DOCUMENT RESUME

ED 119 578 HE 007 318

AUTHOR French-Lazovik, Grace
TITLE Evaluation of College Teaching. Guidelines for Summative and Formative Procedures.
INSTITUTION Association of American Colleges, Washington, D.C.
PUB DATE [75]
NOTE 12p.
AVAILABLE FROM Publications Office, Association of American Colleges, 1818 R Street, N.W., Washington, D.C. 20009 ($0.50)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *College Teachers; *Evaluation Methods; *Guidelines; *Higher Education; Peer Groups; Self Evaluation; Students; *Teacher Evaluation

ABSTRACT

The literature on teaching evaluation has long recognized that it is simply not possible now, or perhaps ever, to isolate from among all the variables that are interacting the individual teacher's contribution to changes in the learner, many of which are complex, subtle, and may not be observable until much later in the student's life. Thus, other criteria, usually judgmental in nature, have formed the basis of efforts to evaluate teaching. A widely stressed admonition is that one should never rely solely on a single source of data, but should use several or all of these forms of judgment. The sources of first-hand data that have been most often suggested (and which are discussed in this document) are faculty self-evaluation, peer evaluations without visitation, and the student's evaluation of teaching. (Author/KE)

Association of American Colleges

an occasional paper . . .





# EVALUATION OF COLLEGE TEACHING
## Guidelines for Summative and Formative Procedures

Grace French-Lazovik, Ph.D.[1]

*This monograph was prepared under the auspices of the Association's Commission on Liberal Learning. It is suggested that it be shared with all those in the academic community who have an interest in the subject. Additional copies may be ordered from the Publications Office, AAC: single copy, 50c; 2-25 copies, 35c per copy; over 25 copies, 25c per copy. All orders must be prepaid.*

The decade of the '70s has brought to the academic community a renewed interest in improving teaching; and the evaluation of teaching, as one means to this goal, is receiving a tremendous surge of attention. Unfortunately, this attention sometimes degenerates into acrimonious argument among faculty members about what procedures should be used or, on the other hand, results in an uncritical and uninformed plunge into collecting all sorts of opinions—nowadays, mostly student opinions.

In the current rush to bring some degree of objectivity to the evaluation of teaching, excellent advice is being ignored.[2] For those who would heed it, the place to start is with *The Recognition and Evaluation of Teaching*, by Kenneth Eble,[3] director of the Project to Improve Teaching, jointly sponsored by the Association of American Colleges and American Association of University Professors. But additional required reading includes McKeachie,[4] the *AAUP Statement on the Evaluation of Teaching*,[5] and Miller.[6] What is now needed, I believe, is a careful examination of some of the experience and findings available from serious efforts to evaluate teaching. Which procedures establish the grounds for meaningful and reliable decisions about teaching quality? What kinds of data contribute to sound assessment of the teaching of individual professors when promotion or tenure is at issue? At least as important are guidelines as to how evaluation can contribute to the goal of *improving* college teaching.

At the outset, we should sharply differentiate these two quite different purposes that evaluation serves. Scriven[7] has called evaluation for improvement "formative evaluation." Necessary to individual instructors on a continuous basis, its requirements differ markedly from those of "summative evaluation" (Scriven), which becomes relevant only periodically, perhaps every two or three years, as a significant contribution to academic deliberations about tenure and promotion. The crucial demand on summative evaluations is that they provide the basis for fair decisions. For unless teaching quality is rewarded in a way which is perceived by faculty as fair, there will be little motivation for formative evaluation or the improvement that it facilitates.

What is the essential first step in establishing sound procedures for the summative evaluation of teaching? Many writers have emphasized that it should be a clear statement of the criteria by which teaching is judged, along with the specification of the weight to be accorded in academic decisions to the teaching performance of each faculty member. (Weights for the other traditional academic functions—scholarly contribution to one's discipline, academic or community service, or any nontraditional responsibilities should, of course, also be clearly specified.) The process by which these individualized statements can be effectively prepared has been so well laid out by others[8] that there is no need to repeat here the principles which govern it, the influences to which it should be sensitive, or the

*Grace French-Lazovik is an Associate Professor of Psychology and Director of the Center of Evaluation of Teaching at the University of Pittsburgh.*



1818 R STREET, N.W • WASHINGTON, D.C. 20009

crucial role of department chairpersons and deans in implementing it.

The literature on teaching evaluation has long recognized that it is simply not possible now, or perhaps ever, to isolate from among all the variables which are interacting the individual teacher's contribution to changes in the learner, many of which are complex, subtle, and may not be observable until much later in the student's life. Thus, other criteria, usually judgmental in nature, have formed the basis of efforts to evaluate teaching. The desirability of using first-hand data in tenure and promotion decisions has been frequently acknowledged, and the sources of first-hand data that have been most often suggested are faculty self-evaluations, peer evaluations based on classroom visitation, peer evaluations without visitation, and students' evaluations of teaching. A widely stressed admonition is that one should never rely solely on a single source of data, but should use several or all of these forms of judgment. Let us examine separately each of the recommended sources of first-hand data.

### Classroom Visitations

Classroom visitation by colleagues has been tried in a number of different forms. The general finding is that it does *not* provide a sound method of evaluating the teacher's in-class activities. A few classroom visits by one colleague cannot be expected to produce a reliable judgment. (The terms reliable and reliability are used simply to mean consistency among judgments, including their repeatability.) Even when the number of colleagues is increased to three, and each makes at least two visits, the reliability of resulting evaluations[9] is so low as to make them useless.

Whether these ratings would ever attain the reliability of the pooled judgments of students in a class, who observe and experience teaching for an entire term, is a question which has not been studied. Knowledge of the conditions which produce sound ratings would lead us to believe that they would not. A major problem is that the anonymity of the raters cannot be preserved. Even if three or four colleagues visit (not many faculty take kindly to the idea of having a team of colleagues present at each class meeting), and the ratings were low, the givers of the low ratings would then be known to the teacher, who would typically have to interact with these evaluators on a daily basis. It is little wonder, then, that where colleague visitation has been tried, all ratings tend to be very high. In a study[10] where 54 teachers were evaluated on the basis of classroom visitation (two visits by each of three colleagues) 94% of all ratings were in the top two categories of a five point scale. Arthur Eastman[11] confirms this effect in his delightful article, "How Visitation Came to Carnegie-Mellon University:" "Visitors were generous ... most (teachers) were encouraged at the approval they received ...." Such a positive bias prevents the attainment of reliability, which depends in part on discerned differences in performance among individuals as against the perceived sameness of everyone.

Scott Edwards[12] has suggested possible reasons for this high positive bias. "What department member conducting a class visit, knowing that he who evaluates today is himself evaluated tomorrow, can fail to see the need of a discreet reciprocity? Even where such concerns are not present, as when an evaluator enjoys a sufficient protection of tenure and rank, the too close acquaintance of department members does not permit the placing of much confidence in their assessments of each other."

Thus absence of adequate reliability renders this source of first-hand data useless for purposes of *summative* evaluation.

On the other hand, colleague classroom visitation can be a valuable source of suggestions for the improvement of teaching. A system of visitation, free from the responsibility to record a formal evaluation and engaged in by an entire department, can stimulate discussion and concern among faculty about their teaching, and may prove a powerful motivator for teachers to be better prepared for each class. While most of us may not mind being thought of by our peers as not very polished in our delivery or skilled in leading discussions, we certainly do not want to be regarded as having given only superficial thought to the organization of our subject matter or to the current developments in our fields. A general sharing of observations and discussion of problems, perhaps at weekly brown-bag lunches, could bring a healthy openness to the traditional "conspiracy of silence"[13] about problems encountered in teaching.

### Self-Evaluation

With respect to self-evaluation, the evidence again does not support the use of this source of first-hand data as a basis for decisions about teaching quality. Blackburn and Clark[14] collected separate evaluations of teaching effectiveness for 45 full-time faculty members from four different sources—students, administrators, faculty colleagues, and from the professors themselves. This study found that self-ratings showed near zero correlations with ratings made by each of the other sources of judgment. The investigators conclude that, "The professor lives with an erroneous perception of how others perceive and assess him."

Centra[15] compared teacher self-evaluations with those made by students. "The results demonstrate a clear discrepancy between the way most teachers describe their instruction and the way students describe it. Not surprisingly, most teachers . . . viewed themselves in more favorable terms, particularly on such matters as whether they stimulated student interest, the extent to which the course objectives were met, and whether the instructor seemed open to other viewpoints. Of course there were some teachers who viewed themselves very much as their students viewed them, and even a few had more negative perceptions. Nevertheless the majority saw themselves in rather glowing terms . . . ." Thus, self analysis cannot provide the kind of data needed for summative evaluation.

As with classroom visitation, self-analysis, along with other sources of feedback, can contribute positively to *formative* evaluations. By comparison of perceptions from other sources with their own self-descriptions, faculty members can be alerted to examine further whatever discrepancies occur. McKeachie[16] has pointed out that feedback that differs from our own perceptions or which adds new information is much more likely to be followed by change in behavior than is feedback that simply confirms what we already know.

**Peer Evaluations (without class visitation)**

There remain two other sources of first-hand data, peer evaluations without visitation and student evaluations of teaching. What has not been sufficiently clarified in most writing on the evaluation of teaching is that both peer and student judgments are essential to summative evaluation; one without the other will lead to unfair decisions. Even when both of these forms of evaluation are carried out, unfair decisions can still result unless very careful thought is given to the role of each.

It does very little good to obtain from faculty peers a single global judgment of a colleague's teaching effectiveness. Some years ago, Edwin Guthrie and I[17] examined, through factor analysis, the relation of colleague judgments of teaching to student ratings of teaching quality. Data were available for 121 faculty members, each of whom had been evaluated by his or her students and also by committees of six or seven faculty colleagues.[18] The six items judged by students and eight items judged by faculty were the following:

| | |
|---|---|
| Evaluations by Students | 1. Teaching effectiveness<br>2. Clear and understandable in explanations<br>3. Active personal interest in the progress of the class<br>4. Friendly and sympathetic manner<br>5. Shows interest and enthusiasm in subject<br>6. Gets students interested in subject |
| Evaluations by committees of faculty colleagues | 7. Teaching effectiveness<br>8. Contribution to field through research and publication<br>9. Contribution to community or state<br>10. Ability to cooperate with other members of department<br>11. Knowledge of subject<br>12. General knowledge and range of interest<br>13. Rate of professional growth<br>14. Recognition by others in his or her profession |

The analysis revealed that these items clustered into three independent factors. The first involved all six items judged by students; the second consisted of faculty judgments on four items:

8. Contribution through research and publication
11. Knowledge of the field
13. Rate of professional growth
14. Recognition by others in his or her profession

The third was measured by items:

7. Teaching effectiveness (judged by colleagues)
9. Contribution to community or state
10. Ability to cooperate with other members of department
12. General knowledge and range of interest

Guthrie called the first factor the teacher's impact on students; the second, impact on one's profession; and the third he called impact on colleagues. The chief point of interest here is that teaching effectiveness as judged by colleagues measures something quite different from the group of items which involve impact on students. Colleagues probably judge one another as teachers on the basis of things they can observe ("Readiness to work with others in the department in arranging schedules, examinations, and the mass of operational detail on which members of a department must agree"[19] as well as on breadth of general interests), *not* on the basis of classroom activities. Shoben,[20] another psychologist, has suggested an equally plausible interpretation of this third factor—that it represents a general likableness and "reputation as a good colleague. It suggests that none of us is likely to designate a likable guy as a poor instructor unless contrary evidence arises to strike us across

the chops." If this is true, then how can peer judgments be used in the evaluation of teaching?

Peer evaluations can be used only if we break down the global judgment of teaching quality into those characteristics which faculty do observe, and if we use certain psychometric controls in obtaining them. Behind the need for both student and peer judgments lies two well established principles of psychological measurement. The first is that judgments of complex human performance cannot be valid unless they are based on adequate observations of the performance or characteristics to be rated. The second is that the rater must have appropriate background against which to compare and evaluate what is observed. Students directly observe what goes on in the classroom and can make judgments about certain aspects of teaching, particularly those relating to their own experience of it. They do not characteristically, however, have the background to judge other essential characteristics. On the other hand, faculty peers typically do not observe the in-class teaching of their colleagues, nor are they capable of experiencing it as do those without their knowledge of the field; but they do observe and have the background to judge characteristics of teachers which students cannot.

One absolute essential of good teaching is the instructor's knowledge of the subject being taught. Students, especially freshman and sophomores, are not in a position to make this judgment and should not be asked to do so. (They may judge whether or not the teacher could answer their questions or whether he or she presented material beyond the textbook. They may also judge whether they felt they had learned something new. But this must be distinguished from whether what they learned was superficial and out of date or represented an in-depth knowledge of the subject.) It is exactly this point that is demonstrated by the widely quoted "Dr. Fox Experiment."[21] In that study, a trained actor delivered a lecture on mathematical game theory to a group of medical educators. He presented incorrect information, cited non-existent references, and used neologisms as basic terms. When his audience rated the lecture, the great majority gave favorable responses to questionnaire items regarding its quality. We can be sure the ratings would have been quite different had the lecture been delivered to professors of mathematics. The essential point is that judgments about the accuracy, currentness, or sophistication of a teacher's knowledge can *only* be made by faculty peers conversant with the same field.

All of us have heard statements which express sentiments such as, "No one who isn't publishing can be a good teacher," or "Those who take all of the time necessary to prepare publishable materials do so at the expense of their students' welfare." The awareness of the importance of a teacher's knowledge to sound teaching has led to some very muddled thinking on this point. It has, I believe, been one of the reasons for the use of judgments of research quality not only as a criterion for the scholarly achievement of faculty but as a criterion for their teaching effectiveness as well. Good teaching requires scholarship—the kind that keeps the instructor in immediate and thoughtful contact with developments in his or her field and with the ideas and findings of other scholars. This may not necessarily be the kind of scholarship which results in publication. But many faculty members do not trust their judgments of a colleague's knowledge unless they can see something he or she has written, or unless they know that editors of respected journals have accepted and published his or her work. Peer judgments of a colleague's publications are a perfectly legitimate criterion in the evaluation of the "professor as scholar," but their substitution as a criterion for the evaluation of the "professor as teacher" simply misses the mark. It is time for the academic community to acknowledge that there are other ways of demonstrating currency and depth in one's field than by publishing, and time also for faculty to have the courage to trust their judgments about the substantive knowledge of colleagues with whom they interact on a daily basis. The active, on-going life of an intellectual community is filled with discussions of recent developments in a field, consultations with others on problems and ideas, colloquia, meetings, attendance of lectures, etc.; one cannot help developing an informed opinion of a colleague's knowledge.

In addition to evaluating a teacher's knowledge, peer judgments are needed for the evaluation of at least three other aspects of teaching. If an appropriately selected group of colleagues[22] reviews such data as a teacher's course outlines, texts, syllabi, reading lists, and statements of objectives, then they can render a useful judgment of the quality of teaching materials. A judgment of this sort does not need to produce fine discriminations, but it can answer relevant questions like, "Are the materials current?" and "Do they reflect the best work in the field?" and "Are they appropriate to the course goals?"

Some record of the performance of students should also be examined by a peer committee. What kinds of tests were used, and how did the students perform on them? Were they all true-false items, or were they more demanding of higher intellectual functions? Were papers written or projects carried out? What was their quality? What did the students learn?

This last question is important for any course, but it has particular significance for many elementary courses in which the content is prescribed as the foundations on

which more advanced courses must build. I will never forget my disbelief at hearing a young instructor in a beginning psychology course say, "My class wasn't interested in the neural basis of behavior or the principles of sensation and perception, so we skipped those topics and discussed something they were interested in, the origins of sex-role identification." The origins of sex-role identification define a perfectly legitimate topic in psychology, usually covered in courses on developmental psychology or personality theory. One can even be sympathetic to the young instructor's desire to encourage his students' interest in a psychological topic. But instead of arranging extra class sessions or informal meetings to pursue their interests, he chose to skip fundamental course content. It is the department's responsibility—not the students'—to see that a teacher does not sacrifice "hard topics" for more naturally appealing ones, and this can be ascertained if peers ask, "What did the students learn?"

Finally, there are aspects of teaching which do not bear directly on a faculty member's classroom activities, but which should be evaluated and rewarded. These relate to the assumption of departmental responsibilities such as service on curriculum committees, supervision of graduate students who are learning to teach, the proposal of new courses, and even service on peer evaluation committees.

If peer evaluations are obtained on additional teacher characteristics, two guiding principles are critical: (a) the judges must be able to observe, outside the classroom, what they are evaluating, and (b) they must have the background against which to compare what they observe. These are such fundamental requirements of valid judgment that their emphasis constitutes an embarrassment. Nevertheless, they seem to be the ones most frequently and persistently violated.

To achieve the validity of which peer judgments are capable, careful and systematic procedures, related to number and choice of judges, instructions, and method of obtaining judgments, are essential. While care must go into their planning, the procedures themselves are relatively simple and less cumbersome than might be imagined.

As mentioned earlier, summative evaluations of teaching should be made only periodically—perhaps every two years for young untenured instructors, every three years for senior faculty. One reason for this timing is to allow enough teaching to occur to provide a representative segment of a teacher's work. Evaluations should not rest on one course, or even on several courses taught in one term. There should be enough data to ascertain trends when a course is taught on several occasions or to see whether improvement is taking place over time. As much as possible, the evaluations should be staggered, so that about a third of the members of any given department are being evaluated each year. This arrangement makes feasible the use of rating controls necessary for reliability but that becomes burdensome if demanded for all members of a department at the same time. In general, the process should guarantee the anonymity and independence of the rater.

A look at the peer rating procedures used by Edwin Guthrie illustrates many of the principles which should be followed. Whenever a faculty member was a candidate for promotion or tenure, Dean Guthrie requested him or her to nominate five colleagues who could serve as evaluators. They could be from the faculty member's own department or from a related department, but the essential requirement was that each evaluator be conversant with the field of the person to be judged. From the five, the Dean chose three and added three more of his own choice. Within this structure, he tried to insure that no rater was in competition for rank or salary with the person evaluated (thus, only tenured faculty served on committees) and that there were at least two members on the committee from outside the department of the candidate, but in a related field. The six[23] constituted a secret committee that never met. No member knew who the other members were nor did he know whether he had been nominated by the ratee or chosen by the Dean to serve on the committee. *He was asked not to reveal his appointment to anyone,* and instructed that this was a matter of academic integrity. Each rater was supplied with a set of materials that the ratee had provided to the Dean. The task of each rater was to arrive at a totally independent judgment on the specified characteristics and to write a general statement about the candidate. Each member returned his or her signed ratings directly to the Dean, and the six judgments were pooled for each characteristic rated.

The principles underlying this set of procedures include these:

1. The person being evaluated had some choice, with the Dean, of his evaluators.
2. Because more faculty were nominated than were chosen, the candidate could not be sure which of his or her nominees had been appointed and thus could not identify any individual as definitely on the committee. This provided a measure of protection for the anonymity of the raters.
3. The secret committee prevented one rater from trying to influence the others. No one could act as an advocate or an adversary.
4. Each rater was forced to rely on his own judgments—not those of others.

5. The knowledge that the Dean, and only the Dean, saw the signed evaluations promoted a good deal of care on the part of the evaluators.
6. The pooling of a set of independent judgments gave maximum reliability—better than a jointly agreed upon judgment.
7. The extra-departmental members acted as a corrective for occasional intradepartmental biases.

Obviously, the Guthrian model is not the only way in which reliable peer judgments can be collected, but the principles illustrated enjoy considerable importance and are often overlooked. Another model in use requires that both the dean's and the candidate's choices come from an elected committee of the college faculty. An important consideration here is that such a committee be large enough to afford choice, and especially large enough that each candidate can nominate more committee members than will be selected; otherwise anonymity cannot be preserved. Other illustrations that incorporate the essential safeguards could be described, but institutions vary so widely in size and organization that no set of models will serve all. As Richard I. Miller[24] often emphasizes, only if a college "adapts; not adopts" will a particular system work within its structure.

The practice of having evaluations arrived at in meetings of peer judges should be discouraged for two reasons: 1) it destroys the independence of judgments and 2) it fails to protect the evaluation process from the subtle and complex interplay of social and psychological variables present in face-to-face meetings. Such a procedure is often followed under the belief that a gathering of the group facilitates information exchange, but this function can be accomplished in other ways.[25] A covert advocacy or oppositional stance on the part of a peer can often be couched in what appears to be an unbiased and reasoned argument. Even seemingly objective committee discussions are not free of personality interactions based on friendship, charisma, or respect for another's status; nor do they prevent the interplay of factors such as a desire to please, a history of exchanged favors, or an unwillingness to speak up in the presence of stronger individuals who thereby "wield disproportionate influence."[26] This is not to say, of course, that many faculty cannot maintain an unbiased position in the presence of these factors; but generally, open meetings do not provide the conditions that maximize objectivity of judgment on the part of all evaluators. When peer evaluators do not know who the other members of the evaluation committee are, the effects of such variables either cannot operate or are held to a minimum.

How many raters are needed? We know that a single rating is not, in general, reliable. Because pooling the ratings from as many as three judges substantially improves reliability, peer committees should probably have at least three members, more if possible. Another important consideration is that the system must have enough flexibility to be used in all departments. If a department is so small that it has only two tenured faculty, then only one can be appointed and the other two judges must come from allied fields. The principle of anonymity of the rater may not be protected perfectly in such cases, but it should be guarded as carefully as is possible. There is abundant evidence[27] that ratings made without the protection of anonymity have neither the validity nor the reliability of ratings made with the guarantee that the rater will remain anonymous to the person being judged. Peer ratings based on the principles outlined here can provide one source of usefully cogent data to be examined along with student ratings and with recommendations from departmental chairpersons.

Even here, where the purpose is so clearly summative in nature, preparation by the ratee of materials for a faculty-peer committee may also serve formative evaluation. Thus, the self-presentation process may contribute to self-evaluation,[28] especially if, in addition to assembling samples of syllabi, tests, graded papers, etc., the faculty member prepared an analytical paper on the development of each course taught, on his or her own development as a teacher, and on the changes made over the years in a particular course. Such self-analysis could become the starting point for efforts to improve.

### Student Evaluations of Teaching

The characteristics of good teaching that colleagues can judge are essential ones but not sufficient, for they tell us nothing of what transpires within the classroom. Much published work has established the reliability and some types of validity of student evaluations of teaching.[29] There is no doubt that if the best known procedures are used, student judgments can provide an excellent source of first-hand data. How much faith can be placed in these judgments will depend on the quality of the instrument and of the procedures employed to collect them. As with anything else, that quality can range from sound and sophisticated to sloppy and inaccurate. Careful planning and discussion, the commitment of resources, and some expert advice must precede their use.

Before any type of student opinion is obtained, two basic issues should be understood. The questions raised by the differing requirements of formative and summa-

tive evaluations contrast sharply with those posed by the second issue. The first focuses on consequences of distinctive evaluation purposes; the second concerns the role that an institution wishes its students to assume in the evaluative process.

If the purpose of obtaining student judgments of teaching is wholly that of providing feedback as a basis for the individual professor to improve, and if the results are not to be used in any administrative decision, then the answers to a whole set of questions regarding procedures are automatically determined. For example, the questionnaire items to which the students are asked to respond can be framed by the individual professor or by a group of faculty. The items may, but need not, go through elaborate processes of refinement. The questionnaire can contain many items that are as detailed and specific as possible to the course taught, so as to give clues for improvement. Items dealing with the teacher's style, the text, the exams, and all aspects of the course are appropriate. The guiding principle in item selection is simply that the information might help a professor improve.

Additionally, administration of a student questionnaire can be left in the hands of the individual teacher and there is no necessity that the ratings be numerical. If the resources for obtaining quantified ratings are available the professor may obtain fairly precise information, but if these resources are not available, then qualitative evaluation can be used. And finally, only the faculty members involved should see the results. It is especially important that results of formative evaluation *NOT* be given to administrators. Lacking the requirements of summative evaluations, student judgments obtained solely for the teacher's improvement can lead to inaccurate comparative assessments of teaching quality.

When student judgments are to be considered in summative evaluations, a wholly different set of procedures is dictated in order to insure the comparability, accuracy, and consistency of the results necessary to their use in the academic decision process. A standard questionnaire, one which has been carefully derived and subjected to considerable refinement, is necessary to provide comparability among professors. Only a very small number of items, covering the qualities common to all good teaching[30] should be used. Items dealing with teaching style should not be included, for in the hands of an administrator they provide a temptation to consider one teaching style better than another. There is abundant evidence[31] that no one style, *per se,* produces superior learning, but the style with which an individual teacher can be most successful depends on a host of variables, including his or her own personality, the subject matter taught, the students' backgrounds, the goals of the course, and many others.

Quantified responses are necessary for comparability, as are norms against which the numerical ratings can be compared. Each college should determine the type of norms needed. This decision rests on discovering what variables, such as faculty rank, class size, and course level result in overall differences in student ratings. The widely quoted studies[32] that imply that the same variables are operating on all campuses misrepresent the evidence. Which variables contribute to differences depend on the evaluative instrument used and the nature of the students and faculty at a particular school.

It goes without saying that for summative evaluation, the anonymity of the student raters must be guaranteed. Just as important, standard procedures of administering the student questionnaire are required. Quantitative ratings can be influenced by the instructions given regarding the ratings. Evidence here is provided by an investigation[33] designed to determine "whether the individual administering an evaluation instrument has any significant effect on the results . . .This study, involving ten sections and 227 students in an introductory educational psychology course, found a significant difference (at .05 level) between whether the instructor or a neutral individual administered the student evaluation form. Higher ratings were achieved when the instructor administered the survey."[34] Thus control of the presentation of the rating instrument and the instructions regarding its completion are essential and cannot be left in the hands of the individual instructor.

Finally, there must be an equitable process, mutually agreed upon by faculty and administrators, by which the ratings are communicated to the department chairperson or dean. The policy used should take into account the realities of any educational institution. It is well known that on occasion there are variables beyond the control of the individual professor which adversly affect the quality of his or her teaching. An especially heavy work load may be assigned in a particular term making necessary class preparation impossible; the size of a particular class may not be appropriate to the skills of the teacher; a personal tragedy in the professor's life may occur during a particular term; or someone may have to fill in for a colleague on leave by teaching courses outside his or her special area of knowledge. Requiring that a sufficient sample of evaluations from all courses taught be submitted but permitting each instructor some choice as to which ones are presented usually prevents unfairness in these matters.

Theoretically, it is possible simultaneously to collect student judgments for formative and summative evalua-

tions, perhaps by using a two-part instrument. If such a dual attempt is made, however, then all of the procedural safeguards for summative evaluations must be observed.

Addressing the second major issue—namely the role of students in the summative process—helps clarify the nature of their judgments. Two general positions are currently prevalent. One of these regards the student as a reporter who observes and transmits information on what takes place within a class. Menges[35] expresses this view when he says, "I believe that the instructor and his faculty colleagues, rather than students are the proper interpreters and weighters of student observations." The alternate view holds that college students are fully capable of functioning as "evaluators" as long as they are asked to judge only those aspects of teaching for which they have the appropriate background to make comparisons. This position allows students to participate with faculty colleagues and administrators in the evaluative process. Its supporters believe that college students have experienced enough teaching to be able to say that one professor is "outstanding, better than most, or only fair in comparison to other teachers I have known" in his or her efforts to promote understanding of the subject matter, or in stimulating or motivating more active intellectual efforts.

Here again, these different views dictate different types of student rating instruments. One of the logical consequences of viewing the student as a reporter is that the items placed on the questionnaire are chosen because *faculty* believe they describe the qualities most important to good teaching. The items may be derived from faculty discussion or from educational theory. They may even be subjected to rather elaborate methods of refinement, but their ultimate justification lies in their origin. The other consequence of this view of students is that they are provided only with descriptive—not evaluative—terms for registering their observations. The response categories on these questionnaires indicate frequency, amount, or agreement (e.g., rarely, sometimes, frequently; less than, about the same as, or more than in most courses; or strongly agree, agree, . . . strongly disagree).

Viewing students as "evaluators" entails selecting items for the questionnaire because they have been shown to carry weight in differentiating those teachers whom the *students* have judged as good from those they have evaluated as poor.[36] Additionally, students register their responses in evaluative terms; outstanding, excellent, better than most, competent, average, only fair, in need of serious improvement, poor.

Many existing instruments contain items for both formative and summative purposes, some of which require the student to be a reporter, some an evaluator. These combinations have resulted in most cases from a lack of awareness of the issues, and they contribute to confusion in how the results can be used.

Once these fundamental distinctions have been addressed, a college can then seek expert help in planning and implementing a sound system of teaching evaluation. This is not a matter to be entered into lightly. Each institution must decide whether improving the quality of its teaching is one of its goals and whether that goal is worth the effort to achieve it, including the effort which must go into evaluation. And here lies an issue so crucial that it cannot be ignored—the problem of how a system of evaluation can be initiated and fostered. Although pressure to reward teaching merit occasionally comes from faculty, the impetus for instituting sound evaluation procedures cannot, in general, be expected to originate with them. The idea of systematic evaluation in an area of professional functioning for which most faculty received little or no formal training, and precious little help or advice, is understandably threatening, and often engenders massive resistance. It is not surprising, then, that many professors prefer to be judged solely on their role as scholars, for which they have had long and arduous training. Nor do some faculty care to be judged in areas of performance which they know will *not* be rewarded. It is well to remember that within institutions of higher education, the visable rewards of salary increase and promotion are primarily controlled by deans and department chairpersons. Thus it is that serious efforts to evaluate teaching, either by peers or students, come about largely through the leadership of informed administrators. Even where resources of expertise exist, these will have limited effect without the administrative support which gently guides a faculty through discussions of the issues basic to evaluating teaching. And only when those who make academic decisions value the teaching role, attend to its different levels of merit, and reward it fairly are sound evaluative procedures sought. It is no accident that at each of the institutions nationally recognized as leaders in teaching evaluation, there are one or more academic officers who understand and stimulate these developments. Where evaluation efforts have floundered or failed, it is often for lack of administrative support.

Responsible concern for teaching quality goes beyond evaluation. Institutions that espouse this goal must provide resources for faculty development in the practice of this vital skill. The relatively recent recogni-

tion by colleges and universities of the nature of this responsibility underlies the currently emerging concept of faculty development, one principal entailment of which is direct assistance to professors who want to improve their effectiveness as teachers.

Attention to the quality of teaching will not solve all of the problems of academia. F.B. Morgan, Jr.[37] is right when he says of evaluation," . . . it will not usher in the Kingdom!" But an understanding of basic issues may reduce some of the controversy surrounding the choice of procedures and improve efforts to reward teaching fairly.

## REFERENCES

[1]The author wishes to express sincere appreciation for their valuable suggestions to Dr. Robert D. Marshall and Dr. Edward J. Shoben, Jr., both of whom read an earlier draft of this paper.

[2]I would like to emphasize that while this paper deals only with evaluation of the teaching responsibilities of faculty, there is no implication that the other academic functions are of lesser importance. For this reason the term "faculty member" or "professor" rather than "teacher" will be used throughout, except where teacher is needed for clarity.

[3]K.E. Eble, *The Recognition and Evaluation of Teaching,* (Washington, D.C.: American Association of University Professors, 1971).

[4]Dr. Wilbert J. McKeachie has written widely on student evaluation of teaching. His address, given at the April 1974 Conference on National Issues in Higher Education gives excellent advice, but can be obtained only as an audio-cassette from Kansas State University's Department of Continuing Education.

[5]"Statement on Teaching Evaluation," *AAUP Bulletin,* 60, 2, (June 1974), 168-170 and 61, 2, (August 1975), 200-202.

[6]R.I. Miller, *Developing Programs for Faculty Evaluation,* (San Francisco: Jossey-Bass, Inc., 1974).

[7]M. Scriven, The Methodology of Evaluation. American Educational Research Monograph Series on Curriculum Evaluation, No. 1, *Perspectives of Curriculum Evaluation,* (1967).

[8]R.I. Miller, *Evaluating Faculty Performance.* (San Francisco: Jossey-Bass, Inc., 1972).

AAUP Statement on Teaching, op. cit.

E.J. Shoben Jr., *Faculty Development, Evaluation, and Academic Recognition: A Proposal Regarding Salary Increments, Promotion and Tenure,* (Pittsburgh: University of Pittsburgh, 1974), Mimeographed.

[9]J.A. Centra, "Colleagues as Raters of Classroom Instruction," (New Jersey: *Research Bulletin,* 74-18, Educational Testing Service, Princeton, 1974).

[10]Ibid.

[11]A.M. Eastman, "How Visitation Came to Carnegie-Mellon University," (*Bulletin of the Association of Department of English,* May 1969).

[12]S. Edwards, "A Modest Proposal for the Evaluation of Teaching," (*Liberal Education,* Volume 60, Number 3, October 1974), 316-326.

[13]J.F. Noonan used this term in describing the institutional taboos against open discussion of teaching problems in his address at the 1974 Conference on National Issues in Higher Education.

[14]R.T. Blackburn and M.J. Clark, "An Assessment of Faculty Performance: Some Correlates Between Administrator, Colleague, Student and Self Ratings," in L.C. Buhl and S.H. Lane (Eds.) *Innovative Teaching: Issues, Strategies and Evaluation,* (Ohio: The Cleveland State University, 1973), 353-374.

[15]J.A. Centra, *Strategies for Improving College Teaching,* (Washington, D.C.: American Association for Higher Education, 1972).

[16]McKeachie, op. cit.

[17]This study was carried out while the author was Director of the Office of Student Evaluations of Teaching at the University of Washington. The factor analysis was performed by Barbara Jacobsen Meyers, and is reported in: E.R. Guthrie, *Evaluation of Teaching: A Progress Report,* (Seattle: University of Washington, 1954).

[18]The way in which these committees were selected is described later in this section.

[19]Guthrie, op. cit.

[20]Personal communication, November 1974.

[21]D.H. Naftulin, J.E. Ware Jr., and T.A. Donnelly, "The Doctor Fox Lecture: A Paradigm of Educational Seduction," *Journal of Medical Education,* (48, July 1973), 630-635.

[22]The principles which should guide the selection of peer evaluation committees are discussed later in this section.

[23]The number was not always six; it was usually six or seven, but very occasionally only five. These committees are described in E:R. Guthrie, "The Evaluation of Teaching," *Training Analysis and Development Informational Bulletin,* (USAF, Fall 1953, 4, 3), 199-206.

[24]Miller, op. cit.

[25]The point here is to provide a way by which the peer judged can call for additional information without interacting with other judges. One procedure that has been used directs such questions to the administrator handling the evaluation process, who obtains and transmits the information called for.

[26]Commission on Academic Tenure, *Faculty Tenure,* (San Francisco: Jossey-Bass, 1973), p. 60.

[27]L.M. Aleamoni, and R.E. Spencer, "The Illinois Course Evaluation Questionnaire: A Description of Its Development and A Report of Some of Its Results," *Educational and Psychological Measurement,* (1973, 33), 669-684.

A.T. Sharon, and C.J. Bartlett, "Effect of Instructional Conditions in Producing Leniency on Two Types of Rating Scales," *Personnel Psychology,* (1969, 22), 251-263.

W.G. Warrington, "Student Evaluation of Instruction at Michigan State University," in A.L. Sockloff (Ed.), *Proceedings: The First Invitational Conference on Faculty Effective-*

*ness as Evaluated by Students,* (Philadelphia, Pa.: Measurement and Research Center, Temple University, 1973), 164-182.

[28]This suggestion was made by Dean Robert D. Marshall in a letter to the Provost's Committee on Undergraduate Programs, University of Pittsburgh, November 1, 1974.

[29]An excellent review of this literature can be found in F. Costin, W.T. Greenough, and R.J. Menges, "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research,* (1971, 41), 511-535.

[30]The literature on what these qualities are is extensive. The following are simply indicative of different approaches to analyzing what these qualities are: K.E. Eble, *Professors as Teachers,* (San Francisco: Jossey-Bass, 1972).

W.J. McKeachie et al., *Research on the Characteristics of Effective College Teaching,* (U.S. Department of Health, Education, and Welfare, Office of Education, Project No. 850, Ann Arbor: University of Michigan, 1964).

[31]W.J. McKeachie "Research in Teaching: The Gap Between Theory and Practice," in B.T.C. Lee (Ed.), *Improving College Teaching,* (Washington D.C.: American Council on Education, 1967), 211-239. In this article, Dr. McKeachie summarizes the findings of over a hundred studies. The following statement is taken from his conclusions: "Where do we stand today with respect to teaching methods? Clearly, no one method is best for all goals, students, or teachers. Rather, what is the best method is a function of each of these variables."

[32]N.L. Gage, *Teacher Effectiveness and Teacher Education,* (Palo Alto, California: Pacific Books, 1972), 170.

[33]R.P. Kirschner, *A Controlled Factor in Teaching Evaluation of Students,* (Lexington, Kentucky: College of Education, University of Kentucky, 1969).

[34]R.I. Miller, *Developing Programs for Faculty Evaluation,* (San Francisco: Jossey-Bass, Inc. 1973), 184.

[35]R.J. Menges, "The New Reporters: Students Rate Instruction," in C.R. Pace (Ed.), *Evaluating Learning and Teaching,* (San Francisco: Jossey-Bass Inc., 1973).

[36]G. French-Lazovik, "Predictability of Students' Evaluation of College Teachers From Component Ratings," *Journal of Educational Psychology,* (1974, 66, 3), 373-385.

[37]F.B. Morgan, Jr., "Evaluating Teaching: Not Easy, Not Avoidable," (Association of Governing Boards of Universities and Colleges, *AGB Reports,* 17, 4, January/February 1974).



12