second aspect of this study was to find out how much instructors actually learn from students about their teaching. To what extent, in other words, do instructors describe or rate their teaching differently than their students do? It may be that instructors failed to change after receiving student feedback simply because they were not learning anything they did not already know.

Items from the student form were reworded slightly for instructor responses. Instructors were asked, for example, whether they thought they had

made objectives clear, whether they were encouraging students to think for themselves, and so on. This information was collected at midsemester.

There was a significant difference between instructor and student responses to most of the items, and in each instance instructors tended to describe or rate their teaching in more positive terms. In particular, instructors and their students did not agree on the extent to which course objectives had been made clear and on whether there was agreement between objectives and what was taught. This would suggest that many teachers need to spend more time clarifying their course objectives and directing their teaching toward those objectives. There was also considerable lack of agreement on whether students had been encouraged to think for themselves, on whether the instructor was actively helpful to students, and on whether students were free to ask questions or express their opinions. These items would suggest that many instructors are not interacting with students as successfully as they think they are.

For several of the items, however, the gap between student and faculty responses was not very pronounced. There were more common views, for example, on whether the instructor was well prepared for class, on whether students' interest in subject areas had been stimulated, and on whether students seem to be putting a good deal of effort into the course. But even on these items, about a fourth of the teachers viewed the course or their teaching much more positively than did their students.

Although there was a good deal of similarity between teachers' and students' views, there were also sufficient differences to warrant the collection of students' opinions. It is possible, in fact, that student feedback results in changes only by those teachers who see themselves much differently

from the way the students see them--a possibility that I am pursuing with current analyses.

Accountability in the Classroom

My comments thus far have focused on the college classroom as the major arena for teacher evaluation, but as we all realize, a teacher's effects extend beyond the classroom walls--or at least they should. Student learning is said to be the ultimate criterion of such effects; in fact, some have argued that we should forget about evaluating teaching and concentrate instead on changes in students. Measure the amount that students learn (or "value-added" in economic terms) is what the proponents say.

At first glance the method seems easy enough to apply: simply administer a so-called pretest to students at the beginning of the course and follow it with an end-of-course test usually referred to as the posttest. Both examinations, of course, must suitably measure course objectives. Of interest are the average gains in students' scores, and the good teachers, naturally, are the ones whose students demonstrate the largest gains. Advocates of the method would argue that these teachers ought to be rewarded just as reseachers are rewarded for the number of publications they produce; both measures are, after all, quantifiable.

This method of evaluating teaching, which seems to be another outgrowth of the cult of accountability, involves a number of problems. First, faculty members within a department or course must agree on the objectives to be measured, and then tests must be devised to assess these objectives adequately and in the proper proportions. This is no small problem, but let's assume it can be done. Now, how do we compare score gains between classes? Is a 20-point gain in physics with good students comparable to a 20-point gain with poor

students? In spite of certain standardizing manipulations (i.e., standard score conversions), any conclusions regarding differences between score gains could be totally unwarranted simply because, as numerous experts have pointed out, there is no way of making proper allowances for uncontrolled preexisting differences between students in each class.[5] That is why the random assignment of subjects is so crucial in experiments. In fact, test score gains are used frequently in research on teaching, but that is a different matter from using them as a regular part of the reward structure.

If gain scores are fed into the reward structure, we can also expect that instructors will begin teaching to the test, an abuse of testing that will neither aid the student nor the institution that seeks to reward good teaching. It must be remembered that tests represent only a sample--indeed a very small sample--of the subject matter included in a course or field, and if instructors merely emphasize that limited domain, they may appear as if they've been effective when in fact they have short-changed their students.

But tests properly used can, indeed, be very beneficial; they play an important role both in giving the instructor periodic feedback on his students' progress and in providing a summative evaluation of each student at the end of the course. More instructors, no doubt, could also profit from knowing more about their students at the beginning of a course--information such as their expectations for the course as well as their knowledge of subject matter. So tests per se are not the problem; what I am cautioning against is the systematic use of test scores as the sole criterion for determining which teachers have been most effective.

---

[5] Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 21-38.

Returning to the question raised in the opening paragraph of this paper: How can teaching be evaluated in a way that will lead to instructional improvement or more valid decisions on promotion? To begin with, we will probably never entirely eliminate subjective judgment in evaluating teaching. What we might best do, then, is to utilize as many sources as possible for those judgments and when possible combine them with whatever "objective" information is available. The use of multiple measures of outcomes or performance is a relatively simple but often disregarded notion in evaluation. Administrators (chairmen and deans), students, and faculty colleagues are, of course, the primary groups available to rate teachers, and there is some research evidence on how these three groups compare in their judgments. In two separate studies in which each of the three groups had evaluated the overall teaching effectiveness of specific instructors with whom they were acquainted, there was substantial agreement among the three.[6]

Specifically, the correlations were in the .60 to .70 range, indicating that while their judgments were not in complete agreement, they did share a common basis in making their ratings. What that common basis is we can't be sure, but one possibility is that they are sharing what Polanyi refers to as "tacit knowledge."[7] Tacit knowledge, he says, includes those unmeasurables that underlie an individual's competence and show up in the impressions he

---

[6] Clark, M. J., & Blackburn, R. Assessment of faculty performance: Some correlates between self, colleagues, students, and administrators (submitted for publication).

Also, Maslow, A. H., & Zimmerman, W. College teaching ability, activity and personality. Journal of Educational Psychology, 1956, 47, 185-189.

[7] Polanyi, M. Tacit dimension. New York: Doubleday, 1966.

makes on others. Those so-called unmeasurables are exactly why subjective judgments will continue to be used in teacher evaluation.

But this is not to say that the actual subjective evaluations cannot be made more precise. In addition to using as many sources as possible to collect judgments on teaching, it would seem that something more than an overall rating needs to be employed and that more direct evidence of performance both in and out of the classroom is needed. In particular, if course or instructional improvement is the goal, something more than a single overall rating is necessary. In the first place, a single rating assumes a unitary dimension of teaching ability--and that we know to be unlikely; secondly, an overall judgment does not give a teacher the kind of specific information needed for improvement. My tennis game would stand little chance to improve if I were simply told I was "below average." But when one of my tennis colleagues points out my backhand grip and stance are not right, then I can do something about it--maybe.

We might thus try to encourage more faculty members to sit in on their colleagues' courses, or in various ways, to help one another improve their courses or teaching. For example, a practice that has had some success is to have each teacher orally present to appropriate colleagues a summary of how he plans to deal with a particular unit in his course; in this way instructional objectives and techniques are shared among the group. While the five-college study reported in this paper did not provide any hard evidence that students had much effect on changing instruction, one's colleagues may have more impact. College professors do not hesitate to review critically each other's research and scholarly outputs; is it not possible to bring that same spirit of criticism to each other's teaching?