Teaching & Learning
74-750

REPORT OF SUBCOMMITTEE ON EVALUATION OF TEACHING

(R. Burke)

I. Existing Procedures for Evaluating Teaching at O.U.

A. Summary of Existing Procedures (see Table 1):

1. Almost all departments (except two: History and Area Studies) currently use a departmental questionnaire.
2. Most departments (16 out of 22) require that the questionnaire be distributed in all courses, and use it in reappointment decisions. Most use it also in determining "personal factors."
3. Many departments (10) use some sort of formal colleague evaluation of teaching other than compiling student opinions. The meaning of "formal" here is not clear, however; the number (10) could be 5 or 15.
4. Many departments (11) use formal interviews with a few students, usually some chosen by candidate and some not.
5. Only a few departments (4) use visits to classes by colleagues as a regular part of their evaluation procedure.
6. Two departments (Chemistry and Linguistics) use actual student performance as one measure of teaching effectiveness.
7. Very few departments seem to have formal procedures for:
   a. evaluating teaching of tenured faculty (other than questionnaires);
   b. weighing teaching against research and service in making reappointment recommendations.

B. Plausible Conclusions:

1. Questionnaires are already used by a large enough proportion of the faculty, and are enough alike, to make a common

questionnaire feasible. Whether it is desirable is a different question, dealt with below.

2. Some sort of common form of colleague evaluation may be possible, but any one method of doing this is likely to meet considerable opposition.
3. Judging from the relative lack of formalized procedures other than a questionnaire, from the vagueness of the answers to questions a-e in the Burke memo of Feb. 13, and from comments volunteered by many chairmen, there is a good deal of scepticism about the validity of any procedure for evaluating teaching effectiveness.

II. Progress Report

We are not yet ready to make specific recommendations about the evaluation of teaching at Oakland, but we can indicate some general principles and the recommendations that seem to us at this point to follow from them. We think the subcommittee and the general Teaching and Learning Committee should continue to study the problem next year, in collaboration with:

1. Those members of the Oakland faculty who have special expertise in this area: William Bezdek (Sociology), Ralph Schillace (Psychology), Daniel Braunstein (Econ.-Mgt.), etc.
2. The members of the FRPC.
3. Anyone interested enough to volunteer.

We cannot respond yet to the request of the Arts and Sciences CAP, for reasons which are explained below.

A. General Principles:

1. The most important purpose or function of the evaluation of Teaching is to enable the instructor to improve his teaching effectiveness. If we decide to adopt a new evaluation procedure, therefore, such as one involving a university-wide student questionnaire, it should be to serve this purpose better, not simply to facilitate personnel decisions. It should, therefore:

   a. It should therefore involve all faculty, including those with tenure, not just those being considered for re-appointment.

   b. It should involve a follow-up program of diagnosis and advice, both within departments and in ways to be devised and sponsored by the Teaching and Learning Committee for the whole university: workshops, symposia on special teaching problems, etc. We strongly urge that funds be sought to bring Dr. Frank Koen of the University of Michigan back to the campus next year to help us devise such programs. We must not give an instructor a poor rating, and leave him to figure out what to do about it.

   c. It should be specific enough in feedback information to enable the instructor, with help, to understand what he needs to improve and how he might do it.

2. The evaluation of teaching should also play an important part in reappointment decisions. Oakland University has always

been committed to high-quality undergraduate instruction above all, and our criteria for reappointment should reflect this commitment. A few departments seem to feel, however, that teaching must be de-emphasized in such decisions because no valid method of measurement exists. While the issues are certainly complex, both philosophically and methodologically, our research and discussion so far leads us to reject this position because:

a. While a few studies of the validity of student ratings of faculty (questionnaires) have been inconclusive or even negative, the recent surveys of the literature indicate overall positive results. Here are three surveys the committee found useful:

W. McKeachie, "Student Ratings of Faculty," AAUP Bulletin, Winter, 1969, pp. 439-444.

F. Costin et al, "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," Review of Educational Research, Vol. 41, no. 5 ( ), pp. 511-535.

R. Miller, Evaluating Faculty Performance (Jossey-Bass, 1972). Costin's conclusion is representative: "A review of empirical studies indicates that students' ratings can provide reliable and valid information* on the quality of courses and instruction." See Appendix I for further details.

---

*"Reliable" here means consistency in spite of differences in students' age, sex, grade, GPA, etc. "Valid" means correlation with other measures of teaching effectiveness: before-and-after testing, instructors generally regarded as good teachers by their colleagues, etc.

b. It implies that other bases of evaluation, such as quantity and/or quality of research, are more valid, which is false. There is just as much disagreement about criteria in this area, and there has been far less reflection and empirical study about problems of methodology.

3. Effectiveness of teaching can only be evaluated in relation to the goals of teaching. This seems obvious, but it has important corollaries:

   a. In interpreting questionnaire results (and interviews with students or faculty), it is essential to have a statement by the instructor of his goals for the particular course, his methods, and the factors in the situation influencing his approach to these goals: for example, the size of the class, the level of preparation and motivation of the students, his own strengths and weaknesses, etc. This does not imply, of course, that the instructor's goals must be accepted by his department or by a committee evaluating his performance, but they must be known, or the data about him can be seriously misleading--especially if it can be precisely quantified!

   b. Whatever methods of evaluation we use should be as flexible as possible, to allow for variations in goals and situations from one department to another, and from one instructor to another within each department.

4. A common university-wide procedure, with room for variations of the kind just mentioned, would serve both of the functions listed

In 1 and 2 above better than our present system with different procedures in each department. There really are serious problems of comparison and sampling at present, despite the fact that we all agree that teaching is important, and we seem to mean roughly the same things by it (judging from the similarity of the questionnaires). What we need is a way of giving it the weight that it deserves, however this may vary from one department or individual to another.

B. Tentative Recommendations:

1. We are moving toward a recommendation that all departments in the university adopt a common core of student questionnaire items, while retaining variation in other items from one department to another and even (if possible) from one instructor to another. There are two good ways of doing this:

   a. Adoption of a well-standardized external system such as the Purdue "cafeteria," using 15-30 items (5-10 in each of the three categories just mentioned) from their catalog of 200 items. This system is currently used by at least 21 schools (6 around our size) and supplies standardized norms for all items updated every year. It is computerized, and they have a service for test construction, print-out, and scoring. There are 3 "blank" items, allowing departments or individuals to construct questions not included among the 200. A study of our current dept. questionnaires suggests that all of them could be translated into selections from Purdue's

200 items, making use of the 3 "blanks." There are also 5 "demographic" items, according to which comparative data can be broken down: school, year, sex, required course or not, grade expected in course. Thus, if we ascertained that ratings of instructors in required courses average .85 of the ratings of all instructors on a particular item, we could program the computer to correct for that deviation. If we suspect that another factor might influence the ratings, say the student's GPA, we could ask that and find out, and correct for it. At the same time, the instructor would find out whether high-GPA students rate him higher than lower-GPA students, and how both compare to those at comparable schools. This is an attractive system, but it might encounter opposition from those departments completely satisfied with their present instrument (such as Education and Economics-Management). Also, it would be expensive (around $2,000 each semester). And the comparative data from other schools may not actually be that useful to us, since teacher ratings (like student grades) probably form a similar bell-curve at each institution.

b. The other alternative is to build our own core of 5-10 items, as recommended by Frank Koen, on the basis of our own institutional goals and values. This might not be too difficult, since the current department questionnaires are much alike, and similar to the models suggested in

the literature. It would allow departments to keep their present questionnaires, minus any items that duplicated those in the common core. It would take some hard thinking to decide whether the questions we are all already asking are all the questions we should be asking. But even with the Purdue system, we would need to agree on a common core, so in this respect there is little difference. The Sociology-Anthropology Department has recently adopted the Purdue system, after considerable study of others like it (for further information, consult William Bezdek). We suggest that we wait and see whether they are satisfied with it next year, while working on our own "core items."

2. We are ready to recommend that the administration of questionnaires by departments should be standardized, to make the results as comparable as possible:
   a. They should be used by all instructors in all classes: not only to permit development of reliable Oakland norms, but to improve the teaching of everyone on the faculty.
   b. They should be filled out at the same time in the semester in all classes: perhaps halfway through, to allow time for suggestions to be used in that course, and to get the opinions of students who may drop out later; or perhaps at the end, to allow evaluation of the whole course, and to eliminate the fear of retribution.
   c. They should be filled out in class, collected by a student monitor and brought to a central place. One study shows

that the instructor's presence alone makes ratings significantly higher; therefore perhaps he should not be in the room. The percentage of the official class enrollment filling out the questionnaire should be noted: anything below 70% may be statistically suspect.

3. In order to ensure that questionnaire results will be interpreted in the light of the instructor's goals and methods and the relevant factors of the situation (3.a. above), we further recommend that each instructor should draw up a statement of such goals, methods, and relevant factors for each course, preferably using a common format. This statement should play an important part in departmental and committee evaluation of teaching. The Teaching and Learning Committee could design a common format for such statements, if requested.
4. Some method of colleague evaluation would be desirable also, to act as a check on the questionnaires and to make judgments that students are not qualified to make, such as how well an instructor knows his material, his choice of methods, textbooks, etc. There are difficultues however, with each method of doing this:
   a. Visits to classes. Even if more than one visitor makes more than one visit per course, there are still serious methodological questions:
      (1) are the visitors biased? (One "friendly" and one "hostile" does not add up to objectivity).
      (2) Is their sample large enough? (No.)
      (3) has their presence introduced a disturbing factor into the situation?

Also, these visits take time (if combined with conference before and after, as they should be). And some faculty feel they are a violation of a traditional right of privacy in the classroom. Finally, they could create tension and hostility between colleagues, where cooperation and friendship should be. For all these reasons, we hesitate to recommend this method for use in reappointment decisions. But we encourage mutual visits as a general practice; and if the atmosphere permits and the visitors keep in mind what can and cannot be learned from 1 or 2 visits, they can be a useful adjunct to other methods. A standard form could be devised by the Teaching and Learning Committee.

b. Examination of syllabi, exams, and other publicly distributed materials. This may be very informative in some cases, but simply inappropriate (and thus unfair) in others. Perhaps each instructor could indicate in his "Goals and Methods" statement whether evaluation of such materials would be fair in his case. But then he might say No whenever he thought they would be given low marks! In any case, if this method is used, we recommend that a standard form be used, which could be designed by the Teaching and Learning Committee.

c. Informal conversations with students and colleagues, impressions from departmental colloquia, etc. Such "methods" are hopelessly unreliable, and should not be admissible as evidence in reappointment decisions.

We feel fairly strongly about this. It may seem a shame to be unable to include the chance favorable comment or impressions, but the harm by the chance unfavorable one is too great, and too irresponsible, to be toleratre. Any avoidable element of "subjectivity" in these judgments is an element of injustice.

5. We are not recommending interviews with students as a method of evaluation of teaching, in spite of the fact that about half of our departments are currently doing it. This method has some of the same problems as visits to classes. The sample is usually much too small to be reliable; the students are not anonymous; and choosing some "friendly" and some "random" does not create objectivity. If the sample were large enough (say 20%) it would take a lot of time. There is a real danger that the vivacity of face-to-face interviews with students would outweigh the cold, dry figures from the questionnaires in the minds of evaluators; whereas in case of conflict, the latter are far more reliable and thus more fair. If this method is used, a standard interview format should be developed to permit comparison, and the size of the sample should be noted.

   As a supplement to questionnaires, interviews with students may be useful in raising questions about the validity of the questionnaire results, and leading faculty committees to seek more data. The burden of proof, however, should always be on any other method if it conflicts with the questionnaire results.

6. Since some people think that the only "true" evaluation of teaching effectiveness would be before-and-after testing, perhaps we should point out that:
   a. Such tests cannot be completely objective either. If they are made up by the instructor himself, this is qbvious; if they are made up by an outside agency (the American Chemical Society, say), their goals do not necessarily coincide with the instructor's. This may not be a serious problem in Chemistry, but it is in Philosophy.
   b. To the extent that a teacher's goals include influencing the attitudes and values of his students, there are no "tests" that are any more reliable than student questionnaires.
7. Finally, we call your attention to the fact that several aspects of teaching are not included in our present system, but we do not yet see any good way to include them:
   a. guidance of independent study projects;
   b. academic advising;
   c. helping colleagues with their teaching;
   d. public presentations--the university community at large, to departmental colloquia, in other people's courses, etc.;
   e. curriculum development (perhaps under "Service"?).

   Candidates could be encouraged to include such things in their "Goals and Methods" statements, but it is hard to see how they could be given much weight from this alone.

TABLE I: EXISTING PROCEDURES BY DEPARTMENT

	Use Dept. Questionnaires	Require Questionnaires	Formal Interviews With Students	Class-room Visits	Formal Colleague Evaluation
Area Studies	--	--	X	--	X
Art	X	X	X	--	--
Biology	X	--	X	--	--
Chemistry	X	--	X	--	--
Classics	X	X	X	--	--
Econ-Mgt.	X	X	--	--	--
English	X	X	--	X	X
History	--	--	X	--	X
Linguistics	X	X	--	--	X
Math	X	--	--	--	--
Mod. Lang.	X	X	X	--	X
Music	X	--	--	--	--
Philosophy	X	X	--	X	X
Physics	X	X	--	--	--
Political Science	X	X	--	--	--
Psychology	X	X	X	X	--
Soc.-Anthro.	X	X	X	--	X
Speech	X	X	--	--	--
Learning Skills	X	X	X	X	X
Engineering	X	X	--	--	--
Education	X	X	--	--	X
NCC	X	X	X	--	X

APPENDIX I

A few empirical findings which may surprise some people (see McKeachie, Costin, and Miller, opera cit.):

1. Students rate instructors about the same 10 years after graduation as they rated the same instructors when they were in college.
2. Student ratings of instructors generally correlate with how much they have learned, not with whether they liked the instructor's personality.
3. There are no significant differences in ratings by--
   a. male and female students;
   b. older and younger students (although graduate students rate instructors higher);
   c. students with high and low grades in the course (although students with generally high or low grades--GPA--may rate a given instructor quite differently).
4. There are no significant differences in ratings of--
   a. male and female instructors;
   b. older and younger instructors (but older get slightly lower ratings);
   c. instructors who grade hard or easy (except on the item "fairness in grading");
   d. instructors who do more or less research.
5. Most studies show instructors of large courses getting lower ratings than of small ones, and those of required courses lower than electives. Also, majors gave instructors higher ratings

In some studies. But some studies contradict these results, and the differences are not as great as one might expect. For a 220-item annotated bibliography of research in this area since January, 1968, see Virginia A. de Wolf, Student Ratings of Instruction in Postsecondary Institutions (Washington University, 1974).

/mem
7-16-75

May 7, 1975

ME ORANDUM

TO: Members of the Teaching & Learning Committee

FROM: Ralph Schillace RS

The evaluation sub-committee's investigation and report of the methods used at Oakland University to assess teaching is the most thorough and thoughtful treatment of the topic ever conducted here. I agree, generally, with the findings and with the spirit of the recommendations. My own viewpoint on the matter includes the following recommendations:

(1) Student questionnaire assessment of teaching should be applied in a systematic and cooperative fashion across departments to correct current problems in sampling and to facilitate comparison of individuals. But I do not believe that student questionnaire data should ever be allowed to stand alone as the essential measurement of teaching performance or effectiveness.

(2) The procedure of colleague visits followed by a written report of teaching performance by the observers should be part of the official documentation for a given instructor. This procedure offers the most promise for interaction among peers and the highest probability of changing teaching behavior for both the observer and the observed. The details of this procedure can be developed so it is practical.

(3) No official report of teaching performance and effectiveness should be allowed to stand and be reviewed without a written statement by the instructor being evaluated, in which he responds to the contents of 1 and 2 above and generally recognizes that he has the right to qualify that content.

I cannot endorse a procedure to evaluate teaching that does not include all three of these components. The efficiency of student questionnaires and the mystique that surrounds quantification of responses are serious distractors to the short-comings of student paper-and-pencil assessment. A procedure that uses peer or colleague evaluations and self-reports in conjunction with student questionnaires promises to offer a system with some reasonable checks and balances. Finally, such a system has the best chance of changing behavior, the avowed, major purpose for the Teaching & Learning Committee's assessment project.

RS/mem