Volume XIX

Agenda and Minutes of the Meeting of November 8, 1982

WICHITA STATE UNIVERSITY SENATE

AGENDA

Meeting Notice: November 8, 1982, 126 Clinton Hall, 3:30 p.m.

Order of Business:

I. Calling of the Meeting to Order

II. Informal Proposals and Statements

III. Approval of the minutes for the meeting of October 25, 1982 (Vol. XIX, No. 4).

IV. New Business:

Postponed Nominations to Senate Committees

Statement on Professional Ethics
(Attachment A)

Report on Teacher Evaluation--Dr. Ben Rogers
(Attachment B)

Summary of Survey on Fringe Benefits--
Dr. James Clark
(Attachment C - will be distributed as soon as available)

V. Adjournment

ATTACHMENT A
To Agenda 11/8/82
Document 8

STATEMENT ON PROFESSIONAL ETHICS

I. Faculty members, guided by a deep conviction of the worth and dignity of the advancement of knowledge, recognize the special responsibilities placed upon them. Their primary responsibility to their subject is to seek and to state the truth as they see it. To this end they devote their energies to developing and improving their scholarly competence. They accept the obligation to exercise critical self-discipline and judgment in using, extending, and transmitting knowledge. They practice intellectual honesty. Although they may follow subsidiary interests, these interests must never seriously hamper or compromise a faculty member's freedom of inquiry.

II. As teachers, faculty members encourage the free pursuit of learning in their students. They hold before the student the best scholarly standards of their discipline. They demonstrate respect for the student as an individual and adhere to their proper roles as intellectual guides and counselors. They make every reasonable effort to foster honest academic conduct and to assure that they evaluate students according to their true merits. They respect the confidential nature of the relationship between teacher and student. They avoid any exploitation of students for private advantage and acknowledge significant assistance from them. They protect the student's academic freedom.

III. As colleagues, faculty members have obligations that derive from common membership in the community of scholars. They respect and defend the free inquiry of their associates. In the exchange of criticism and ideas they show due respect for the opinions of others. They acknowledge their academic debts and strive to be objective in their professional judgment of colleagues. They accept their share of faculty responsibilities for the governance of their institution.

IV. As members of their institution, faculty members seek above all to be effective teachers and scholars. Although they observe the stated regulations of the institution, provided these do not contravene academic freedom, they maintain their rights to criticize and seek revision of them. Faculty members determine the amount and character of the work they do outside their institution with due regard to their paramount responsibilities within it. When considering interruption or termination of their service, they recognize the effect of their decision upon the program of the institution and give due notice of their intention.

V. As members of their community, faculty members have the rights and obligations of any citizen. They measure the urgency of these obligations in the light of their responsibilities to their subject, their students, their profession, and their institution. When faculty members speak or act as private persons they avoid creating the impression that they speak or act for their college or university. As citizens engaged in a profession that depends upon freedom for its health and integrity, faculty members have a particular obligation to promote conditions of free inquiry and to further public understanding of academic freedom.

Received by Senate 11-8-82

ATTACHMENT B
To Agenda 11/8/82

Document 9

REPORT

of

The Ad Hoc Committee on Teaching Evaluation

TABLE OF CONTENTS

Recommendations amended and passed 11-22-82

		Page
I.	Preamble Committee charge - modus operandi	1
II.	Questionnaires Types of student questionnaires - facts known about them	2
III.	The Use of Information on Student Perceptions Improvement of teaching - personnel Judgments	4
IV.	Formative Evaluation Improving teacher performance	4
	A. Standardized Questionnaires Advantages - disadvantages - individual tests reviewed	4
	B. Forms Developed by Instructors Advantages - disadvantages	7
	C. The Effect of Student Evaluation Studies on this question reviewed	7
V.	Summative Evaluation Evaluation for personnel judgments	7
	A. The Place of Student Perceptions in Summative Evaluation Rationale for their use	8
	B. Difficulties with the Present Situation Findings of the committee concerning local practices	8
	C. Minimum Criteria for Adequate Student Input into Summative Evaluation Basis for committee recommendations	10
VI.	Recommendations University-wide questionnaire, technical committee	10
VII.	Appendix Description of forms used on campus	14

REPORT

of

The Ad Hoc Committee on Teaching Evaluation

Preamble

The original charge of this committee was to review the various procedures which have been developed, used, and evaluated for the purpose of providing information concerning the student perception of teaching in individual courses. The purpose of the review was to provide information to the Senate concerning various features of these procedures relevant to their use in teaching improvement and in the evaluation of teaching for tenure and promotion.

Our initial inquiries concerning types of procedures used on this and other campuses quickly showed that there are innumerable forms and questionnaires that have been used to solicit information from students about teachers and their teaching. However, these seem to break down into only a few basic types. Also, a quick search of the literature on the subject revealed a few other approaches to gaining such information, such as surveying alumni, and established that the results from the latter correspond well to the results obtained from surveys of current students (Centra, pp. 41-42). Since various kinds of questionnaires administered to current students are the cheapest and least obtrusive means of obtaining such information, and seem to produce the same results as other methods, we have concentrated on these.

Fortunately for the committee, a good deal of research has been done on methods for obtaining student perceptions, and in this report we rely heavily on recent authoritative reviews of this research by Centra and by McKeachie. As a consequence of these initial findings, the committee felt that it could not only provide the information requested by the Senate in its original charge but that it could also make some specific recommendations about the kind

of procedures which could best be put to specific uses. It petitioned the Agenda Committee for permission to enlarge its charge to make such recommendations, and this petition was approved.

After the Senate enlarged the charge of the committee, a preliminary draft of this document was written which reported the findings of the committee to that point but which did not contain specific recommendations. This draft, accompanied by reprints of material by Centra, Scriven, and McKeachie, was circulated to a representative faculty bodies in each of the colleges. Also, each of these groups was given for discussion a sketch of possible, but mutually incompatible, recommendations which had been canvassed by the committee. Then the committee, or a subcommittee thereof, met with each of these groups to discuss the work of the committee and recommendations it might make to the Senate. No formal action was requested of any of these groups with respect to the report or possible recommendations. Subsequent to meeting with these college bodies and considering their concerns and advice, the committee formulated its recommendations.

Questionnaires

The most widely used procedure for obtaining student perception of teaching is the administration of questionnaires of the type which are represented by LASTIC and IDEA forms on this campus. Such questionnaires have been proven to be statistically reliable by a number of tests (Centra, pp. 26-28). Practically speaking, this means that the questionnaires are a good means for getting student perceptions of teaching. The crucial question which everyone wants answered is whether student perception of the quality of teaching has any relationship to the actual quality or effectiveness of teaching.

As soon as the question is raised, there is an immediate difficulty. If one is to relate the student perceptions to teaching effectiveness, there must be some independent way to judge and rate teaching effectiveness. That is very

difficult to do. First, there are usually a great number of different things a teacher wishes to accomplish in a single course. Socond, there is the further question as to the desirability of the things the teacher wishes to accomplish. Third, there is the difficulty of measuring what the students bring to the class and what they gain from the class. Finally, in order to judge teacher effectiveness one must know what is a typical gain, in order to have a meterstick for comparison. Usually, most of these things are not known. There is some measure of comfort available from the fact that in the few well-designed experiments which have been performed on one accepted criterion--amount of relevant material learned, there is significant positive correlation between the amount of material learned and student ratings of their teachers (Centra, pp. 36-38).

Another area of frequent faculty speculation about such questionnaires concerns the effect on student ratings of teaching effectiveness of factors other than the quality of teaching. There have been literally hundreds of studies of this kind done. The best-established results follow (Centra, pp. 28-34).

Student characteristics having little or no relationship to ratings of teacher effectiveness are: sex, grade point average, college year, academic ability, and age. Class size has an effect: classes of less than 15 have the highest ratings, followed in order by 16-35, over 100, 36-99. There are slightly higher ratings in humanities classes compared with social sciences and natural sciences. Also students give slightly higher ratings in major courses than the courses taken to fulfill college requirements.

Grade anticipated in courses did not directly affect ratings of teacher effectiveness. However, in questionnaires given after final grades were known there is a difference in ratings given by those who got the grade they expected or higher and those who received a lower grade than they were expecting.

Teachers in their first year as a group get the lowest ratings. After that, there is no appreciable difference, until there is a slight decline after twelve years of teaching. Academic rank, sex, teaching load, and research productivity have little effect on student perception of teaching effectiveness.

The Use of Information on Student Perceptions

There are two principal uses to which the information obtained concerning student perceptions are put: teaching improvement, and summative evaluation for determination of tenure, promotion, and salary. Those procedures, forms, and questionnaires which are appropriate for various aspects of one of these purposes are not necessarily appropriate for the other purpose. Some questionnaires, properly administered, are plausible candidates to use for both purposes. At this university, various items have been used for both, sometimes correctly, sometimes not. It is the purpose of this report to make some recommendations to improve this situation.

Formative Evaluation

It is useful to have a term to describe that kind of evaluation which is aimed primarily at obtaining information to be used directly for changing course parameters and teaching styles, i.e., aimed at improving teaching performance. In the literature on evaluation this kind of evaluation is called formative evaluation. We shall adopt this terminology. The principal means to gather such information is by the use of standardized questionnaires and instructor-developed forms.

Standardized Questionnaires

There are considerable advantages in using standardized questionnaires to obtain student perceptions of one's teaching. These advantages flow from the

standardization itself. First, the scores generated allow teachers to compare themselves to the group on which the form was standardized. If the standardization group is representative in the appropriate ways, teachers have a good idea of how their teaching stands with respect to their peer group on a number of different items or factors. Second, standardized tests of this kind have generally been shown to be reliable, and the reliability of the particular test in use may be known. Third, the data from the standardization group can be grouped by statistical techniques, of which factor analysis is one, giving a better indication of what area of teaching performance each item of the form samples. This kind of analysis increases the reliability of the scores that the individual instructor receives. Fourth, the effect of different variables, such as class size, on the student perception of teaching effectiveness can be studied and account of these effects can be taken in interpreting the evaluation.

Standardized tests have their weaknesses, however. They are not open-ended, and valuable information about the course or the teacher may be missed. Futher, the standardized test is not responsive to information about particular features of the course design, different teaching styles or methods, and may be biased toward one particular teaching mode (such as lecturing).

A comparison of actual standardized forms is useful in understanding the strengths and weaknesses of different approaches to sampling student perceptions. The LASTIC questionnaire, for example, has a local data base, which allows WSU instructors to compare themselves to the local peer group in interaction with the local students. Having a local base allows one to study the effect on student perceptions of local conditions, which may be different from those found nationally, which results were reported in an earlier section. LASTIC results are presented in the form of scores on six factors, which is a reliable kind of presentation. However, there is a possible bias built into the data base in that

the base was formed of teachers who volunteered to administer the tests. This may mean that the base group had higher scores than the group of all faculty would have, since those in doubt about their performance would be less likely to give the questionnaire. LASTIC does not permit the instructor to enter new questions into the form, but it does seek open-ended reponses.

The IDEA form has national data base which is useful for seeing how WSU teaching compares to the national norm base but is less useful for comparison with the local talent teaching the local student. The absence of a local data base also precludes studying the effects of local conditions. Another disadvantage of IDEA is that the scores are reported only on each item rather than being grouped, say by factor analysis. This reduces the reliability of the result and is more likely to lead to misinterpretation. The universities and colleges in which IDEA was given in order to establish a norm base were chosen because the writers of IDEA had friends there. In addition, there is no evidence that the teachers who used IDEA in these colleges and universities were not self-chosen. So there is reason to suspect that the norm base for IDEA is every bit as biased as we suspect LASTIC of being with respect to being normed over teachers who get good student response on an informal basis. And since the colleges and universities in which IDEA was normed were not chosen randomly, they probably do not constitute a representative national norm base.

One interesting type of questionnaire in use at Purdue, Michigan, and Illinois allows each teacher to choose from 200 items which are to be added to a non-optional core of five items. This allows for considerable flexibility in tailoring the questionnaire to the needs of a particular instructor but one loses the advantages of standardization on the optional questions. Many questionnaires allow the instructor to add five questions to the standard group. Of course, the scores on these questions do not allow for inter-peer-group comparison. Most other forms in general use seem to be only minor variants of the types exemplified by the forms discussed above.

Forms Developed by Instructors

In some ways the strengths and weaknesses of forms developed by individual instructors are mirror images of standardized questionnaires. Their strengths are flexibility and specificity. They can be open-ended, formed to the needs of the individual course or instructor, and designed to provide information about specific aspects of texts, courses, or methods of presentation. On the other hand, the interpretation of scores is more difficult because there is no standard against which to judge the results. Also, the scores are of unknown reliability, and the sophistication of the forms and its analysis is heavily dependent on the sophistication of the instructor who develops the form.

There are also considerable ethical responsibilities which the instructor bears directly, with regard to neither seeking personal information about the student nor otherwise disregarding their anonymity. These safeguards are usually assured for the standardized test.

The Effect of Student Evaluation

Does student evaluation bring about improvement of teaching? It is clear that the questionnaires give a certain amount of information about how the students perceive the course and the instructor. It is often less clear what needs to be changed to bring about a better score. Studies of the effectiveness of student evaluation show that if they are obtained in mid-course and if changes are effected as a result of the inquiry then there are improvements in scores on subsequent tests (Centra, pp. 38-41). There is no direct evidence that teacher effectiveness improves, largely because the study of this question is difficult and does not seem to have been done.

Summative Evaluation

Summative evaluation is the jargon for evaluation for the purposes of tenure, promotion and salary. The comments on standardized questionnaires and instructor-

constructed forms made in the above section of formative evaluation are largely applicable with respect to summative. Additional aspects of these forms relevant to summative evaluation are considered below.

The Place of Student Perceptions in Summative Evaluation

Since the students are the persons most directly affected by good or poor teaching, it seems only just that they have a say about the quality of teaching received. Thus, it is necessary to have an appropriate vehicle for students to make known their evaluation of the instruction they receive.

However, there are aspects of teaching effectiveness that students are not likely to be in a position to judge, for example, whether or not the course content is current. Hence, on this and similar matters, there is no substitute for appropriate peer judgement. Hence, evaluation of teaching by students should not be the only evidence used in the determination of teaching quality. Other kinds of evidence are necessary. A large list of kinds of relevant evidence is given in the report of the Blue Sky Committee. More specific recommendations are given by Scriven (see bibliography).

Difficulties with the Present Situation

The committee has received a large number of remarks from faculty, students, and administrators expressing their concerns with present practices in regard to the collection and use of student perceptions of teaching effectiveness in summative evaluation. It judges the following concerns and difficulties to be valid. They constitute important issues which any proposal concerning future practices must address.

1. There is no single standard for presenting student perceptions, so comparisons among faculty are difficult. Yet evaluators are called upon to make comparative judgements.

2. LASTIC is probably the form most widely presented as evidence of teaching effectiveness. While it is intrinsically sound in concept as a measure of student perception of teaching performance and has the advantage of being normed over a local population, the present method of administration allows for the significant manipulation of its results. Also as previously mentioned, the base group is likely biased toward high scores, and no study has been made of the effects on scores of conditioning on different subgroups. A revision of the LASTIC form is underway, and some of the defects will be corrected.
3. IDEA is also widely used here. There is no local data base so local comparisons cannot be made directly. The item presentation of scores easily leads to misinterpretation of the scores for summative evaluation. The usual means of administration is open to manipulation.
4. Results on instructor-generated forms, while useful for certain specific needs in formative evaluation, provide no valid basis for comparison with other faculty, so they have a very limited use in summative evaluation.
5. Quite often instructors provide an insufficient or unrepresentative series of questionnaire results for summative purposes. According to Centra, (p. 27) at least five courses must be sampled in order to insure dependability if the number of students in the courses is at least 15. If as many as ten courses are sampled, the number of students in the courses makes little difference.
6. Quite frequently student perception questionnaires are the only form of evidence presented for summative evaluation. This leaves out important factors in the proper evaluation of teaching.
7. The committee received a number of complaints that indicated students in substantial numbers were not taking seriously filling out LASTIC and IDEA forms. Two hypothesis as to the cause are: the forms are too long, the students are not made adequately aware of the real use and importance of the questionnaire results.

8. Certain questions on many forms are biased toward certain styles or methods of teaching, e.i. lecturing (see Scriven, p. 253).

Minimum Criteria for Adequate Student Input into Summative Evaluation

1. The administration of questionnaires must insure the integrity of the results and the anonymity of the students.
2. Ideally, the results should be interpreted with respect to an adequate, local, unbiased data base.
3. The means of gathering the information must be relatively unobstrusive. Specifically, there should be control of the length of the form, the means of administration, and the frequency of administration.
4. How to interpret the results correctly must be widely understood by those making decisions. In particular, small differences in scores should not be used to rank persons.
5. A fairly large sample of instructor's courses should be represented, and these should not be chosen by the instructor. Either the sample should be random or exhaustive.
6. Sufficient statistical analysis of the norm group should be carried out on the effects of different teaching conditions to allow reasonable interpretation of individual scores. In interpreting the scores allowance should be made for biases discovered in such statistical analyses.

Recommendations

On the basis of the foregoing considerations, the committee makes the following recommendations:

1. That the faculty mandate the adoption of a university-wide questionnaire for obtaining student perceptions of the teaching effectiveness of individual instructors for use in summative evaluation.

It is only just that students have a reliable vehicle for the transmission of their views concerning the effectiveness of their instructors to those who are responsible for the professional evaluation of these instructors. Such a questionnaire, in wide-spread use, is an appropriate means to accomplish this end.

Members of the faculty, about whom judgments regarding tenure, promotion and salary are made, deserve to have available to them reliable information concerning their student's perceptions of their teaching. Presently such information does indeed enter into their evaluations in all sorts of ways, many of which are of dubious reliability. The creation and implementation of a well-designed questionnaire, correctly administered, will provide a much more reliable, systematic, and fair means of providing this component of teacher evaluation than do current practices.

2. A technical committee should be created to design such a questionnaire, or adopt an existing one, to design a process for its correct administration, and to oversee its administration.

   This technical committee should:

   a) be composed of persons professionally qualified to create or judge the proper design of such tests and of other members of the faculty from areas not directly concerned with such design;

   b) consult with colleges and departments with regard to problems they have experienced in the use of previous questionnaires of this type so as to try to avoid these difficulties in the construction or adoption of this one;

   c) provide faculty groups and administrators involved in faculty evaluation with guidelines for the correct interpretation of questionnaire outcomes in the evaluation of individual faculty members; in particular they should investigate the effect of evaluation with respect to different norm groups and other such

studies as they deem relevant to determining correct interpretive guidelines;

d) be a continuing committee charged with the periodic review of the questionnaire and the responsibility to revise it as experience proves necessary.

This committee is necessary for the creation and maintenance of the questionnaire mandated in the first recommendation.

BIBLIOGRAPHY

Centra, John A., Determining Faculty Effectiveness, San Francisco: Jossey-Bass, 1979.

McKeachie, Wilbert J., "Student Ratings of Faculty: A Reprise" in Academe: The AAUP Bulletin, Vol. 65, No. 6 (October, 1979).

Scriven, Michael, "Summative Teacher Evaluation" in Millman, ed., Handbook of Teacher Evaluation.

Report of the Blue Sky Committee, WSU. Available from Dr. Tom Maher.

Appendix

IDEA

The IDEA survey form for obtaining student reactions to instruction and courses was developed by the Center for Faculty Evaluation and Development at Kansas State University. It is a questionnaire form with 45 items. It uses a five point Likert-type scale (that is, for each item there is a choice of five responses, say, from strongly agree to strongly disagree). The basic data base includes responses from roughly 17,000 classes. One of the unusual features of IDEA is that it attempts to measure the degree of motivation of each student taking the course. The information returned to the instructor is normed against other courses of similar size in which students report similar levels of motivation. The report to the instructor consists of an item by item analysis of student response, which is a generally less reliable form of information for summative purposes than is a factor analysis report. It may provide an instructor with valuable information about a specific strength or weakness for formative purposes. The form is administered by the instructor. Current costs are $0.33 per student form which includes the cost of processing and report to the instructor.

LASTIC

The LASTIC student survey form was developed at WSU by the Liberal Arts and Sciences Teacher Imporvement Committee. It consists of a questionnaire that contains 11 demographic items and 39 instructor/course evaluative items. The responses to the evaluative items are recorded on a five-point Likert-type scale. Provision is also made for student response to open-ended questions. The data base consists of 42,019 questionnaires

collected from 2,115 class sections. Responses are reported to instructors both by an analysis of individual items and a factor analysis that shows the insturctor's standing with respect to six factors. The format for reporting results is currently under review. The factor analysis format is a statistically reliable means of providing information for summative review, provided that the the questionnaire is properly administered. Currently, the questionnaire is administered by student proctors from the class under evaluation. It has been difficult to determine the cost of administering LASTIC.

College of Business Administration's SPTP Form

The SPTP form was developed at WSU in the College of Business Administration and has been adopted by its faculty specifically for teacher evaluation. It consists of a questionnaire with six demographic items and ten evaluative items, two of which are open-ended. For the remaining evaluative items, responses are recorded on a five-point Likert-type scale. In order to establish a college-wide basis for evaluation, the Business faculty voted to have a mandatory evaluation in each class in the college for a two semester period, with the stipulation that the mandatory evaluation for all classes would not constitute a precedent establishing procedure. This was done last year. The primary advantage of this questionnaire, aside from it universal data base, is its brevity. It requires only five to ten minutes of class time to administer. The main disadvantage is that it is a relatively unsophisticated statistically. There is only one item per area of evaluation, and this poses problems with reliability and validity. Costs are not yet available.

Music Performance Department Forms

Forms for student evaluation of music performance instruction were developed because most of the instruction in that department takes place in the private studio or in the rehearsal of musical ensembles and they perceived that neither activity is completely compatible with the approach and questions of other forms used in the university. The form consists of 18 to 21 questions, the responses to which are recorded on a five-point Likert-type scale. In addition, there is one open-ended question. No body of statistical data has been assembled from these evaluations, so interfaculty comparisons are impossible as well as judgments as to reliability. Costs of administration are unavailable.

CHRP Course/Instructor Evaluation Questionnaire

The College of Health Related Professions developed a Course/Instructor Evaluation Questionnaire. It consists of 20 questions, of which four are demographic. The responses to the remaining questions are recorded on a five-point Likert-type scale. The use of this form was mandatory for several semesters in all lecture classes with 10 or more students and three or fewer instructors. The data base from this use is available. Mandatory student evaluation of teaching has since been discontinued. The cost of printing and administering this form is $0.074 per student, with no charge for scoring.

In addition, there was developed in CHRP a questionnaire for use in clinical courses. The form consists of 30 questions. Responses are recorded on a five-point Likert-type scale. The form was developed because, like the music performance faculty, the clinical faculty felt that the more widely used forms were not appropriate in the clinical setting. It was used on a trial basis by several faculty, but no comprehensive record exists. Costs are unavailable.

to give more importance to peer judgments at other institutions and less importance to the sheer number of articles in all types of journals.

Overview

In the remaining chapters the evaluation of faculty performance will be discussed as it relates to using the collected information to make personnel decisions or to improve faculty performance in their various activities. Evaluators refer to the former as *summative evaluation* because it "sums up" performance at the end of a time period and results in some kind of overall judgment. Evaluation to improve performance can be called *formative* because it is meant to help "form" performance while it is in progress. Chapters Two through Five discuss the evaluation of teaching—the strengths and limitations of student ratings (Chapter Two), self-assessment and self-analysis, which includes audio and video playback as well as the so-called individualized development plan (Chapter Three), colleague evaluations (Chapter Four), and the assessment of student learning (Chapter Five). Chapter Six considers the evaluation of research, advising, and public service, and Chapter Seven discusses legal concerns in personnel policy. Key questions are raised in each of the chapters, research related to each of these questions is summarized, and recommendations are offered at the chapter's conclusion.

The final chapter summarizes some key points and details a comprehensive model of faculty evaluation with reference to other models. It also discusses the administrator's role and methods of using evaluation information for personnel decisions.

Senate Office

John A. Centra, Determining Faculty Effectiveness, San Francisco: Jossey-Bass, 1979.

TWO

Uses and Limitations of Student Ratings

Despite some strong opposition to incorporating student ratings in faculty evaluation, they are widely used and endorsed by both students and faculty members. Seventy-two percent of responding college freshmen in the 1977 annual survey by the American Council on Education (ACE) felt that they should help to evaluate faculty performance (Astin, 1978). The past half dozen ACE surveys reported similar findings. And in 1972 nearly 70 percent of a national sample of faculty members agreed that "faculty promotions should be based in part on formal student evaluations of their teaching" (Bayer, 1973).

Student ratings take place at the end of the course and are generally described as informal or formal. Informal student evaluations involve occasional comments to the instructor or the dean by a few students in the class. Such opinions may not represent the views of all the students and are therefore less useful than those comprehensive or formal systems that ask students for written responses to a set of

This and following two documents Distributed with Report on Teacher Evaluation

short-answer, open-ended questions concerning the course and the teaching methods.

What Are the Characteristics of Effective Teaching?

Many techniques have been used over the years to identify potentially useful items for inclusion in formal systems of rating teaching and courses. A widely used method requests the opinions of faculty members, students, administrators, and alumni. Using such consultation, a University of Toledo study (Perry and Baumann, 1973) identified some sixty teacher behaviors that students, faculty, and alumni associated with effective teaching. Ranked high in teaching value by all three groups were such behaviors as being well prepared for class and exhibiting interest in the subject under study, while items with low teaching value included being neatly dressed or having irritating personal mannerisms.

Hildebrand, Wilson, and Dienst (1971) sought to describe effective teaching by a survey of students and faculty at the University of California, Davis. Students—asked to identify the best and the worst teachers that they had had in the previous year and to describe their teaching—included as distinguishing features of good teaching such items as "explains clearly, . . . seems to enjoy teaching, . . . makes difficult topics easy to understand, . . . knows if class is understanding the teacher or not, . . . keeps well informed about class progress, . . . is sensitive to student's desire to ask a question." Faculty members queried in the same survey listed such comments on good teaching by their colleagues as "seems to have a congenial relationship with students, . . . uses well-chosen examples to clarify points, . . . emphasizes *ways* of solving problems rather than solutions, . . . is an excellent public speaker" (emphasis added).

In 1975, Wotruba and Wright summarized twenty-one studies in which various groups had been asked to identify the qualities of effective teaching. The resulting list, typical of what would be found in most studies of this kind, included ten most frequently named characteristics:

- Communication skills—clearly interprets abstract ideas and theories
- Favorable attitudes toward students

- Knowledge of subject
- Good organization of subject matter and course
- Enthusiasm about subject
- Fairness in examinations and grading
- Willingness to experiment—flexible
- Encouragement of students to think for themselves
- Interesting lecturer—good speaking ability

Studies such as these have generated a pool of items that institutions can use in developing the hundreds of rating forms that have been developed over the years. (The best-known of these instruments are briefly described in the Appendix.) Two commercial forms currently used by the largest number of colleges are the Educational Testing Service's (ETS) Student Instructional Report (SIR) and the Instructional Development and Effectiveness Assessment System (IDEA) from Kansas State University (see Exhibits 5 and 6).

What Should Be the Content of Rating Forms?

Factor analysis studies of student ratings published over the past twenty-five years (examples are Coffman, 1954; Hodgson, 1958; Isaacson and others, 1964; Centra, 1973a) have identified several common dimensions or groups of items. Three of those appearing in the majority of instruments devised are (1) organization, structure, or clarity, (2) teacher-student interaction or rapport, and (3) teaching skill, communication, or lecturing ability (see Exhibit 7). Other categories included occasionally in rating instruments are evaluations of the course workload or difficulty, grading and examinations, impact on students (self-rated student accomplishment), and global or overall effectiveness.

Teachers can add five to ten optional items to most machine-scored forms in order to obtain student reactions to such areas as particular assignments, tests, and techniques used. With optional items, students can also rate the achievement of specific course objectives. Grasha (1977) describes one such procedure. The instructor, after reviewing each objective, asks students to identify any objective that has not been made clear during the course and to explain the reasons for this lack of clarity. Students are asked to indicate their difficulties in achieving each objective and to name the factors re-

Exhibit 5. Student Instructional Report

SIR Report Number

This questionnaire gives you an opportunity to express anonymously your views of this course and the way it has been taught. Indicate the response closest to your view by blackening the appropriate oval. Use a soft lead pencil (preferably No. 2) for all responses to the questionnaire. Do not use an ink or ball point pen.

SECTION I Items 1-20. Blacken one response number for each question.

- NA (0) = Not Applicable or don't know. The statement does not apply to this course or instructor, or you simply are not able to give a knowledgeable response.
- SA (4) = Strongly Agree. You strongly agree with the statement as it applies to this course or instructor.
- A (3) = Agree. You agree more than you disagree with the statement as it applies to this course or instructor.
- D (2) = Disagree. You disagree more than you agree with the statement as it applies to this course or instructor.
- SD (1) = Strongly Disagree. You strongly disagree with the statement as it applies to this course or instructor.

		NA	SA	A	D	SD
1	The instructor's objectives for the course have been made clear					
2	There was considerable agreement between the announced objectives of the course and what was actually taught					
3	The instructor used class time well					
4	The instructor was readily available for consultation with students					
5	The instructor seemed to know when students didn't understand the material					
6	Lectures were too repetitive of what was in the textbook(s)					
7	The instructor encouraged students to think for themselves					
8	The instructor seemed genuinely concerned with students' progress and was actively helpful					
9	The instructor made helpful comments on papers or exams					
10	The instructor raised challenging questions or problems for discussion					
11	In this class I felt free to ask questions or express my opinions					
12	The instructor was well prepared for each class					
13	The instructor told students how they would be evaluated in the course					
14	The instructor summarized or emphasized major points in lectures or discussions					
15	My interest in the subject area has been stimulated by this course					
16	The scope of the course has been too limited; not enough material has been covered					
17	Examinations reflected the important aspects of the course					
18	I have been putting a good deal of effort into this course					
19	The instructor was open to other viewpoints					
20	In my opinion, the instructor has accomplished (is accomplishing) his or her objectives for the course					

SECTION II Items 21-31. Blacken one response number for each question.

21. For my preparation and ability, the level of difficulty of this course was:
   - Very elementary
   - Somewhat elementary
   - About right
   - Somewhat difficult
   - Very difficult

22. The work load for this course in relation to other courses of equal credit was:
   - Much lighter
   - Lighter
   - About the same
   - Heavier
   - Much heavier

23. For me, the pace at which the instructor covered the material during the term was:
   - Very slow
   - Somewhat slow
   - Just about right
   - Somewhat fast
   - Very fast

24. To what extent did the instructor use examples or illustrations to help clarify the material?
   - Frequently
   - Occasionally
   - Seldom
   - Never

Questionnaire continued on the other side

572MRC116P200X

283562

Exhibit 5. (continued)

25. Was class size satisfactory for the method of conducting the class?
   - Yes, most of the time
   - No, class was too large
   - No, class was too small
   - It didn't make any difference one way or the other

26. Which one of the following best describes this course for you?
   - Major requirement or elective within major field
   - Minor requirement or required elective outside major field
   - College requirement but not part of my major or minor field
   - Elective not required in any way
   - Other

27. Which one of the following was your most important reason for selecting this course?
   - Friend(s) recommended it
   - Faculty advisor's recommendation
   - Teacher's excellent reputation
   - Thought I could make a good grade
   - Could use pass/no credit option
   - It was required
   - Subject was of interest
   - Other

28. What grade do you expect to receive in this course?
   - A
   - B
   - C
   - D
   - Fail
   - Pass
   - No credit
   - Other

29. What is your approximate cumulative grade-point average?
   - 3.50-4.00
   - 3.00-3.49
   - 2.50-2.99
   - 2.00-2.49
   - 1.50-1.99
   - 1.00-1.49
   - Less than 1.00
   - None yet

30. What is your class level?
   - Freshman
   - Sophomore
   - Junior
   - Senior
   - Graduate
   - Other

31. Sex
   - Female
   - Male

SECTION III Items 32-39. Blacken one response number for each question.

		Not applicable or don't know or there were too few	Excellent	Good	Satisfactory	Fair	Poor
32.	Overall, I would rate the textbook(s)						
33.	Overall, I would rate the supplementary readings						
34.	Overall, I would rate the quality of the exams						
35.	I would rate the general quality of the lectures						
36.	I would rate the overall value of class discussions						
37.	Overall, I would rate the laboratories						
38.	I would rate the overall value of this course to me as						

39. Compared to other instructors you have had (secondary school and college), how effective has the instructor been in this course? (Blacken one response number.)

One of the most effective (among the top 10%)	More effective than most (among the top 30%)	About average	Not as effective as most (in the lowest 30%)	One of the least effective (in the lowest 10%)

SECTION IV Items 40-49. If the instructor provided supplementary questions and response options, use this section for responding. Blacken only one response number for each question.

	NA											NA									
40.												45.									
41.												46.									
42.												47.									
43.												48.									
44.												49.									

If you would like to make additional comments about the course or instruction, use a separate sheet of paper. You might elaborate on the particular aspects you liked most as well as those you liked least. Also, how can the course or the way it was taught be improved? PLEASE GIVE THESE COMMENTS TO THE INSTRUCTOR.

Exhibit 6. Instructional Development and Effectiveness Assessment System (IDEA)

IDEA SURVEY FORM -- STUDENT REACTIONS TO INSTRUCTION AND COURSES

Your thoughtful answers to these questions will provide helpful information to your instructor.

• Describe the frequency of your instructor's teaching procedures, using the following code.
1 — Hardly Ever 3 — Sometimes
2 — Occasionally 4 — Frequently 5 — Almost Always

The instructor:

1. Promoted teacher-student discussion (as opposed to mere responses to questions).
2. Found ways to help students answer their own questions.
3. Encouraged students to express themselves freely and openly.
4. Seemed enthusiastic about the subject matter.
5. Changed approaches to meet new situations.
6. Gave examinations which stressed unnecessary memorization.
7. Spoke with expressiveness and variety in tone of voice.
8. Demonstrated the importance and significance of the subject matter.
9. Made presentations which were dry and dull.
10. Made it clear how each topic fit into the course.
11. Explained the reasons for criticisms of students' academic performance.
12. Gave examination questions which were unclear.
13. Encouraged student comments even when they turned out to be incorrect or irrelevant.
14. Summarized material in a manner which aided retention.
15. Stimulated students to intellectual effort beyond that required by most courses.
16. Clearly stated the objectives of the course.
17. Explained course material clearly, and explanations were to the point.
18. Related course material to real life situations.
19. Gave examination questions which were unreasonably detailed (picky).
20. Introduced stimulating ideas about the subject.

• On each of the objectives listed below, rate the progress you have made in this course compared with that made in other courses you have taken at this college or university. In this course my progress was:
1 — Low (lowest 10 per cent of courses I have taken here)
2 — Low Average (next 20 per cent of courses)
3 — Average (middle 40 per cent of courses)
4 — High Average (next 20 percent of courses)
5 — High (highest 10 per cent of courses)

Progress on:

21. Gaining factual knowledge (terminology, classifications, methods, trends).
22. Learning fundamental principles, generalizations, or theories.
23. Learning to apply course material to improve rational thinking, problem-solving and decision making.
24. Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course.
25. Learning how professionals in this field go about the process of gaining new knowledge.
26. Developing creative capacities.
27. Developing a sense of personal responsibility (self-reliance, self-discipline).
28. Gaining a broader understanding and appreciation of intellectual/cultural activity (music, science, literature, etc.).
29. Developing skill in expressing myself orally or in writing.
30. Discovering the implications of the course material for understanding myself (interests, talents, values, etc.).

• On the next four questions, compare this course with others you have taken at this institution, using the following code:
1 — Much Less than Most Courses
2 — Less than Most
3 — About Average
4 — More than Most
5 — Much More than Most

The Course:

31. Amount of reading
32. Amount of work in other (non-reading) assignments
33. Difficulty of subject matter
34. Degree to which the course hung together (various topics and class activities were related to each other)

• Describe your attitudes toward and behavior in this course, using the following code:
1 — Definitely False
2 — More False than True 4 — More True than False
3 — In Between 5 — Definitely True

Self-rating:

35. I worked harder on this course than on most courses I have taken.
36. I had a strong desire to take this course.
37. I would like to take another course from this instructor.
38. As a result of taking this course, I have more positive feelings toward this field of study.
39. I have given thoughtful consideration to the questions on this form.

• Describe your status on the following by blackening the appropriate space on the Response Card.

A. To which sex-age group do you belong?
1 — Female, under 25 3 — Female, 25 or over
2 — Male, under 25 4 — Male, 25 or over

B. Do you consider yourself to be a full-time or a part-time student?
1 — Full-time
2 — Part-time

C. Counting the present term, for how many terms have you attended this college or university?
1 — 1 term 3 — 4 or 5
2 — 2 or 3 4 — 6 or more

D. What grade do you expect to receive in this course?
1 — A 3 — C
2 — B 4 — D or F 5 — Other

E. What is your classification?
1 — Freshman 3 — Junior or Senior
2 — Sophomore 4 — Graduate 5 — Other

F. For how many courses have you filled out this form during the present term?
1 — This is the first course
2 — 2 or 3 courses 3 — 4 or more courses

G. How well did the questions on this form permit you to describe your impressions of this instructor and course?
1 — Very well 3 — Not very well
2 — Quite well 4 — Poorly

If your instructor has extra questions, answer them in the space designated on the Response Card.

Your comments are invited on how the instructor might improve this course or teaching procedures. Use the back of the Response Card (unless otherwise directed).

Exhibit 7. Factors or Categories of Ratings and Examples of Items

1. Organization, Structure, or Clarity
   - Material presented in an orderly manner
   - Instructor well prepared for each class
   - Class time well spent
   - Course well organized
   - Instructor made clear what we were expected to learn
   - Considerable agreement between announced objectives and what was taught
2. Teacher-Student Interaction or Rapport
   - Instructor readily available for consultation with students
   - Instructor seemed to know when students didn't understand the material
   - Instructor actively helpful when students had difficulty
   - Students felt free to ask questions or express opinions
   - Instructor seemed concerned with whether students learned the material
3. Teaching Skill, Communication, or Lecturing Ability
   - Instructor used examples or illustrations to clarify the material
   - Instructor spoke audibly and clearly
   - Instructor presented material clearly
   - Instructor summarized or emphasized major points in lectures or discussions
4. Workload, Course Difficulty
   - In relation to other courses, this workload was heavy
   - Instructor tried to cover too much material
   - Reading assignments were very difficult
   - Course challenged me intellectually
   - I put a great deal of effort into this course
5. Grading, Examinations
   - Instructor told students how they would be evaluated
   - Examinations reflected the important aspects of the course
   - Instructor made helpful comments on papers or exams
   - Instructor assigned grades fairly and impartially
6. Impact on Students, Student Self-Rated Accomplishments
   - I learned a great deal in this course
   - This course generally fulfilled my goals
   - This course stimulated me to want to take more work in the same or a related area
7. Global, Overall Ratings
   - Instructor's effectiveness as a teacher was: (excellent to poor)
   - Overall value of the course was: (excellent to poor)
   - Instructor made a major contribution to the value of this course
   - General quality of lectures was: (excellent to poor)
   - General quality of class discussions was: (excellent to poor)

sponsible for such problems. An alternate approach is to ask students to rate individually their progress toward each specific content objective.

In format, most rating instruments encourage students to add comments about the course and the way it was taught. Open-ended questions—such as "How do you think the course can be improved?" or "What did you like least (and most) about the course?"—can be used to elicit student reactions that single-response questions fail to tap. Some instructors and departments prefer to use such open-ended questions. Students might understandably feel constrained, however, about making negative comments, fearing that their course grades would be affected; typing would ensure anonymity, but the cost and time involved in this practice limits its use. Responses to open-ended questions are also vulnerable to more subjective interpretation than are answers to single-response questions. If the comments are used in a summary evaluation of the instructor, care must be taken not to put undue weight on a few highly negative or highly positive remarks.

Including an optional section or open-ended questions on rating forms allows instructors greater flexibility in improving the course and their teaching practices. Another way to give individual teachers some flexibility and choice, devised at Purdue University, is known as the "cafeteria system" (see the Appendix). This computer-assisted system allows teachers to choose from a catalogue listing 200 items. Selections are added to a nonoptional core of five items, and the computer prints individually designed questionnaires. As with standard forms, the computer then scores and processes student responses. Other institutions such as the University of Illinois and the University of Michigan have similar rating programs.

Although the major use of student ratings is to guide instructional improvement and faculty summative evaluations, another use is to help students in their choice of courses and teachers. Forms for this use include items rating the evaluation of class sessions, the relevance of the course, the amount of work required, and the degree of teacher interest in the class. Such information is also useful to teachers for course improvement. In practice, most of the items typically found in forms designed for student use are included in instruments designed for teacher use, so the contents of the two types of instruments generally differ much less than do their intended audiences.

Some student government groups and other student organizations collect and publish student rating information. As might be expected, student-produced critiques vary considerably in quality from one institution to another and from one year to the next, depending on the students involved. A few student-sponsored rating programs allow teachers to see their personal ratings prior to publication in order to include their comments on the findings or to provide additional information about the courses. This approach is more impartial, and it provides information about the courses that is not usually found in college catalogues. Unfortunately, certain student-sponsored course critiques are based on ratings by small and nonrepresentative samples of students or may emphasize some of the more critical comments about a teacher, regardless of how generalized such comments may be, thus leading many faculty members to object to student-conducted course ratings. Critics also argue that student-sponsored ratings result in "instructor hostility, resentment, and distrust," and thus alienate faculty members from their work with the class (Kerlinger, 1971, p. 354).

A recent and less common use of student ratings is to determine whether or not teachers possess the minimal competencies and behaviors expected of all faculty members. At Michigan State University, for example, the Academic Council approved a code of teaching responsibility that included seven specific provisions recommended for use in making salary, promotion, and tenure decisions. The council also approved a student rating form to reflect specific points in the code, including the following (Olson, 1977):

1. Were the instructional objectives stated either in writing or orally at the beginning of the term? Was instruction consistent with stated objectives?
2. Did the instructor explain course grading procedures either in writing or orally? Were the stated grading procedures followed?
3. Were graded materials returned to you soon enough to be useful in your learning?
4. Did the instructor meet your class at the scheduled or agreed upon time?
5. Did the instructor have scheduled office hours for consultation? Was the instructor or teaching assistant available during office hours?

6. Was the instructor available for prearranged appointments with you?

The vast majority of faculty members would receive positive responses to these and other items on the form. For the few who might not, the information obtained would be useful in making administrative decisions and in enforcing a reasonable teaching code.

Regardless of their purpose, student rating forms should be succinct. Ten to fifteen minutes should be the maximum time needed to complete a set of questions; anything longer strains student interest and tolerance and diminishes the quality of responses, especially if forms have been completed for several courses. As teachers, too, would resent too much class time being spent in this way, institutions might offer them a short form of no more than ten questions. A slightly longer form designed particularly for instructional improvement could be administered when teachers want more detailed information.

Are Student Ratings Reliable?

An instrument of poor reliability can be thought of as an elastic yardstick that would provide a different reading each time it was used. Student ratings are not elastic yardsticks. Their reliability or consistency, as indicated by numerous studies, is very good, providing enough students in a class have made ratings. For personnel decisions, it is also important to base judgments on several courses taught by a specific teacher.

Feldman (1977) discussed several procedures used by researchers to determine the reliability of student ratings. Each procedure estimates the extent of student agreement on ratings within a class or their internal consistency. The results are fairly similar, regardless of the specific procedure. One method draws pairs of students at random from a course and correlates their ratings, with higher correlations indicating greater consistency among student respondents. A similar method computes the mean scores for random halves of a class and then correlates these means across a number of classes. A third and more frequently used method computes the intraclass correlation coefficient (Winer, 1962). This index compares the variation within

classes with the variation across classes to provide an estimate of the relative homogeneity of ratings. A low reliability estimate, according to this method, usually indicates a wide variation in ratings among students in a typical class. Ideally, it is desirable to have more differentiation among mean scores for teachers than among individual student responses in each class. As with the other procedures, this does not measure the extent to which students give exactly the same rating.

Applying the intraclass correlation procedure, I calculated reliabilities for each of thirty instructional rating items and for varying numbers of student raters (Centra, 1973a). With ten raters, the reliability coefficients were about .70 for most items and .78 for ratings of a teacher's overall effectiveness. The estimated reliability for fifteen student raters was above .80 for most of the items; for twenty raters, the reliabilities were close to .90. These average reliabilities are similar to those reported in other studies.

Given this information, how many students are needed to provide a sound basis for a reliable average rating? The answer depends on how the rating results are to be used. For instructional improvements, average ratings based on as few as eight or ten students could provide the instructor with useful information, but larger numbers are preferable. To evaluate a teacher's instructional effectiveness for a promotional decision, both the number of students rating each course and the number of courses to be considered are critical. A study by Gilmore, Kane, and Naccarato (1978) shows that the use of ratings from five or more courses in which some fifteen students responded will result in a "dependable" assessment of teaching effectiveness; they found that ratings from more than five courses are required for dependability if teachers have taught relatively small sections (fewer than fifteen students). If as many as ten courses are considered, then the number of student raters for the various courses makes little difference. If the ratings of only one or two courses are considered, however, the researchers concluded that the results should not be used as a measure of teaching effectiveness for personnel decisions, regardless of the number of raters.

The proportion of a class that rates an instructor is as important as the number of raters. If only twenty out of sixty students in a class respond to a rating form, it is possible that they do not represent the reactions of the entire class (unless raters are selected on a random

basis). Even with responses from two thirds of the enrolled students in a course—the minimum desirable proportion—there could be some response bias. Using the evaluations of a sufficient number of students, however, will reduce the effects of a few divergent raters.

The stability of student responses has also been investigated to determine the influence of student moods and other invalid effects. Rating forms given twice to the same students over a short period of time produced fairly stable results (Costin, 1968; Centra, 1972); the mean ratings for 296 teachers collected about five weeks apart, for example, correlated an average of .70 for twenty-three rating items. Ratings collected a year apart from the same students also correlated significantly, though the later ratings tended to rate the teacher as less effective than those collected at the end of the course (Overall and Marsh, 1978).

Do Student and Class Characteristics Affect Student Ratings?

The fact that student ratings are reasonably reliable and stable when based on responses from a sufficiently large sample of students does not guarantee that they are immune from contamination. Whether the ratings actually assess teaching behavior or reflect characteristics of (1) a course over which the teacher has little control or (2) of the raters themselves are critical questions—especially when ratings are used administratively. Student characteristics that could affect ratings include: age, sex, college year (freshman, sophomore, and so on), academic ability, grade point average, expected grade in the course, reason for taking the course, and personality differences. Course characteristics that could have an effect include: type of course requirement (major requirement, general college requirement, or elective), subject matter area, class size, and method of instruction. The relationship between these factors and student ratings has been investigated in hundreds of studies, several of which are reviewed by Costin, Greenough, and Menges (1971), Kulik and McKeachie (1975), and Doyle (1975).

For a closer look at the possible effects of student and course factors on ratings, we drew on responses from over 300,000 students and approximately 16,000 classes at over 100 colleges that had used

the Student Instructional Report (Centra and Creech, 1976). To facilitate computations, a random sample of approximately 15,000 students and 9,000 classes was selected and analyzed, focusing on the following five-point global question: Compared to other instructors you have had (secondary school and college), how effective has the instructor been in this course?

- One of the most effective (among the top 10 percent)
- More effective than most (among the top 30 percent)
- About average
- Not as effective as most (in the lowest 30 percent)
- One of the least effective (in the lowest 10 percent)

Findings from this survey were similar to those from many other studies: the relationships between student or course characteristics and student ratings were generally insignificant or small enough not to have any practical significance. For several factors, however—class size, subject area of the course, and (occasionally) the course in relation to the students' curriculum—the correlations with ratings were high enough to recommend that they be considered in interpreting ratings.

Student characteristics having weak or insignificant relationships with ratings of teacher effectiveness were: sex, grade point average, college year, academic ability, and age. For example, the mean ratings given by females and males were almost identical: 3.74 and 3.73. But this is not to say that teachers who direct their teaching toward students of a particular sex or ability level will not be rated differently by those groups, as Elliot (1950) suggests. He finds a positive correlation between grade point averages and teacher ratings only when the teacher taught to the better students in the class. Another study (Centra and Linn, 1976) analyzed separately the student ratings within each of three large classes. Subgroups of students in each of the three classes rated differently such things as course examinations, class discussions, and assignments, but in only one of the courses (titled [significantly?] "Social Inequality") did such student characteristics as grades, gender, and college year differentiate the subgroups. Similarly, Yonge and Sassenrath (1968) investigated the relationship between student personality factors and ratings for each

of three instructors; they found personality factors to be somewhat related to ratings, but the relationships were not consistent for all three instructors.

These analyses within classes, as well as the numerous studies based on data pooled across classes, indicate that student characteristics, although not systematically affecting the ratings given, may on occasion have a significant effect. Teachers who use ratings for self-improvement thus may find it useful to look beyond the average ratings from the entire class and inspect the responses of identifiable subgroups of students, such as those with high or low grade point averages. By examining subgroups' responses, instructors may discover a rating pattern that suggests that one segment of the class is being slighted. If so, some adjustments in teaching methods or course design may be needed.

In personnel decisions, ratings should span a number of courses taught by a particular instructor in order to reduce the likelihood that the ratings have been systematically biased for or against the instructor. Providing norms or comparison information for similar courses can be equally valuable. This is especially important for factors, including the following four, that are more highly related to ratings.

Class Size. Very small classes—those with less than ten or fifteen students—are most highly rated. Those with fewer than fifteen students clearly received the highest ratings, followed by those with sixteen to thirty-five students and those with over a hundred students. The lowest ratings are found in classes with thirty-five to a hundred students (Centra and Creech, 1976). Classes of more than a hundred students may receive higher ratings because colleges or departments assign their best teachers and resources to such large classes and because the teachers themselves may prepare more thoroughly for particularly large groups of students than they would for smaller ones. Classes with thirty-five to a hundred students may not receive such attention and may also be too large to facilitate teacher-student interaction. Smaller classes, especially those with fewer than fifteen students and thus with fewer demands and less student variation, allow more questions to be posed and answered and enable teachers to adjust material more closely to student needs than do larger classes. For the same reasons, discussion courses are generally

rated higher in both course value and teacher effectiveness than are lecture courses.

Subject Matter. In comparing thousands of classes in each of the fields of study, slightly higher student ratings of course value and teacher effectiveness are found in the fields of the humanities than in the social sciences and the natural sciences (Centra and Creech, 1976; Educational Testing Service, 1975, 1977). Why this general pattern exists is not clear, but it may be due to the relative importance accorded to teaching and to research by instructors in each of the different fields. In one study (Parsons and Platt, 1968), teachers in the natural sciences judged research to be three times as important as teaching and social science teachers saw research as being "four thirds" as important as teaching; only in the humanities did instructors see research and teaching as equal in importance.

Students in more than four hundred classes from five colleges rated courses in the natural sciences as faster paced, more difficult, and less stimulating than those in the humanities, social sciences, and education (Centra, 1972). They reported natural science teachers to be less open to new viewpoints than teachers in other disciplines. Humanities teachers, compared to those in the other three areas, were rated as less likely to inform students of evaluation methods and less likely to teach toward announced objectives.

Type of Course Requirement. Students give slightly higher ratings to their majors or electives than to courses taken to fulfill a college requirement (Centra and Creech, 1976). Their motivation and their personal interest in their major courses and in subjects they have chosen to study would lead them to rate the courses as more valuable and effective. In addition, some teachers have less interest in lower-level, college-required courses and thus put less effort into their teaching.

Grade Expected. Of major concern in rating programs are the influence of students' grades on their ratings and the possibility that students will reward easy-grading teachers with higher ratings. There seems to be no overriding evidence that students rate an instructor more favorably or unfavorably on the basis of the grades they anticipate receiving—although there may be occasions when that occurs, as shown by Holmes (1972), who found that students give the instructor lower ratings if their actual grades are lower than those they had

expected. Similarly, they give lower ratings when their expected grade is lower than the grades they have received in other courses, as indicated by their cumulative grade point average (Bausell and Magoon, 1972; Centra and Creech, 1976).

The correlation between grades and ratings is usually in the .20 range. In one study (Centra and Creech, 1976), students expecting an A grade gave teachers an average rating of 3.95 (on a five-point scale), while those expecting a C grade gave them an average rating of 3.41. The correlation between expected grade and student ratings of the course value are generally a little higher than the correlation between expected grade and teacher rating. Both correlations are based on data pooled across classes, but evidence indicates that the same relationships would hold for analyses within classes as well (Centra and Linn, 1976).

One way to interpret the association between expected grade and ratings is to view it as partial evidence for validity: If a grade or an expected grade reflects how much a student knows about the subject matter at the end of the course, then there should be some relationship between that grade and the student's ratings of the teacher and of the course. In part, then, the relationship can be viewed as modest evidence that students rate higher those courses in which they learn more.

Are Ratings Affected by Teacher Characteristics?

Whether student ratings reflect characteristics of instructors that should have no effect on their teaching effectiveness is a major concern. (We would hope, for example, that students would not rate full professors highly simply because of their rank or status on campus.) Research evidence indicates that teacher characteristics are generally not related to the ratings they receive. The one exception is the number of years of teaching experience, but the pattern of ratings for teachers with varying years of experience is clearly explainable and probably does not reflect bias.

Analysis of the ratings of overall teaching effectiveness for more than 8,000 teachers with varying years of experience shows that those in their first year of teaching generally receive the poorest ratings (average 3.54 on a five-point scale [Centra and Creech, 1976]). Teachers with one or two years of experience and those with more

than twelve years receive similar ratings, an average of about 3.75. Slightly higher are teachers in the three- to twelve-year range, with an average of 3.83. First-year teachers are usually learning on the job; most of them have had little formal training in graduate school on the process of teaching. In using ratings for administrative purposes, then, it should be recognized that first-year teachers may improve considerably with experience. While instructors with very poor ratings may not become exceptional teachers, there could be critical changes for first-year teachers who receive only average ratings.

The slight decline in rated effectiveness in the later years of a teaching career (there is no significance in the twelfth year per se; my recent analysis with a new sample indicates that teachers with over twenty years of experience received even lower ratings on the average) has implications for teaching improvement programs. Some teachers acquire substantial administrative or research responsibilities in their later years, along with a decline in teaching involvement; others become bored and indifferent. Faculty development programs, therefore, need to be concerned with revitalizing older teachers and with assisting those just entering the profession. Highet (1976) points out that changes occur over time in the subject matter within a discipline, as do teachers' relationships with their students. He believes that many teachers in their later years assume too much knowledge on the part of the young, while their own grasp of the subject matter has become more automatic. Eble (1971) provides an excellent discussion of career development for faculty members in mid-career and later.

Numerous studies have correlated other teacher characteristics with ratings, such as academic rank, sex, teaching load, and research productivity. None of these, however, are significantly and consistently related to ratings.

Academic Rank. The mean scores for more than 8,000 teachers at four academic ranks (instructor through professor) are virtually identical (Centra and Creech, 1976). Only teaching assistants received significantly lower scores—probably due to their limited teaching experience, as discussed earlier.

Sex. Male and female teachers are occasionally rated differently, but the differences do not have much practical significance. Students in one study (Centra, 1972) rated women teachers higher than men teachers on items dealing with teacher-student interaction; however,

they found courses taught by men more stimulating. A few studies reported ratings to be slightly higher when teacher and student gender are the same (Ferber and Huber, 1975; Elmore and LaPointe, 1975), but even these small differences are inconsistent and may depend on the particular course (Wilson and Doyle, 1976).

Teaching Load. One might expect that faculty members with the heaviest teaching loads would receive lower student ratings because of less time for preparation and other teaching-related activities. Yet the opposite is true. Analysis of ratings for the more than 8,000 teachers studied by Centra and Creech (1976) indicates that teachers with a credit-hour load of thirteen or more were given the highest ratings. There is little difference in the ratings of teachers with less than a thirteen-hour teaching load. Faculty with loads of thirteen or more hours are generally located at two-year or four-year colleges where teaching is the major faculty activity. Indeed, the ratings of teachers at two-year colleges are slightly higher than for teachers at four-year colleges and universities (Educational Testing Service, 1975, 1977). For teachers at the same college or at the same type of college, teaching load would probably have little effect on ratings and need not be considered in their interpretation. Some colleges, however, do take teaching load into account in determining faculty rewards and promotions (see Chapter Eight).

Research Productivity. Research and writing help to keep teachers refreshed and on top of their fields, a good reason to expect a positive relationship between scholarly productivity and teaching effectiveness as assessed by students. A few studies support this expectation (see Stallings and Singhal, 1970). Several other studies, however, report no association between research productivity—as reflected by the numbers of books and articles published—and student ratings of teaching effectiveness (see Guthrie, 1954; Voeks, 1962; Aleamoni and Yimer, 1973). Publications apparently are not essential for good teaching; therefore the use of publication counts in teacher evaluation will not reflect teaching performance as judged by students. Publications also appear to be unrelated to colleague ratings of teaching (Aleamoni and Yimer, 1973).

Are Teachers Who "Entertain" Rather than "Teach" Rated Highly by Students?

This question, frequently raised by faculty members, raises an important issue: what students perceive as good teaching. A study by Naftulin, Ware, and Donnelly (1973) tested the entertainment question by employing a professional actor to deliver a graduate-level lecture that was nonsubstantive and contradictory in content. Dr. Fox, as the actor was called, was a very entertaining and dynamic lecturer. The high ratings that he received, the researchers reasoned, supported their contention that "Given a sufficiently impressive lecture paradigm, an experienced group of educators participating in a new learning situation can feel satisfied that they have learned despite irrelevant, conflicting, and meaningless content conveyed by the lecturer" (1973, p. 634).

The students gave Dr. Fox high ratings in organization, stimulation, and interest in the subject—probably accurate reflections of the actor's performance. It was the content of the lecture that was faulty. In a sense, then, the Dr. Fox results underscore findings from other studies showing "lecturing ability" to be an important part of teaching effectiveness as rated by students. Many of them rate a good teacher as one who has a great deal of interest and enthusiasm in the subject, organizes the material well, and is stimulating in presentation. So is a good entertainer. To some extent, then, teaching and entertainment do overlap, at least in lecture presentations. But, as Guthrie (1954) found, highly rated teachers tend to be "substance teachers" and not merely good entertainers. The Costin, Greenough, and Menges (1971) conclusion would still seem applicable: sheer "entertainment" is not what most students see as good teaching.

The Dr. Fox study has implications as to who should evaluate a teacher's knowledge of subject matter. Given the fact that the study was based on only one lecture, it is conceivable that the students would eventually have rated the content as poor. Certainly it is much easier to delude a group of students for one session than for an entire semester. Even so, the results support the view that colleagues are more appropriate than students as judges of a teacher's subject knowledge.

Do Students Learn More from Teachers Whom They Rate Highly?

Global ratings of teacher effectiveness and course value correlate more highly with student learning than do the ratings of such specific instructional practices as teacher–student interaction. Global ratings may be more valid estimates of student learning because they are not tied to a specific instructional style. The research results also suggest that some instructional practices work well for some teachers but not all. Close student relationships are not needed by all teachers, for example, to facilitate learning in their courses. For some it is part of their teaching style, and it may well contribute to their effectiveness as measured by ratings or student achievement. Other practices may account for the effectiveness of other teachers. This means that in using ratings in personnel decisions, global ratings could be more defensible than ratings of specific practices. Although global ratings and achievement are generally correlated highly for most courses in the studies, the exceptions underscore the need to supplement ratings with additional criteria of teaching effectiveness.

Are highly rated teachers those from whom students learn most? This question focuses on the critical issue of what student ratings actually mean. It has spawned numerous validity studies that employ the so-called criterion-related approach to validity: the amount that students have learned at the end of a course is the criterion of good teaching—an argument with strong support (for example, Cohen and Brawer, 1969; Rose, 1976). Student ratings, hopefully, will be at least moderately related to learning. Multisection courses with common final examinations in each are used for these studies; for each section, mean student ratings are correlated with mean final exam performance at the end of the course. Students generally select their sections or teachers rather than being assigned at random, which requires some kind of statistical adjustment to compensate for initial differences in student ability or achievement. For example, Elliot (1950) adjusted for academic aptitude and found moderate correlations between the adjusted achievement scores and ratings of some aspects of instruction. Cohen and Berger (1970), Morsh, Burgess, and Smith (1956), McKeachie, Lin, and Mann (1971), and Doyle and Whitely (1974) also report moderate correlations between ratings and learning, but again students were not assigned to teachers on a random basis.

A high negative correlation between learning and ratings is reported by Rodin and Rodin (1972), who found that students actually learned *less* from teachers whom they rated highly. Because of this unusual result—and probably because of its publication in *Science*—the study received a great deal of undeserved attention. The sample studied consisted of twelve teaching assistants who were teaching in only a peripheral sense: they met with their classes two days a week (the professor lectured to the entire group on the other three), primarily to help those students who needed aid in solving assigned calculus problems. Since the Rodins used as their criterion the number of calculus problems done correctly by students at the end of the term, the negative relationship with ratings is understandable. It is likely that students who had little need of help not only obtained the best grades but also rated the teaching assistant lowest in teaching performance (they may, in fact, have skipped many class sessions). Students most in need of help solved the fewest number of problems and probably rated their teachers highest. In other words, aid provided by the teaching assistant—or teaching performance—may have had little to do with the number of problems that students completed.

A proper study includes teachers in more typical multisection instructional settings and also includes a random assignment of students to each class section. Randomization helps to ensure that differences in final exam scores will be due to teacher effectiveness rather than to differences in student motivation. Highly motivated students might do better than was predicted by a pretest and might seek out teachers with good reputations and rate them higher regardless of teaching performance (Leventhal, 1975). In short, randomization of students is one of the steps needed to draw a cause and effect relationship between rated teacher effectiveness and student learning.

In two studies (Sullivan and Skanes, 1974; Centra, 1977b), students were assigned at random to multisection courses in which a common final examination was used. The Centra study also included some courses in which randomization was not used. Both studies took place at Memorial University in Newfoundland and together included analyses of 202 sections in seventeen courses—the subject areas being

first-year chemistry, biology, mathematics, physics, and psychology. Sullivan and Skanes report that the average correlation between ratings of teacher competence and student learning as measured by the final exam was modest but significant: .39. The relationship, however, was much higher for full-time instructors (.53) than for teaching assistants (.01), and for experienced teachers (.69) versus those in their first year of teaching (.13). Inexperienced teachers, Sullivan and Skanes reason, have not yet developed a consistent teaching style, thereby contributing to the low validity correlation.

The Centra analysis finds a significant relationship between ratings of teacher effectiveness and student achievement: half of the correlations are .60 or higher and all but one are positive. Student ratings of the course value show similar results. Ratings of course objectives, of organization, and of the quality of lectures correlate fairly well with achievement. Ratings of the teacher-student relationship, of course examinations, and of student effort do not correlate strongly with achievement: the median correlation is .30. The weakest or most inconsistent correlations with achievement are for ratings of teaching assignments and for course difficulty and workload.

The relationships between ratings and student achievement are significant, but they might have been higher if the range for both types of variables was greater. The restricted range of achievement scores between the sections and the limited variability in mean ratings across instructors, evident in both studies, suppress the correlations.

Do Student Ratings Improve Instruction?

There is a good deal of skepticism regarding the effect of student ratings on changes or improvements in instruction—particularly when the results are seen only by the individual teacher. It is assumed that teachers value student opinion enough to alter their instructional practices when needed. But do they? Although the ratings that individual teachers receive often improve over time, it cannot be assumed that the initial ratings caused that improvement; additional teaching experience by itself often results in instructional changes. Student ratings may lead to some changes when only the teachers see the results, but there are probably many ways to increase their impact.

To investigate the effects of student ratings on a teacher's practices requires an experimental design in which random groups of teachers receive feedback from students while other teachers—those in control groups—do not. Such a study (Centra, 1973b) involved more than 400 teachers at five different types of college. In every department, teachers were randomly assigned to one of three groups: (1) *The feedback group*, in which teachers administered the 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, the treatment in this instance being what is done at most colleges using student ratings for instructional improvement—the results are seen only by the instructor. (2) *The no-feedback group* used the rating form at midsemester but saw the summary of results at the end of the semester. This is the "control" group. (3) *The post-test group*, in which the rating form was used only at the end of the semester to determine whether midsemester ratings had a sensitizing effect on teachers in the no-feedback group—whether their use of the form resulted in changes even though they had not had any feedback.

In addition to using the form at midsemester, teachers in the feedback and no-feedback groups administered the form at the end of the semester. Both sets of ratings were collected during the single 1971 fall semester, so that the same students could provide both sets of ratings. Teachers were also asked at midsemester to rate their own instructional practices on a self-rating form that paralleled the student rating form.

The major conclusion of the study is that, for instructors whose self-evaluations were considerably better than were their student ratings, changes in instruction (as assessed by repeated student evaluations) occur after only a half semester. If, in other words, teachers are "unrealistic" in observing their teaching—unrealistic relative to their students' view, that is—then they tend to make changes in their instructional practices. A second finding is that a wider variety of instructors change if given more than a half semester of time and if they have information to help them interpret their scores. These changes are most evident in their preparation for class, use of class time, summarization of major points in lectures or discussions,

openness to other viewponts, and making helpful comments on papers or exams.

The reason to relate changes in teaching procedures to the discrepancy between self-evaluation and student ratings can be found in social-psychological theory—in particular, in equilibrium or self-consistency theory, the central notion of which is that an individual's actions are strongly influenced by self-evaluation. Thus, when student ratings are much poorer than an instructor's self-rating, a condition of imbalance (Heider, 1958) or dissonance (Festinger, 1957) is created in the instructor. In an attempt to become more consistent—or, in theoretical terms, to restore a condition of equilibrium—the instructor changes in the direction suggested by the student ratings.

These theories assume that most teachers value collective student ratings and that they know how to make changes. The study results indicate that to some extent this occurs. Undoubtedly, however, some teachers write off student judgment as unreliable or unworthy, and for these individuals changes are unlikely even though they may be needed. Other types of evaluation or analysis may be more effective for these teachers, such as colleague reactions or use of in-class videotapes.

Still other means of treating student rating responses might have greater impact. If instructors rate themselves on most items, it is possible to produce a discrepancy score between student and self-ratings that can highlight aspects of instruction in special need of attention. Publicizing ratings may also draw increased attention to them. Response summaries in the five colleges studied (Centra, 1973b) were seen only by the individual teacher; publicized ratings might have pressured more teachers to change their methods. Including ratings in personnel decisions will also increase their importance, though some instructors may not know how to go about improving their teaching methods. It is therefore possible to use ratings in a counseling situation as part of a faculty or instructional development program. According to some research evidence (Aleamoni, 1974), accompanying rating results with some kind of counsel—such as that of a faculty development specialist or a master teacher—is effective.

The impact of ratings can be increased by evaluating continuously rather than only at the end of the course. Parent and others (1971) describe a program tried at the University of Minnesota. The

instructor plans the course content and instructional methods on what he or she has learned of student expectations at the start of the course, their prior preparation, and personal descriptive information. A course and instructional rating form at midterm elicits student reactions to course content and organization, the adequacy of methods used, the value of texts and assignments, and the like. In addition, six to ten students act as ombudsmen, funneling the reactions of other students to the teacher. According to the authors, this procedure has the advantages of involving students directly in course development, of providing teachers with information that allows them to adapt course content to the enrolled students, and of promoting good relationships between students and faculty—thus increasing student motivation and learning experiences.

Do Alumni Rate Teachers Differently?

It is frequently said that student ratings do not adequately reflect the long-term effects of instruction. Student immaturity or lack of perspective are often blamed for this shortcoming, and it is assumed that later ratings—say, when the students are alumni—are more valid measures of teacher effectiveness. There is the example of the hard-driving, demanding teacher supposedly not appreciated by students until they have gained more real-life experience. Though there may be times when this is indeed the case, research suggests that it is rare; most teachers rated poorly by students are also rated poorly by alumni.

Because of the general agreement between student and alumni ratings, there seems to be little need to use the latter in faculty tenure and promotion decisions. Ratings by current students provide similar information and are much easier to collect. Alumni ratings are also less useful in instructional improvement since alumni may not be able to recall the kind of specific information needed by teachers. Alumni, on the other hand, may be able to provide useful reactions to the relevance of courses in the curriculum and to other college experiences, reactions which could be useful in modifying department offerings.

Three studies demonstrate the similarity between student and alumni ratings of instruction. Drucker and Remmers (1951) surveyed

graduates who had been out of college for ten or more years and found positive correlations between ratings given to seventeen instructors by students and alumni. Correlations ranged from .40 to .68 on ten such teacher traits as the presentation of student matter, interest in the subject, sympathetic attitude toward students, and fairness in grading. Had these correlations been based on an overall assessment of teaching rather than of specific traits and had the length of time away from the courses not been as great, they might have been higher. Student ratings of overall teacher effectiveness compared with similar ratings by alumni who had been out no more than five years (Centra, 1974) showed agreement between the two. In this study, some 500 alumni named the best and worst teachers they had had in college. The rank correlation for student and alumni responses based on the ratings for twenty-three teachers was .75. The agreement between current studies and alumni regarding specific effective or ineffective teachers thus was substantial, particularly at the extremes: very good and very poor teachers were identified as such by both students and alumni.

In some instances, however, a teacher was seen as both "good" and "bad" by the same groups; the same teacher was occasionally nominated as one of the best teachers by some alumni and as one of the worst by other alumni. Obviously, some teachers have a special appeal or lack of appeal to specific kinds of students. Other researchers indicate that it is not enough to speak just of "good" or "bad" teachers; one might also ask "good" or "bad" for which students (McKeachie, Lin, and Mann, 1971; Dowaliby and Schumer, 1973). While this admonition would seem justified for some of the teachers in the Centra (1974) study of alumni ratings, most teachers in the sample appeared to be effective with a wide variety of students—at least, as measured by a single overall rating provided by alumni and students.

A third study of long-term effects of instruction (Overall and Marsh, 1978) compares responses from alumni who had rated a senior-level course with their ratings of the course and instructor a year after graduation. Although the alumni ratings tend to be lower than ratings given at the end of the course, the two sets of ratings are consistent in that they correlate significantly. Unlike the previous two studies, the same individuals rated the teacher at both times.

Are Student Ratings Affected by Their Intended Use?

Because of the serious implications for the teacher, one might expect students to rate the teacher more leniently if the results are to be considered in salary, tenure, or promotion deliberations. On the other hand, they might be more frank and perhaps more severe in their ratings and criticism if the results are to be used for course or instructional improvement; such information, they might logically assume, could lead to needed changes.

Ratings collected in forty-one classes at a midwestern university showed that students tend to rate teachers similarly under both circumstances (Centra, 1976b). Random halves of each of the classes were given different written directions regarding the intended use of the results—i.e., "will be used in *salary, promotion or tenure* considerations for this teacher," or "will be used *only by the instructor* to evaluate and improve his or her teaching." A comparison of responses indicated that only in a few instances did students give more favorable ratings when they understood that the results would be used administratively. Later studies reported similar results, so the findings seem to be valid.

Although the small differences in ratings indicate that students are generally not influenced by written instructions, it may well be that oral directions given by a teacher, especially if given with a subtle appeal to generosity, could have a sizable effect. A study by Fentress and Swanson (1973) found that instructors got higher ratings by giving reinforcement and praise to the class at the time the forms were administered. They also found that the teaching assistants who participated in the study could influence their ratings favorably by combining praise of the class with outright reinforcement in the form of soft drinks, pretzels, and potato chips.

Should Ratings Be Anonymous?

To use ratings for tenure, promotion, or salary considerations, it is advisable to establish standardized procedures for the administration of forms, procedures such as the requirement that a student or some type of proctor distribute, collect, and place the questionnaires

in a sealed envelope and that the teachers not be present during the administration of the questionnaires. It is advisable also to keep rating forms anonymous in all cases, thus ensuring that a student cannot be penalized for giving low ratings.

Some teachers argue that signed rating forms motivate students to give thoughtful responses and allow teachers to obtain detailed information from selected students regarding critical responses. And, if ratings are used for administrative purposes, they argue that they should know their accusers—a specious argument, since only the average class ratings are reliable and not individual student responses. Although students who identify themselves are expected to be far more generous in their ratings, especially if the forms are distributed and returned to the instructor prior to final grading, evidence does not totally support this expectation. One study finds that students who identify themselves rate their teachers no differently than those who remain anonymous (Stone, Rabinowitz, and Spool, 1977), while another study by the same researchers finds the expected higher ratings by students who completed signed rating forms (Stone, Spool, and Rabinowitz, 1977).

What Are the Limitations on Student Ratings?

Limitations of student rating programs other than those already given in this chapter should be considered:

1. Because most student rating instruments elicit numerical responses that can be scored and quantified, it is easy to assign them a precision they do not possess. In a discussion of standardized tests, Turnbull (1978) terms this tendency the "micrometer fallacy." Decision makers therefore should guard against overinterpreting small variations between teachers; there is little practical difference between a teacher whose mean rating is at the 60th percentile and another who is at the 65th percentile. Another fault is giving student ratings too much weight in relation to other criteria. Because they can be quantified, the temptation to assign them undue importance is understandable.

2. The manipulations of ratings by teachers must be considered when ratings are used for personnel decisions. At issue is whether

teachers can influence ratings but not student learning. Teachers who argue, as one did, that they can improve their ratings by inviting students to their homes for informal discussion accompanied by refreshments can also be improving student learning and motivation as well. Their attempt to improve ratings is, in this instance, also a good educational practice. But the teacher who is lenient in assigning grades and out-of-class work is not improving learning, yet may be better rated by some students. The extent to which lenient grading consistently causes higher ratings is still in question. As one safeguard, a teacher's grade distribution ought to be examined; in particular, the course grades for students in the class should be compared with their average grades in other courses. An inflation-proof grading system along these lines has been proposed at the University of California, Berkeley (Carnegie Council on Policy Studies in Higher Education, 1979).

3. Some institutions point to their student rating system as proof that they are concerned about improving teaching yet do little else to help teachers develop their skills. In short, student ratings have misled some institutions into thinking that nothing more is needed to upgrade instruction. While some teachers can use the rating information to make needed changes, others need faculty and instructional development services.

4. Because of the positive bias in student ratings, teachers who need to improve may not realize their weaknesses. Providing comparative data is one way to minimize misleading interpretations.

Recommendations

1. For personnel decisions, ratings of a teacher across courses should be considered, the minimum number of courses depending on the number of student raters in each course. In general, five or more courses are needed for a dependable assessment if at least fifteen students have rated each course.
2. Do not overuse student ratings. Students will get bored and will respond haphazardly or not at all. Use can be limited by recommending that tenured staff collect ratings in only one course each year and in new courses; nontenured staff could collect ratings in their different courses but not in every section.

3. A rating form should not be excessively long. Ten minutes to complete a form is all most students will want to spend, and teachers are often reluctant to use up too much class time.
4. For personnel decisions, items that rate the overall effectiveness of the teacher and course (global items) are especially useful. Other items might be used in making judgments if they reflect a teaching code that all teachers are expected to fulfill.
5. If a common set of rating items are adopted or developed by a college, teachers and departments should have the option of adding their own specific items. Written comments by students should also be encouraged for instructional and course improvement.
6. Decision makers and teachers need to be aware of possible influences on rating responses due to specific characteristics of the course, of students, or of teachers—characteristics that have little to do with actual teacher or course effectiveness. Most extraneous variables have a relatively weak relationship to ratings. But a few—small classes, for example—seem to get higher ratings and are generally advantageous. Such characteristics should be considered in interpreting results.
7. Standardized procedures in administering forms are recommended if the results are to be used in personnel decisions. One method is to have a student or someone other than the teacher distribute, collect, and place the questionnaires in a sealed envelope, the teacher not being present during the process. The timing—preferably during the last week or two of class—should also be standard. Mailing the forms to students usually results in a poor response rate.

THREE

Benefits of Self-Assessment and Self-Analysis

Teacher self-evaluation or self-reports are generally minor factors in tenure and promotion evaluations. The department chairmen surveyed by Centra (1977a) ranked self-evaluation as ninth among fifteen criteria, and few department chairmen gave them more importance than that. Other studies concur. According to Larson (1970), fewer than 10 percent of the English department chairmen surveyed by him collected written self-evaluations. Seldin (1978) reports that about 20 percent of the liberal arts college deans in his 1973 study used self-evaluation, but by 1978 the number reached 36 percent, indicating that the use of self-assessment or self-reports in tenure and promotion decisions is increasing among some colleges.

Senate office

from:
AAUP Bulletin, Academe, Vol. 65, No. 6
(October, 1979)

Student Ratings of Faculty: A Reprise

Wilbert J. McKeachie

IN 1969 I was commissioned by the AAUP Committee C on College and University Teaching, Research, and Publication to write an article on student ratings of faculty for the *AAUP Bulletin*.[1]

In the ten years since that article appeared, a great deal of research has been done, and colleges and universities have accumulated much experience with student ratings. In addition, Committee C, of which I was a member, prepared a "Statement on Teaching Evaluation" adopted by the AAUP Council and approved at the Annual Meeting in June, 1975. The Project to Improve College Teaching, jointly sponsored by AAUP and the Association of American Colleges, also published, in 1971, *The Recognition and Evaluation of Teaching* by Kenneth E. Eble. The purpose of this article is to bring the reader up to date on the evidence with respect to the issues discussed earlier, as well as covering additional issues that have come to the fore more recently.

VALIDITY: DO STUDENT RATINGS MEASURE TEACHING EFFECTIVENESS?

There is now a good deal of evidence supporting a positive answer to our question, but it has also become evident that the question is overly simple. With respect to the general validity question, the cumulating evidence continues to support the conclusion that highly rated teachers tend to be those whose students achieve well. Such a statement is, however, better understood in the context of two more analytic questions.

1. How are different aspects of teaching effectiveness related to student ratings? Or put in other words, "What educational outcomes are related to student ratings of effectiveness?" Ratings may be differentially valid for different educational goals. There is ample evidence that we achieve some goals at the expense of others. Teachers effective in teaching a good deal of knowledge are not necessarily effective in teaching critical thinking. So one needs to make value judgments about the importance of differing goals of education.

2. What is the intended use of student ratings of teaching? For personnel decisions? For improving teaching? For facilitating student choice of courses and teachers? Ratings may be differentially valid for different uses.

For personnel decisions we want student ratings to be valid measures of teaching effectiveness. For improving teaching we want student ratings to be valid in terms of accurate diagnosis of problems and, perhaps, for prescription of solutions. For guiding student choices of courses we want student ratings to provide information valid for enabling students to choose more valuable educational experiences. Let us consider these purposes in turn.

WILBERT J. MCKEACHIE is Professor of Psychology and Director of the Center for Research on Learning and Teaching at the University of Michigan.

This article was commissioned by the Association's Committee C on College and University Teaching, Research, and Publication. The author gratefully acknowledges useful suggestions from members of Committee C and from colleagues at the Center for Research on Learning and Teaching.

[1] W. J. McKeachie, "Student Ratings of Faculty," *AAUP Bulletin*, 55 (December, 1969), pp. 439–44.

J2 65

What Do We Mean by Teaching Effectiveness?

Obviously if we are to answer the question "Are student ratings valid measures of teaching effectiveness?" we need to define "teaching effectiveness." Simply put, we take teaching effectiveness to be the degree to which one has facilitated student achievement of educational goals. But assessing teaching effectiveness is not simple. Much of student achievement is determined by factors other than teaching; for example, student ability or previous experience. Moreover, student achievement in different courses is not comparable since there is no way of estimating how many units of mathematics achievement equal a given number of units of achievement in English.

One might expect that one could simply judge effectiveness in terms of how well students achieve course goals. But such "criterion referenced" measurement ultimately rests upon a judgment about what sort of achievement it is reasonable to expect, and what is reasonable to expect depends upon knowledge of what other teachers have done with similar classes. Thus, to validate a measure of teaching effectiveness, such as student ratings, we must have a number of teachers teaching the same course to comparable groups of students. Only in such a situation can we determine whether those teachers whose students learn the most are rated highest by their students. Let us now examine the evidence from such situations with respect to teaching effectiveness for different educational goals.

Validity of Student Ratings as a Measure of Teaching Effectiveness in Achieving Cognitive Goals

Most college professors who have thought about their goals describe both cognitive and motivational goals for their courses. We want students to make gains toward such cognitive goals as knowledge, skill in solving problems, and ability to evaluate. Typically we also want to achieve affective goals, such as increasing students' interest in the area studied, so that they will be motivated to continue learning after they leave college.

When we speak of student learning as the ultimate criterion of teaching effectiveness, we usually think of the cognitive outcomes. Frequently we assume that these outcomes are measured by the final examination for the course. In fact, however, final examinations typically weigh knowledge much more heavily than application, problem solving, or other cognitive objectives. Moreover, since students are strongly motivated for grades, they will do the best they can to pass the examination regardless of the quality of teaching they have had. If the teacher has been confusing or unhelpful, students often will make up for deficiencies by extra studying. Thus performance of students on a final examination is not an ideal measure of teaching effectiveness. Nevertheless, this is the best we have in most of our validity studies.

In my 1969 article, the most persuasive evidence cited for the validity of student ratings of instruction was the research of Elliott[2] and Russell[3] who demonstrated that student ratings of instruction were related to teaching effectiveness in terms of student achievement in multi-section courses in chemistry.

A substantial amount of research on validity has been carried out since 1969. The results are mixed, but taken as a whole they confirm our earlier conclusion that teachers rated as effective by students are generally those teachers whose students achieve most.[4]

The study most widely cited as evidence of lack of validity of student ratings is that of Rodin and Rodin[5] in which a high negative correlation between mean student ratings of effectiveness and mean performance of students on a mathematics test was found. Frey and others point out that both the methodology and generalizations of the Rodin findings with respect to other college courses are dubious. Even more impressive are replications of the study with better research designs by Frey,[6] Doyle and Whitely,[7] and by Marsh, Fleiner, and Thomas,[8] in which substantial positive correlations between mean

---

[2] D. H. Elliot, "Characteristics and Relationships of Various Criteria of Colleges and University Teaching" (Ph.D. dissertation, Purdue University, 1949).

[3] H. E. Russell, "Inter-relations of Some Indices of Instructor Effectiveness: An Exploratory Study" (Ph.D. dissertation, University of Pittsburgh, 1951).

[4] W. J. McKeachie and J. A. Kulik, "The Effectiveness of Instruction in Higher Education," and J. A. Kulik and W. J. McKeachie "The Evaluation of Teachers in Higher Education," in *Review of Research in Education*, vol. 3, edited by F. N. Kerlinger (Itasca, Illinois: Peacock, 1975).

[5] M. Rodin and B. Rodin, "Student Evaluations of Teachers," *Science*, 177 (1972), pp. 1164-66.

[6] P. W. Frey, "Student Ratings of Teaching: Validity of Several Rating Factors," *Science*, 182 (1973), pp. 83-85.

[7] K. O. Doyle, Jr. and S. E. Whitely, "Student Ratings as Criteria for Effective Teaching," *American Educational Research Journal*, 11 (1974), pp. 259-74.

[8] H. W. Marsh, H. Fleiner, and C. S. Thomas, "Validity and Usefulness of Student Evaluations of Instructional Quality," *Journal of Educational Pyschology*, 67 (1975), pp. 833-39.

student ratings and mean student performance were found, thus providing support for the validity of student ratings.

The studies by Sullivan and Skanes[9] and Centra[10] are the only studies in which students were randomly assigned to instructors. Such random assignment is an important feature in designing validity studies since differences in mean student achievement between teachers may otherwise be the result of differences between the students rather than differences in teaching.[11] Thus, the substantial positive relationships (.4 to .6) between mean student ratings and mean student achievement found in these two studies are particularly significant.

Of especial interest in the Centra study is the finding that global ratings of value of the course to the student tended to have higher validities than items assessing specific aspects of teaching. Ratings of the difficulty of the course, for example, had no significant relationship to student achievement. The Sullivan and Skanes study is also interesting in that greater validity was found for student ratings of regular faculty than for rating of teaching assistants. It may well be that when students feel that they have learned a good deal in a course they are more likely to attribute their success to their own efforts if their teacher was a teaching assistant and more likely to give the teacher some credit if the teacher is a professor.

I doubt that student ratings will ever account for the majority of the variance between classes in student cognitive achievement. Most student rating forms ask students to evaluate *teaching*, not their own *learning*. Even though Lathrop[12] demonstrated that students' perception of their own learning was a *major* factor in determining ratings of instructors and courses, it is not the only factor. Moreover, students are not always able to judge how effectively they have learned, particularly with respect to higher-level cognitive objectives. Students undoubtedly estimate their learning partially by their performance on course examinations. (A weak reed in many courses!) One important educational goal is that of helping students develop the ability to evaluate their own learning. If we are at all successful, we would expect greater validity for student ratings in advanced courses than in elementary courses. Unfortunately, validity studies require multi-section courses in order that the effectiveness of several teachers can be compared on common measures of student achievement. Such multi-section courses are most commonly found at the elementary level so that we cannot compare the validities of student ratings in elementary courses with those in advanced courses.

In any case, we should not expect student ratings of *teaching* to be perfectly related to achievement of any one of several course objectives. But equally important, and seemingly overlooked in bemoaning the modest size of most of these correlations, is the question, "How much of the differences between achievement of classes is produced by teaching differences?" Even in the best controlled studies it seems unlikely that teaching accounts for all the differences between classes. Consequently the correlations found between student ratings and achievement may be about as high as could be expected in accounting for differences produced by teachers.

The studies of the validity of student ratings are thus reasonably encouraging with respect to the goal of achievement on course examinations measuring cognitive goals. What of other educational outcomes?

Validity of Student Ratings as a Measure of Teaching Effectiveness in Achieving Attitudinal and Motivational Goals

One important goal of higher education is that of motivating students for continued learning—lifelong learning. In most courses one would not be happy if students mastered the content of the course but wanted never to learn anything more about that sort of material. Rather, we hope that a given course will be but a step on a path of lifelong learning. The ultimate measure of achievement of this goal is later learning behavior, such as buying books or magazines related to the discipline, later reading habits, attendance at lectures, workshops, or other opportunities for learning, election of further courses, etc. For end-of-the-course evidence it is hard to conceive of a measure with more face validity than answers

---

[9] A. M. Sullivan and G. R. Skanes, "Validity of Student Evaluation of Teaching and the Characteristics of Successful Instructors," *Journal of Educational Psychology*, 66 (1974), pp. 584-90.

[10] J. A. Centra, "Student Ratings of Instruction and Their Relationship to Student Learning," *American Educational Research Journal*, 14 (1977), pp. 17-24.

[11] L. Leventhal, P. C. Abrami, and R. P. Perry, "Do Teacher Rating Forms Reveal as Much about Students as about Teachers?" *Journal of Educational Psychology*, 68 (1976), pp. 441-45.

[12] R. G. Lathrop, "Unit Factorial Ratings by College Students of Courses and Instructors" (paper presented at the 1968 meeting of Western Psychological Association).

to such items as: "This course is increasing my interest in learning more about this area."

With respect to the goal of motivating further student learning, a small study by McKeachie and Solomon[13] found that students of highly rated introductory psychology teachers tended to elect more advanced courses in psychology. Sullivan and Skanes[14] reported that students of highly rated psychology teachers were more likely to major in psychology, but students of teachers who were effective in terms of student achievement also influenced course elections positively, even when the instructors did not receive high ratings.

In the attitudinal domain, Mann[15] found that students in classes of highly rated teachers developed more sophisticated attitudes about economics than students of less highly rated instructors.

Other Data Relevant to Questions of Validity of Student Ratings as Measures of Teaching Effectiveness

One common criticism of student ratings is the plaint, "You can't really appreciate good teachers until you have been out of college awhile. We hated 'Old So-and-So' when we were students, but now we know he really was a great teacher."

The evidence is that such cases are the exception rather than the rule. Drucker and Remmers[16] and Centra[17] found that alumni ratings of faculty correlate highly with those of current students. Aleamoni[18] and Marsh[19] report similar results in comparing current ratings with those by graduating seniors. Additional evidence supporting the validity of student ratings comes ironically from a series of studies widely believed to attack their validity. The Dr. Fox studies[20] demonstrated that even professors, professionals, and administrators are unable to tell when a single lecture not directly in their field of expertise is not authentic.

The series of studies carried out following this finding indicate that students, too, are not always good judges of whether teachers present more or less material than normal—not a surprising finding since the students haven't been through the course before. However, when content is equivalent, students tend to rate higher the teachers from whom they learn most.

Perry, Abrami, and Leventhal[21] carried out a well-controlled study replicating the Dr. Fox research. Their results, however, did not replicate those reported in the original Dr. Fox studies. In only one of four situations were the results similar. In general students both learned more and rated instructors higher in sections with more content and in sections in which the instructor was more expressive. However, ratings were influenced more than achievement by the instructor's expressiveness. This finding was confirmed by Meier[22] and fits with Frey's finding that student ratings of instructor skill are more highly related to student learning criteria than are student ratings on the "rapport" dimension.[23] This

---

[13] W. J. McKeachie and D. Solomon, "Student Ratings of Instructors: A Validity Study" *Journal of Educational Research*, 51 (1958), pp. 379-83.

[14] *Op. cit.*

[15] W. R. Mann, "Changes in the Level of Attitude Sophistication of College Students as a Measure of Teacher Effectiveness," (Ph.D. dissertation, University of Michigan, 1968).

[16] A. J. Drucker and H. H. Remmers, "Do Alumni and Students Differ in Their Attitudes Toward Instructors?" *Journal of Educational Psychology*, 42 (1951), pp. 129-43.

[17] J. A. Centra, "The Relationship Between Student and Alumni Ratings of Teachers," *Educational and Psychological Measurement*, 34 (1974), pp. 321-26.

[18] L. M. Aleamoni, "Typical Faculty Concerns About Student Evaluation of Instruction" (paper presented at the Symposium on Methods of Improving University Teaching, Haifa, Israel, The Technion Institute of Technology, 1974).

[19] H. W. Marsh, "The Validity of Students' Evaluations: Classroom Evaluation of Instructors Independently Nominated as Best and Worst Teachers by Graduating Seniors," *American Educational Research Journal*, 14 (1977), pp. 441-47.

[20] D. H. Naftulin, J. E. Ware, and F. A. Donnelly, "The Dr. Fox Lecture: A Paradigm of Educational Seduction," *Journal of Medical Education*, 48 (1973), pp. 630-35; J. E. Ware, and R. G. Williams, "The Dr. Fox Effect: A Study of Lecture Effectiveness and Ratings of Instruction," *Journal of Medical Education*, 50 (1975), pp. 149-56; R. G. Williams, and J. E. Ware, "Validity of Student Ratings of Instruction Under Different Incentive Conditions: A Further Study of the Dr. Fox Effect," *Journal of Educational Psychology*, 68 (1976), pp. 48-56.

[21] R. P. Perry, P. C. Abrami, and L. Leventhal, "The Effect of Instructor Expressiveness and Lecture Content on Student Ratings and Achievement," *Journal of Educational Psychology*, 70 (1979, in press).

[22] R. S. Meier, "Student Ratings of Instruction: Characteristics That Influence Evaluations of Teachers" (Ph.D dissertation, Purdue University, 1977).

[23] P. W. Frey, "A Two-dimensional Analysis of Student Ratings of Instruction," *Research in Higher Education*, 9 (1978), pp. 60-91.

does not mean that student ratings are invalid measures of "rapport." It simply means that "rapport" is not highly related to student achievement. In fact, the Dr. Fox studies provide further evidence that students rate teacher *behavior* validly, since the item showing the largest difference between their high-expressive and low-expressive conditions was "The lecturer was enthusiastic about the subject."[24]

To sum up, students know when they are learning, but they do not know whether what they are learning is current, biased, or appropriate for course goals. They do rate lower an instructor who provides less content, but their ratings of effectiveness are probably less affected by amount of content than by other characteristics of teaching. These findings suggest that when student ratings are used as evidence of teaching in promotion decisions, peers should check term papers, examinations, syllabi, etc., to determine that the content is appropriate. (Although we do not know how reliable and valid such peer judgments are, one hopes that peers can provide useful data.)

Faculty critics of student ratings sometimes complain that students cannot evaluate academic competence. Such criticism seems oddly misdirected. Surely students should not be expected to be better judges of subject-matter competence than the department chairperson, or other administrative official, who assigned the instructor to the course. One does not need to be an internationally famous researcher to teach an undergraduate course effectively, and it seems that students have the right to assume that an instructor assigned to a course will have at least minimally adequate subject-matter competence. If faculty members have doubts about an instructor's competence in the subject matter, it seems illogical for them to turn to students for such judgments. On the other hand, when there are questions about what instructors do in the classroom or how they affect students, the students themselves seem a plausible source of information.

Can Student Ratings Help Teachers Improve?

The ultimate test of the usefulness of student ratings as a measure for improving teaching is whether teaching becomes more effective as a result of the use of student ratings. Although some studies have reported no improvement, a few have reported positive results.[25] The most impressive results are those reported by Overall and Marsh.[26] As compared with a control group, students of instructors receiving feedback from student ratings not only gave their instructors more favorable ratings at the end of the year, but also scored higher on an achievement test and on a measure of motivation for further learning and application of the material learned.

Failures of improvement after feedback from student ratings may be due to any of three factors:

1. The ratings may not provide new information.
2. Low ratings and critical comments may create anxiety, discouragement, and lack of enthusiasm for teaching—lowering rather than improving motivation for teaching.
3. Even when faculty members want to improve, they may not know what to do.

Centra[27] and Pambookian[28] demonstrated that new information was important. Their research revealed that teachers whose self-ratings were higher than their students' ratings improved after receiving the student rating; teachers who were accurate or who underestimated the student ratings did not improve. Braunstein, Klein, and Pachla[29] obtained similar results. Pambookian[30] found that instructors in the middle range of ratings tended to benefit from feedback while the top and bottom teachers did not, suggesting that teachers receiving low ratings may become discouraged.

---

[24] J. E. Ware and R. G. Williams, "Discriminant Analysis of Student Ratings as a Means of Identifying Lecturers Who Differ in Enthusiasm or Information Giving," *Educational and Psychological Measurement*, 37 (1977), pp. 627-39.

[25] For example, see: Marsh, Fleiner, and Thomas, *op. cit.*; and T. M. Sherman, "The Effects of Student Formative Evaluation of Instruction on Teacher Behavior," *Journal of Educational Technology Systems*, 6 (1978), pp. 209-17.

[26] J. U. Overall and H. W. Marsh, "The Relationship Between Students Evaluation of Faculty and Instructional Improvement," (paper presented at Third International Conference on Improving University Teaching, Newcastle-upon-Tyne, 1977).

[27] J. A. Centra, "Effectiveness of Student Feedback in Modifying College Instruction," *Journal of Educational Psychology*, 65 (1973), pp. 395-401.

[28] H. S. Pambookian, "Discrepancy between Instructor and Student Evaluations of Instruction: Effect on Instructor," *Instructional Science*, 5 (1976), pp. 63-75.

[29] D. N. Braunstein, G. A. Klein, and M. Pachla, "Feedback, Expectancy and Shifts in Student Ratings of College Faculty," *Journal of Applied Psychology*, 58 (1973), pp. 254-58.

[30] H. S. Pambookian, "The Initial Level of Student Evaluation of Instruction as a Source of Influence on Instructor Change after Feedback," *Journal of Educational Psychology*, 66 (1974), pp. 52-56.

In a study[31] at the University of Michigan, we attempted to meet the conditions governing improvement following feedback of ratings by giving counseling to provide encouragement and suggesting alternative teaching strategies. This proved to be superior to a printed report of the results. Aleamoni[32] obtained similar results. Centra[33] also found that instructors were more likely to improve if they had information to help interpret their scores.

We have little research on what items, what format, or what conditions influence the usefulness of student rating forms for improving teaching. Nor do we know much about how accurate student perceptions are (although one might argue that the student perceptions are important in their own right). There is evidence that student perceptions correlate with those of trained observers.[34]

Can Students Use the Results of Student Ratings to Make Better Choices of Courses and Teachers?

So far as I can ascertain, no one has studied the validity of student ratings with respect to student uses. Not only do we not know whether they enable students to choose courses or instructors more wisely, we do not even know whether the ratings provide valid descriptions of characteristics that make a difference for student choices—right or wrong. An article in preparation by Coleman and the author on the effects of instructor course evaluations in student course selection is apparently the only study to date to demonstrate that student ratings can influence student decisions, although there are several studies indicating that instructor reputation is a factor in student choice and rating of teachers. The research by Borgida and Nisbett[35] indicated that ratings have less effect than face-to-face comments.

General Thoughts About Validity

Even though the data are now strongly supportive of the validity of student ratings for certain goals, this does not mean that they are impervious to influences by other factors. Validity studies are carried out within a given course in which a group of teachers with comparable students and comparable teaching conditions is working toward common goals. In such circumstances, student ratings provide good evidence of teaching effectiveness. But promotions committees and administrators want to use student ratings to make judgments comparing individuals in different courses and in different departments. Obviously, comparing the effectiveness of a mathematics teacher with that of a teacher of history is comparing apples and oranges. Even though one may be able to evaluate apples or oranges validly, one cannot as easily evaluate the relative worth of an apple versus an orange.

In everyday life we do, nevertheless, make such judgments regularly, deciding whether a new television set should be purchased rather than a new hi-fi, whether the head lettuce is better than the romaine, and so forth. And in academia comparisons of two professors' research are made without concern that the research may deal with different problems in different fields. If I am deciding whether to hire one professor rather than another, I make a judgment of relative merit. In our own grading of students we may worry and vacillate, but we still are able to assign grades to students who differ in terms of how well they do on objective tests, how well they have written their term papers, or how well they have participated in class discussion. So we *are* able to make judgments about relative excellence even though the excellence may be achieved along different dimensions.

Student ratings can provide information that may help make such judgments, but we should remember that ratings by different groups of students about different teachers do not necessarily provide valid comparisons between two teachers even though the ratings result in numbers that appear to be compa-

---

[31] W. J. McKeachie and Y. G. Lin, "Use of Standard Ratings in Evaluation of College Teaching," Final Report to National Institute of Education, Grant NE-6-00-3-0110 (Ann Arbor: Department of Psychology, University of Michigan, 1975).

[32] L. M. Aleamoni, "The Usefulness of Student Evaluations in Improving College Teaching," (Tucson: Office of Instructional Research and Development, University of Arizona, 1974).

[33] See note 27 above.

[34] W. M. Stallings and R. E. Spencer, "Ratings of Instructors in Accountancy 101 from Video Tape Clips" (Research Report 265). Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1967; M. S. Touq, and J. F. Feldhusen, "Validity of Student Ratings of Instructors," *College Student Journal*, 8 (1974), pp. 2-5.

[35] E. Borgida and R. E. Nisbett, "The Differential Impact of Abstract vs. Concrete Information on Decisions," *Journal of Applied Social Psychology*, 7 (1977), pp. 258-271.

rable. The evaluative judgments need to be made by peers or administrators using *evidence* from student ratings but not mechanically assigning certain values to certain numbers. Moreover, when using student ratings to evaluate teaching, we should remember that students cannot judge all aspects of teaching effectiveness equally well. Student ratings are highly valid as indices of achievement of attitudinal and motivational goals of education. They are reasonably valid as indices of achievement of cognitive goals. Judgments of the appropriateness of content, goals, and level of achievement are probably more competently made by peers.

WHAT FACTORS INFLUENCE STUDENT RATINGS OF TEACHING?

We have now seen that student ratings can provide valid evidence with respect to important aspects of teaching effectiveness, but, if we are to make good use of them, we need to know what factors may influence student ratings. Some of these factors may be valid in the sense that students may learn more and rate teachers higher in certain situations; other variables may contribute to misinterpretation. In general, the results to be reported are encouraging in that most of the factors which might be expected to invalidate ratings have relatively small effects and those factors which affect ratings also affect learning. We shall examine the evidence with respect to characteristics of students, of courses, of teachers, and of the scales themselves.

Student Characteristics

A common misconception is that only more mature, more experienced students can be expected to rate instructors validly. As indicated in my 1969 article, and in more recent reviews,[36] relatively few student characteristics have significant effects on student ratings. Age, sex, and level of student are among the variables that have been shown to have little effect upon student ratings of teaching.

Probably the single most important student variable affecting satisfaction is student expectations. Students who expect a course or teacher to be good generally find it to be so.[37] As Leventhal, Abrami, Perry, and Breen[38] have shown, students may choose certain classes or sections of classes because of the reputation of the instructor. Thus a professor's current student rating may well be a function, in part, of the reactions of former students. But students who expect a teacher to be good may be more attentive, more highly motivated, and more likely to learn than those with poor expectations.

Some writers appear to believe that students should all evaluate teachers the same way; i.e., that a teacher is equally effective with all students. When they find that some kinds of student rate a teacher higher than others do, they assume that student ratings are invalid. However, since there is some evidence that teachers may be differentially effective for different students, within-class correlations between student characteristics and ratings are not necessarily indications of invalidity of ratings. Within-class correlations between student needs and course ratings may arise because teachers who met the relevant needs were indeed more effective for those students.

For example, some studies have found that particular types of student respond differently to different teaching styles. Domino,[39] for example, found that students scoring high on the Achievement via Conformance scale of the California Psychological Inventory achieved more and rated the teaching higher in psychology sections taught in a conforming manner; students high in Achievement via Independence did relatively better and rated the teaching as more effective in sections taught in a manner emphasizing independence. Most studies of student characteristics related to ratings have lumped together students and teachers across courses in such a way that it is difficult to know what the results mean.

Course and Class Characteristics

The size of a class, whether or not it is required,

[36] F. Costin, W. T. Greenough, and R. J. Menges, "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research*, 41 (1971), pp. 511-35; K. O. Doyle, Jr., "Student Evaluation of Instruction," (Lexington, Massachusetts: Lexington Books, 1975); Kulik and McKeachie, *op. cit.*

[37] R. P. Perry, R. R. Niemi, and K. Jones, "Effect of Prior Teaching Evaluations and Lecture Presentation on Ratings of Teaching Performance," *Journal of Educational Psychology*, 66 (1974), pp. 851-56.

[38] L. Leventhal, P. C. Abrami, R. P. Perry, and L. J. Breen, "Section Selection in Multi-section Courses: Implications for the Validation and Use of Teacher Rating Forms," *Educational and Psychological Measurement*, 35 (1975), pp. 885-95.

[39] G. Domino, "Interactive Effects of Achievement Orientation and Teaching Style on Academic Achivement," *Journal of Educational Psychology*, 62 (1971), pp. 427-31.

and the subject matter—these are all characteristics that may affect ratings. While some studies have shown these variables to make a difference, others have shown no effect,[40] so that the amount of effect seems to be smaller than might be expected. Nevertheless, it seems wise not to lay heavy weight on comparisons of ratings in courses differing greatly in such characteristics. Centra[41] reports that classes of size fifteen and less are more effective in producing student learning and are also rated higher by students. Required courses tend to be rated lower than electives.

Many teachers recognize that some classes go well and others more poorly simply because of key individuals in a class or particular combinations of individuals. One student continually raising anxious questions about tests and grades can demoralize a whole class. Such characteristics of classes have not been assessed, but the possibility of such effects suggests that when student ratings are used in personnel decisions, ratings should be obtained from several classes.

Teacher Characteristics

What characteristics of instructors are related to student ratings of teaching effectiveness? For example, are certain personality characteristics related to effective teaching or to inflated student ratings? Do instructors who are easy graders get higher ratings?

Research shows relatively small effects of instructor characteristics. For example, sex of instructor makes little difference in the student ratings; conflicting results have been found with respect to faculty rank; and personality characteristics do not show consistent relationships to ratings of effectiveness.[42] In one of our studies of student ratings at the University of Michigan, we did find that teaching assistants rated by their peers as high in general cultural attainment were rated as more effective by students.[43]

Some other personality characteristics may be related to student ratings. For example, Hart and Driver[44] found that teachers scoring high in extraversion, intuitiveness, and "feeling" on the Myer Briggs Type Indicator tended to receive higher student ratings. Similarly, Murray[45] found that peer ratings of instructor extraversion, lack of anxiety, leadership, and objectivity correlated positively with mean student ratings of teaching effectiveness. Morstain[46] found that congruence of student and instructor educational orientation resulted in higher ratings. Sherman and Blackburn[47] found that highly rated teachers were perceived to be dynamic, amicable, and highly intellectual. However, we do not know whether these teachers were, or were not, effective in influencing student learning, nor can we explain the apparent contradiction between "intuitiveness" in the Hart and Driver findings and "objectivity" in the Murray results.

The research on grading practices has produced mixed results. A number of studies have found no overall effect of grading practices, although an instructor who is a hard grader is more likely to be rated low on the item, "Fairness in grading."[48] Some studies have found a tendency for teachers giving higher grades to get higher ratings. However, one might argue that in courses in which students learn more the grades should be higher and the ratings should be higher so that a correlation between average grades and ratings is not necessarily a sign of invalidity. Palmer, Carliner, and Romer[49] controlled for student achievement and found no effect of severity of grading on student ratings. My own conclusion is that one need not worry much about

---

[40] Kulik and McKeachie, *op. cit.*

[41] J. A. Centra, "Using Student Assessments to Improve Performance and Vitality," *New Directions for Institutional Research*, 20 (1978), pp. 31-49.

[42] Kulik and McKeachie, *op. cit.*

[43] R. L. Isaacson, W. J. McKeachie, and J. E. Milholland, "Correlation of Teacher Personality Variables and Student Ratings," *Journal of Educational Psychology*, 54 (1963), pp. 110-17.

[44] J. Hart and J. Driver, "Teacher Evaluation as a Function of Student and Instructor Personality," *Teaching of Psychology*, 5 (1978), pp. 198-99.

[45] H. G. Murray, "Predicting Student Ratings of College Teaching from Peer Ratings of Personality Types," *Teaching of Psychology*, 2 (1975), pp. 66-69.

[46] B. R. Morstain, "Relationship of Student and Instructor Educational Orientations with Course Ratings," *Journal of Educational Psychology*, 69 (1977), pp. 388-98.

[47] B. R. Sherman and R. T. Blackburn, "Personal Characteristics and Teaching Effectiveness of College Faculty," *Journal of Educational Psychology*, 67 (1975), pp. 124-31.

[48] J. D. Heilman and W. D. Armentrout, "The Rating of College Teachers on Ten Traits by Their Students," *Journal of Educational Psychology*, 27 (1936), pp. 197-216.

[49] J. Palmer, G. Carliner, and T. Romer, "Leniency, Learning and Evaluations," *Journal of Educational Psychology*, 70 (1978), pp. 855-863.

grading standards within the range of normal variability. If, however, grading standards seem unusually lenient, one might want to look more closely at the standards of achievement and the bases for grading in the course.

Instructor knowledge of the subject matter and knowledge of correct teaching procedures may be, but are not invariably, reflected in ratings by students. In fact, Elliott[50] found a significant negative correlation between the instructor's actual knowledge of the subject and student ratings of effectiveness. Riley, *et al*,[51] however, found that professors with published research were rated higher than those without publications, and Centra[52] found lower ratings for instructors in their first and second years of teaching. Obviously, there are some limits below which instructor knowledge is important, but, given a good textbook and an instructor willing to work, differences in instructor knowledge are probably not major determiners of student learning in most introductory courses.

In general, it seems unwise to assume that certain characteristics denote good teaching and to use student judgments about these characteristics to evaluate teaching. There is ample research evidence that good teachers come in many styles. Most presumed essentials of good teaching, such as organization, warmth, or research ability, are not highly valid.

Other Factors

One would expect that the validity of ratings would be affected by the time when they are collected. This seems not to be a critical variable. Frey[53] found that ratings collected the last week of classes were not significantly different from those collected the first week of the following term.

It may make a difference, however, whether the ratings are to be used for improving the course or for evaluating the instructor for promotion. In the latter case ratings may be higher.[54]

RELIABILITY OF STUDENT RATINGS

Questions are often raised about the reliability of student ratings, indicating that the questioner believes that a high degree of reliability is "good," but often with little understanding of why reliability is important or when it is important.

There are a number of ways of arriving at an index of a test's reliability. The use of any particular measure of reliability is, in at least a broad sense, also a measure of *construct validity*. By construct validity, we mean that a test does what it should do theoretically. We use a particular measure of reliability because we are making the theoretical assumption that the construct being measured should be stable or consistent under the conditions in which the reliability measure is obtained. Obtaining several different kinds of measures of reliability gives us some understanding of the factors influencing ratings. For example, one possible measure of the reliability of student evaluation of teaching would be the degree to which groups of students would rate teachers in the same way at two different points in time. For six classes we found that the correlation between mean student ratings of the teachers at the end of the course and mean ratings of the teachers by the same students fifteen months later was .94.[55] The classic study in this respect is that of Drucker and Remmers[56] who found correlations of about .6 between ratings at the end of the course by current students and ratings by alumni who had graduated at least ten years earlier. Overall and Marsh[57] found that the ratings of individual students one year after graduation correlated .59 with those given at the end of a course. These correlations are in one sense measures of reliability; i.e., they indicate that student ratings are not given randomly, but these results are also relevant to the problem of validity since one of the frequent criticisms directed at the validity of student ratings is that the true value of the faculty member can be apparent to students only some time

---

[50] *Op. cit.*

[51] J. E. Riley, B. F. Ryan, and M. Lifshitz, *The Student Looks at His Teacher* (New Brunswick, New Jersey: Rutgers University Press, 1950).

[52] See note 41 above.

[53] P. W. Frey, "Validity of Student Instructional Ratings: Does Timing Matter?" *Journal of Higher Education*, 47 (1976), pp. 327-36.

[54] L. M. Aleamoni and P. Z. Hexner, *The Effect of Different Sets of Instructions on Student Course and Instructor Evaluation*, Research Report No. 339 (Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1973).

[55] W. J. McKeachie, Y-G. Lin, and C. N. Mendelson, "A Small Study Assessing Teacher Effectiveness: Does Learning Last?" *Contemporary Educational Psychology*, 3 (1978), pp. 352-57.

[56] See note 16 above.

[57] J. U. Overall and H. W. Marsh, "Long-Term Stability of Students' Evaluations of Instructions: A Longitudinal Study" (paper presented at 1978 Annual Meeting of the Association for Institutional Research, Houston, May, 1978).

after a course has been completed.

On the other hand a test-retest measure of reliability is obviously not what we want if our theory expects change between the two administrations of a student rating form. For example, a high test-retest correlation for administrations of a scale of teaching behavior in two different class periods may not be appropriate if we expect effective teachers to vary their behavior from class period to class period. Reliability in terms of student ratings on more general characteristics is reasonably high (.6 to .9) when students are asked to fill out the same scales with two weeks to four weeks intervening.[58]

Some of the most common indices of reliability are computed from correlations between items. One would probably not want a high reliability coefficient on such a measure because this would indicate that we are measuring only one thing, and we usually design student rating forms to assess various characteristics of teaching rather than a single dimension. Despite this, such measures of reliability are usually high, indicating that on most student rating forms there is a strong evaluative dimension running through most of the items.

Another common confusion about reliability is the notion that students should agree on the ratings they assign an instructor. This is a reasonable expectation if the item asks students to report their observation of an instructor's behavior; e.g., "the instructor calls students by name." When, however, the item is evaluative, such as "rate the instructor's effectiveness," we can expect agreement only if we expect the instructor to be equally effective for all students, an assumption that research shows to be generally untrue. However, when one looks at *mean* ratings given by *groups* of students, the correlations between the mean ratings given by groups of students in the same class are very high, ranging from the mid-.80's for classes of ten to the .90's for classes of twenty.[59]

Other measures of reliability might deal with the degree to which a teacher's ratings tend to be the same from semester to semester or from course to course. If one believes that classes have unique characteristics which make a difference in teaching, one should not expect correlations between ratings of a teacher by different classes to be high. In fact such correlations range from .34 to .67.[60]

My point is that a reliability index is meaningful only in terms of (1) the goals for which one is administering a student ratings form; and (2) one's theoretical conception of what the constructs related to teaching effectiveness should be doing in the situations in which the measures are to be administered.

For the use of student ratings in personnel decisions, the measures of reliability one probably wants could be any of the following:

a. Would the same students rate the instructor as well if they had been given a different, but equivalent, rating form?

b. Would other students taking the same course have rated the instructor the same way?

c. Is the teacher's teaching rated the same in this course as in other courses?

Note that there are valid reasons why ratings on a different form, by different students, or in a different course might well be different. Thus we should not expect perfect correlations. Nevertheless all of these correlations tend to be quite high. Moreover, the reliability of classes of students rating teachers is substantially higher than that of colleague ratings.[61] Reliability is not likely to be a concern for most uses of student ratings.

Despite these positive findings, it still seems wise in using ratings for personnel decisions to get ratings from more than one class and from more than one course and to have the ratings interpreted by individuals who know something about the difference between courses; e.g., the difference between a content course and a methodology course.

CHOOSING SCALES OR ITEMS FOR SCALES

Establishing the Purpose of Collecting Student Opinion.

Whether one plans to use one of the ready-made scales or construct one's own, a necessary prerequisite is to examine one's goals in gathering student impressions, for different goals imply different items.

---

[58] K. O. Doyle, Jr., *Student Evaluation of Instruction* (Lexington, Massachusetts: Lexington Books, 1975) p. 39.

[59] K. A. Feldman, "Consistency and Variability Among College Students in Rating Their Teachers and Courses: A Review and Analysis," *Research in Higher Education*, 6 (1977), pp. 223-74.

[60] J. E. Morsh, G. G. Burgess, and P. N. Smith, "Student Achievement as a Measure of Instructor Effectiveness," *Journal of Educational Psychology*, 47 (1956), pp. 79-88.

[61] K. O. Doyle, Jr., and L. I. Crichton, "Student, Peer, and Self Evaluations of College Instructors," *Journal of Educational Psychology*, 70 (1978), pp. 815-26; J. A. Centra, "Colleagues as Raters of Classroom Instruction," *Journal of Higher Education*, 46 (1975), pp. 327-37.

if the goal is to assist in personnel decisions, two to five general items are probably sufficient; if the purpose is to improve instruction, a more detailed, behaviorally oriented set of items relevant to particular kinds of courses is probably more appropriate.

I shall discuss particular items useful for each of these purposes. But as an alternative to developing your own scales, you might wish to consider the use of a scale developed elsewhere, such as those developed at Educational Testing Service, Northwestern, Kansas State, or other universities included in the book by Genova, Madoff, Chin, and Thomas.[62]

In choosing items appropriate for different goals you may be helped by differentiating five types:

1. items in which students report classroom events or teacher behaviors
2. items reporting the student's perception of his or her achievement of course goals
3. items reporting the student's own evaluation of the effectiveness of different aspects of the course
4. items reporting the student's own behavior in the course
5. items reporting student satisfaction.

Different kinds of items are useful for each major purpose for which student ratings scales are used. Moreover, differing items are appropriate for differing types of instruction, such as lectures, seminars, laboratory, or tutorial instruction.

Choosing Student Rating Items for Improving Instruction

As we saw earlier, the use of student ratings is likely to result in improvement when: (a) the ratings provide new information; (b) the teacher is motivated to improve; (c) the teacher can use alternative methods of teaching effectively.

This has implications for choice of items. Items chosen by the instructor because he or she wishes the information are more likely to be informative than items on scales written for more general purposes. Aside from an item or two to indicate general feelings of satisfaction, more specific items reporting perceptions or evaluations of teacher behaviors or specific aspects of the course are likely to be more helpful than very general items.

Moreover one would guess that items worded in an evaluative fashion would elicit more defensiveness than those that are worded descriptively. Thus one might prefer such an item as:

"The instructor writes key points on the blackboard," to an item such as:

"Lectures are well organized."

Factor analyses reveal the major dimensions students and faculty members use in thinking about teaching. In constructing a scale it seems reasonable to include one or two items from each of the major factors. Among the factors commonly identified[63] are:

Skill—Enthusiasm
"The course stimulated my interest."
Structure
"The teacher defined the objectives of the discussion."
Rapport
"The teacher is friendly."
Work load—Difficulty
"The teacher expects too much from students."
Group interaction
"The teacher encourages class discussion."

Mazuca and Feldhusen[64] found in a survey of students that the first two factors above were particularly important to students.

To facilitate adaptation of student ratings to the needs of individual instructors, Purdue University developed the Purdue Cafeteria System. This system permits instructors to choose items from a catalogue of items that have been previously used. Other universities have now adopted or adapted the Cafeteria System to provide flexibility in obtaining student ratings likely to give the instructor useful information.

Choosing Student Rating Items for Personnel Decisions

The current press for teacher accountability is one of the factors leading to attempts to mandate the use of standard, uniform student rating scales for assessing teaching effectiveness. As we have already seen, student ratings of teaching *are* related to teacher effectiveness as measured by the achievement of the teacher's students. Nevertheless this does not

---

[62] W. J. Genova, M. K. Madoff, R. Chin, and G. B. Thomas, *Mutual Benefit Evaluation of Faculty and Administrators in Higher Education* (Cambridge, Massachusetts: Ballinger Publishing Co., 1976).

[63] J. A. Kulik and C. L. Kulik, "Student Ratings of Instruction," *Teaching of Psychology*, 1 (1974), pp. 51-57; and Kulik and McKeachie, *op. cit.*, note 4 above.

[64] S. A. Mazuca and J. F. Feldhusen, "Effective College Instruction: How Students See It," *College Student Journal Monograph*, 12 (1978), pp. 1-12.

mean that student ratings are sufficient evidence of teaching effectiveness. Ideally one would gather evidence from a number of sources, giving most weight to those sources most expert with respect to different aspects of teaching. For example, it is hard to conceive of anyone more expert than students themselves with respect to the degree to which the teacher has stimulated intellectual curiosity and interest in the subject matter field—an important educational goal; on the other hand, one would expect peers to be most competent to judge the scholarly content of a course, assuming that they have examined syllabi, examination papers, instructor and student lecture notes, or other sources of evidence. Thus personnel decisions inevitably involve value judgments using data from several sources with respect to a teacher's effectiveness in achieving a number of different goals.

As suggested earlier, a uniform standard scale is not likely to be very helpful for improving teaching—nor is a lengthy standard scale likely to be helpful for personnel decisions since it is unlikely to be equally well suited for different disciplines or different courses within a discipline. Since comparisons between instructors in different courses can at best be only very general, one should probably not attempt much more than to determine whether students rate an instructor as excellent, adequate, or poor. An item or two of the degree to which a course stimulated interest or curiosity, and perhaps another item or two on general effectiveness, should be sufficient for personnel purposes.

Certain items have proved to relate to teacher effectiveness as measured by mean student performance on an examination. Unfortunately such studies must be done in a multi-section course, and it is often difficult to know how much the results can be generalized to other courses. Frey's validity studies[65] included both calculus and psychology courses so that his items are particularly worthy of consideration. Three likely candidates are:

"Each class period was carefully planned in advance."

"The instructor presented the material clearly."

"This course has increased my knowledge and competence."

Other validated items include:

"Teacher Control"[66]

"Does the professor make students feel free to ask questions, disagree, express their ideas, etc.?"[67]

"Does the professor use examples from his (her) own research or experience?"[68]

One of the commonest misuses of student rating scales in personnel procedures is to sum ratings on a group of items and compute a mean. Such a procedure assumes that the items are all measuring the same thing—effective teaching. In fact most scales contain several types of items each providing useful information, but their purposes are different and they cannot meaningfully be lumped together.

Choosing Student Rating Items to Assist Students in Choosing Courses

One of the first universities to collect student ratings was Harvard, where students began in the 1920's to publish a book reporting student opinion of courses and professors. Such books are now commonly found on college and university campuses. If this is to be the primary purpose of the scale, one would presumably want to include items that are likely to provide information that will make a difference for student decisions. So far as I know there are no data describing items that made a difference, but there have been a number of studies of qualities students believe to be related to superior teaching. Feldman's[69] and Miller's[70] reviews of these studies list the following qualities as being consistently reported:

Stimulation of interest
Clarity
Knowledge of subject matter
Fairness
Preparation
Enthusiasm
Friendliness
Helpfulness
Openness to other's opinions

[65] See note 6 above.

[66] L. A. Braskamp and D. Caulley, "Student Rating and Instructor Self Ratings and Their Relationship to Student Achievement," (Mimeographed, November, 1978).

[67] M. G. Massey, *The Relationship between Instructor Ratings and Learning: Some Extensions*, (Los Angeles: California State University and Colleges, 1975).

[68] R. C. Wilson, J. G. Gaff, E. R. Dienst, L. Wood, and J. L. Bavry, *College Professors and Their Impact on Students* (New York: Wiley, 1975).

[69] K. A. Feldman, "The Superior College Teacher from the Students' View," *Research in Higher Education*, 5 (1976), pp. 243-88.

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

Items Designed to be Educational for Students

One of the potential unfortunate outcomes of the use of student ratings is that students are influenced to focus upon the instructor as the person chiefly responsible for student learning. In fact, however, learning should be a joint responsibility of students and instructor. Before blaming the instructor for failure to achieve educational goals, students should consider whether they have done all they could to make the course a valuable experience.

On my own student rating form I include a section on student responsibility for learning. My purpose is to increase students' sense of responsibility for their own learning and to encourage them to think about their own educational goals. I include such items as:

I attend class regularly.

I have created learning experiences for myself in connection with the course.

I have helped classmates learn.

General Comments About Procedures for Student Ratings of Teaching

1. Allow space for comments. Faculty members uniformly report that these are helpful. Students need the chance to express feelings that do not quite fit the prestructured questionnaire format. Frequently comments give examples or incidents which clarify the meaning of ratings or indicate what changes need to be made.

2. Indicate in the instructions who will read the comments. Students will be more focused if they understand to whom they are writing.

3. I prefer items worded in terms of the individual student's perception or evaluation to more general statements. I think a faculty member is likely to resent global evaluations more than those worded in terms of the impression made on a particular student, and I think a student is generally better able to report how he or she felt than to make global judgments.

4. Faculty members should have the right to participate in the selection of items or forms to be used in evaluating teaching in order that the form may be appropriate for the goals of the particular classes in which the data are to be collected.

5. If ratings are to be used in personnel decisions, some control should be exercised over conditions of administration. Rumors circulate about instructors who roam up and down the aisles, lose poor evaluations, or introduce the ratings by announcing that the students will be determining not only the instructor's fate but that of spouse and children.

6. I believe teachers are likely to be more effective and are more likely to improve if they enjoy teaching. This implies that reports of student ratings should be in a format that encourages good feelings rather than discouragement. Thus a report emphasizing percentile ranks in relation to norms may be less helpful than a report emphasizing the distribution of student responses (since typical classes like their instructor). It probably is not very helpful to tell a teacher rated as "good" by his students that this is only "average."

7. If teaching ratings are to be published in a booklet for students and made available to colleagues, experience indicates that faculty motivation is enhanced if the booklet reports teacher strengths rather than weaknesses.

8. Effective teaching is a skill that can be learned and that develops over time. Teaching ratings can help development. Basing career decisions on ratings in a single course early in a teaching career is not wise or fair. Basing decisions on a single visit by a peer or superior is even more unwise.

9. If student ratings are used in personnel decisions, the instructor should have an opportunity to present his or her interpretation of the data as well as whatever additional evidence seems relevant.

10. Student ratings tend to focus on classroom teaching. Evidence with respect to out-of-class educational functions such as course planning, advising, etc., also is needed.

11. When student ratings are used in personnel decisions, they should be evaluated by peers who are familiar with the courses in which the ratings were gathered, who know the teaching methods used, and who can take into account the circumstances under which the course was taught.

SUMMARY

Student ratings of teaching can be useful for several purposes, such as:

1. improving teaching
2. providing data relevant to judgment about teaching effectiveness
3. aiding student choice of course and instructor
4. stimulating students to think about their education.

Student ratings are not automatically valid and useful for any of these purposes. Thus we need to understand what student ratings can and cannot do before embarking upon large-scale institutional programs of student ratings.

We use ratings to improve the quality of education. No matter how technically sophisticated our ques-

tionnaire and our evaluation system, they are worthless if their use generates such conflict, anxiety, or confusion that education is affected adversely. Student ratings should not be used as the single measure of teaching. Rather we should think of them as data valuable for problem solving, and their impact upon the climate for teaching is more important than the technical excellence of the form to be used.

If students are to provide careful ratings and if faculty members are to make good use of the information, both need to have confidence in the methods used. Thus even a good form may need to be reexamined and revised frequently by student-faculty committees if new generations are to have a sense that the system is theirs rather than one imposed upon them.

Millman, ed., Handbook of Teacher Evaluation

CHAPTER 14

SUMMATIVE TEACHER EVALUATION

MICHAEL SCRIVEN
Evaluation Institute
University of San Francisco

INTRODUCTION

Teacher evaluation is a disaster. The practices are shoddy and the principles are unclear. Recent work has suggested some ways to clarify the issues and to make the procedures more equitable and reasonably valid, but one cannot yet point to a single exemplary system in which the practices come near to matching our knowledge. This state of affairs is a terrible indictment of the universities since the problems are sufficiently tractable that a modest investment of only approximately 2% to 3% of their *internally* controlled research funding would have solved them 50 years ago. That the great research and teaching centers of learning never put serious efforts into research on teaching, though they had to evaluate teachers all the time, is one more scandal in the history of the guilds, of professions that fail to practice what they profess, in this case, research-based decision making. Once more, the pressure to reform is coming from outside, from the Congress and the courts, not the conscience. But belated is better than never, and these notes sketch the traps and paths in the jungle, for those who want to avoid impotence, inefficiency, and inequity.

The emphasis throughout the chapter is on teacher evaluation for personnel decisions (summative evaluation), but there is some explicit reference to evaluation for faculty development (formative evaluation), and, since the two are not intrinsically distinct, there is substantial *indirect* relevance to formative evaluation. But summative evaluation is primary because (1) human careers are at stake, not "mere" improvement; (2) if it is not possible to tell when teaching is bad (or good) overall, it is not possible to tell when it has improved; (3) if it *is* possible to tell when it is bad or good, personnel decisions can be made even though it is not known how to make improvements. In short, diagnosis is sometimes easier than healing, and an essential preliminary to it.

The chapter in general applies to elementary and secondary teacher evaluation as well as to postsecondary; where key differences exist, they are discussed. To avoid cumbersome circumlocutions, the language is sometimes simplified, and the reader with, for example, K-12 interests will occasionally have to translate "dean" into "principal" and ignore references to the evaluation of research.

INSTITUTIONAL PREREQUISITES

It is improper to proceed with a system for evaluating teaching that is supposed to operate well or ethically regardless of the administrative and legal context. The following are some of the conditions that must be present in that context.

1. A system for the evaluation of administrators must be *in place* in order to avoid the entirely justifiable resistance of the "serfs" to being evaluated by those in the castle, who are above such things themselves. Administrator evaluation, if it occurs at all, is chiefly remarkable for being even worse than the evaluation of faculty. Where it is done in what is thought to be an especially enlightened fashion, it is usually based upon some bastardized version of management by objectives (MBO) which, even in optimal form, is totally inappropriate as a basis for the evaluation of administrators. But that is another story.

2. The evaluation of teaching must be part of a system that also has an appropriate process for evaluating research (if research is valued at all) and service (defined clearly), and a specific commitment to their relative weighting. Otherwise one is playing in a game for which only a few of the rules have been stated. It is notable that, for example, personnel manuals for the University of Texas system and the University of California at Berkeley (UCB) have (as of September 1980) no such commitment to the relative worth of teaching and, hence, no binding commitment toward its having *any* worth. The fact that UCB requires, and enforces the requirement, that data on teaching merit, including student evaluations, must be included in a dossier before it will be considered for personnel decisions looks as if it shows a valuing of teaching, but it does not. The data may show that the teacher is a *bad* teacher, yet the data requirement does not specify any penalty; if the data show the teacher to be a good teacher, no determinate benefit results. Specified relative weightings of all the relevant criteria of faculty assessment is the only nonvacuous procedure for objective evaluation and, not incidentally, the only equitable one. It is hardly surprising that in the two systems mentioned (and in most others), the standards vary by campus and department and, indeed, by who happens to be the current chair

or provost or vice-president. One cannot say that Berkeley does or does not value good teaching; one can only say that the official personnel manual, despite a page or two of rhetoric, does not require it for favorable action. Systems in which the weightings vary from individual to individual (and even from year to year) are fine *if* each individual's contract specifies them explicitly. It is not rigidity that is desirable, only clarity.

3. The institution must have clearly understood and defined the difference between the evaluation of worth and the evaluation of merit, with respect to the evaluation of teaching in particular. (The distinction also applies to the evaluation of research and service.) The impact of this distinction is crucial in half a dozen areas, but two examples should suffice. Suppose that an institution employs a teacher whose teaching is superlative by every reasonable standard and who is also producing substantial quantities of absolutely first-rate research, while rendering impressive service to the profession, the campus, and the community. If the institution is not absolutely clear that in spite of all this, and even though money is available, it may have to deny tenure to this teacher, the institution does not understand the difference between worth and merit.

The decision whether to make an initial appointment, the decision to grant tenure, and the decision to push for early retirement by the various means available for that should all depend upon the worth of a faculty member to the institution and not merely upon the faculty member's *professional merit*. The worth to the institution is essentially connected with such issues as the income generated and likely to be generated by the student enrollment (or grant support) that the instructor produces, the special services to the institution's general mission that are provided by the instructor (e.g., an uniquely talented Latin scholar or teacher at a Jesuit institution, a woman mathematician in an otherwise all-male department that was anxious to recruit women as mathematics majors, etc.). Neither private nor public institutions can today afford to be awarding tenure on merit alone as if the enrollments were irrelevant, but many of them are locked into a system of faculty evaluation that makes it impossible to deviate from merit considerations even in extreme cases (except at the appointment decision). Several other cases of worth are important; for example, an interesting case can be made that multidisciplinarians are worth more to an institution than most specialists, in these uncertain times, because they are less likely to be left high and dry by the tides of changing interests and emerging new disciplines.

If worth is to be given any weighting at all, as it should be, then the exact size and limit of this weighting must be as carefully defined as possible in order that one does not get into obvious abuses—such as firing teachers whose political activities lead to complaints or the loss of support from alumni, school boards, or legislatures—the sort of abuse that results from thinking that worth means only financial worth. Worth also has something to do with the integrity of the mission of the school. There is no worth left in an institution that sells its curriculum content to the highest bidder.

Incidentally, salary level will commonly involve yet another value dimension, namely, market value (or cost, or external worth). It is as well to specify the limits on these market adjustments, for example, in separate schedules for various areas.

4. There should be an independent support system of some kind (e.g., a consultant or teaching services unit) available to faculty to assist them in the effort to improve, so that the system of faculty evaluation is neither merely punitive nor seen as merely punitive. This support system should be independent of both administrators and peers, including (where relevant) department chairs and deans or assistant principals/curriculum specialists, and so on in order to avoid the disincentive to use it that results if instructors know that administrators will know (or can find out) when this support system is being used by a particular instructor. The consultations must not only be absolutely confidential but also professionally based and supportively oriented. Finding somebody who can provide the appropriate kind of assistance is very difficult because few might be said to be qualified in the first place, and many of them are too inclined to think that there is one true solution to the problem of how best to teach.

The cost of providing this support system is low. One professional and one secretary/assistant for several thousand students can do this and also do a number of other jobs that should be done in any college or school system, such as keeping up with research and practice in the fields of faculty evaluation and teaching research, maintaining a small consulting library for faculty, improving the student questionnaire (if one is used) and managing its administration, and supervising the data synthesis process in any multidimensional evaluation. That involves research-reading competence, not research performance. The helping role that works with the prima donna type of faculty is not something with which prima donna teaching researchers are familiar, so a nice balance of skills is required. It is therefore rash to make appointments to this job for more than a trial period at first.

5. The background system must have consistent and appropriate practices, not just consistent and appropriate rhetoric. A typical (college) example of how not to achieve consistent practice is to have a system in which the quality of teaching is said to be important, whereas in fact, departments are issued new positions and replacements for retiring or departing old appointees almost entirely on the basis of enrollments. Although this is admirable in the sense that it reflects an attention to a reasonable consideration that a decade or two ago was almost totally disregarded, in many contexts it amounts to rewarding the departments for, among other things, inflated grading, which is inconsistent with the rhetoric of respect for high standards.

The same type of inconsistency in a K-12 setting occurs when teachers are called in to justify failing grades but not passing ones. Another example is the use of a student questionnaire that attends to the extent to which instructors get to know the names and personalities of individual students, provide ample office hours, and so on, all of which are results easily achieved for small classes but not for large classes. Having thus provided an incentive for faculty to restrict enrollment in their classes, administrators then cry about declining enrollments, which this approach reinforces. Consistent systems are hard to set up, but the inconsistencies in systems are extremely expensive.

6. A reasonable *modus vivendi* with student government must be worked out so that both parties obtain mutual benefit from student evaluation.

7. The use of evaluation must not only be consistent but also comprehensive. In particular, each type of personnel decision must use it, especially the major ones: selection, retention, promotion/demotion, salary action, tenure and its revocation, early or postponed retirement, and layoff. In the many systems in which contracts restrict the administrator's options with respect to the major decisions, careful study always reveals a large second set of benefits or penalties, which are also options. They should be distributed with care and with regard to merit/worth; the effect of doing so can be great, and the effect of not doing so speaks loudly of the true values of the administrator and the system.

THE DEFINITION OF GOOD TEACHING

The best teaching is not that which produces the most learning, since what is learned may be worthless. There is a connection between teaching and learning, but not a simple one.

No definitions of good teaching in the literature avoid a series of counterarguments and counterexamples. The following definition avoids the so-far-identified counterexamples, but possibly has some of its own. Teachers are meritorious to the extent that they exert the maximum possible influence toward beneficial learning on the part of their students, subject to three conditions: (1) the teaching process used is ethical, (2) the curriculum coverage and the teaching process are consistent with what has been promised, and (3) the teaching process and its foreseeable effects are consistent with the appropriate institutional and professional goals and obligations.

The intent of the three qualifications can be illustrated with some examples of what they are intended to exclude. (1) Unethical processes include not only cruel and unusual punishment but also those processes that are completely nonuniversalizable, for example, putting so much pressure on the students that they abandon their work for other classes in order to do the work for this one. Since it would be impossible for all teachers to do likewise, it is an unethical procedure. (Getting students to do more homework at the expense of an already extensive leisure is, on the other hand, admirable.)

It is often argued on weak a priori grounds that the use of a reward system such as a token economy or grades is unethical; in fact, the *failure* to use some kind of "reward system"—at least a grading system—is usually both unethical and unprofessional. It is unethical because it fails to inform the students about their progress or competence, and it is unprofessional because it fails to show that the institution (or profession or future employer) values quality work as opposed to minimally competent work. It also happens that grades provide a legitimate incentive for many students; not to use them is thus poor pedagogy unless the teacher has direct evidence for a better approach. Grading can even be arranged to make interstudent competitiveness impossible, while retaining the feedback required in striving for excellence. So, even if grading did reduce learning, it would have to be done and done properly. Properly managed it should increase learning.

(2) Certain contracts are made by instructors and institutions, both explicitly and implicitly, to which they frequently do not adhere. Such contracts are involved in promises as to what a course will cover, made in, for example, a course catalogue, school or departmental handouts, the faculty handbook, the union contract, the class handouts or in the language of the opening presentation in class, or at a parents, faculty, or counselors meeting. Apart from misleading advertising, much of the hierarchical structure of sequential curricula has been made laughable by the failure of instructors to adhere to these commitments. Maximizing learning cannot be given automatic and complete precedence over these obligations.

(3) It would usually be appropriate, if the maximization of learning were the only obligation of an instructor, to adjust the level of instruction to the class average so that more people would be able to benefit from it. But there are institutional and professional commitments that transcend this commitment. For example, if one is faced with a class of medical students who will graduate at the end of this year, and if one's obligation is to instruct them in professional procedures that they will be practicing at that time, and if it is clear that most of them are so far off the pace that no instruction within the time available can get them to a level of competence, then it is professionally obligatory to focus the instruction upon those who *can* be brought up to the appropriate standard and fail the rest.

Other obligations that are important and must be considered very seriously include the problems of providing compensatory justice for minorities and women, which may mean focusing more effort on them than would maximize total class learning; providing knowledge that will be expected by the instructor of a higher level course; and so on. Ultimately there is always an obligation to quality of learning rather than to quantity which precludes

any simple maximization criterion from legitimacy, since it is essentially a quantitative criterion.

Note that the learning referred to in the definition above can certainly cover attitudes, in the rare cases in which it is possible to justify the claim that teaching a certain attitude is an appropriate part of an instructor's obligations. Teaching the scientific attitude might be a case in point, as might be teaching the motivation to learn. Note, too, that teaching can often be done, and often is done to a very important extent, outside the classroom; hence, evaluating it must involve looking for and at this out-of-classroom teaching. Remember also that the value of a teacher must include worth as well as merit, and worth involves very different considerations from maximizing learning.

The most important feature of our definition of teaching is that it does not *identify* good teaching with the production of (even good) learning, though it does not break all connections either. I have discussed several overriding considerations already, but there is yet another type of example that should be mentioned. If you have a number of students in the class who for various reasons unconnected with your own performance are not working hard enough to keep up with the class, you should not be downgraded as a teacher for the failure of these students to learn. If you happen to get half the basketball team in your class, it should not count against you if they are not planning to graduate or are not capable of it. A teacher's task is only to provide the *best possible* environment, not to guarantee that the results will be effective no matter how little effort is made by the students. It follows that one must never use on a student questionnaire the question "How much did you learn from this class (that you think was valuable to you)?" or cognates of that question. One must instead use questions such as "How well do you think the instructor taught the course?"

It is partly for the above reasons that the answers to student questionnaires are not to be regarded as intrinsically inferior to or as substitutes for studies of the learning gains of students in class. They are *more valid* in one respect, in that they address the correct question.

Another incorrect definition of good teaching which has some supporters, perhaps deserves a footnote. This involves the identification of teaching with the transfer of learning from the teacher to the student. But, as has often been pointed out, inspiring the student to seek learning elsewhere may be the best approach for maximizing learning even in the short run and more likely in the long run.

HOW NOT TO EVALUATE TEACHING

It will be clear from the preceding discussion that any method of evaluating teaching that simply identifies teaching merit with the amount learned is

oversimplified and can be extremely unfair to some teachers, in particular to teachers of slow learners. It should also be clear that evaluating a teacher's worth only, or merit only, is, on occasions, also entirely inappropriate. And in the absence of certain institutional requirements, good evaluation is either difficult, inappropriate, or impossible. We should now look at a number of other attractive errors that have also been widely incorporated into systems for evaluating teachers.

Classroom Visits

Using classroom visits by colleagues (or administrators or "experts") to evaluate teaching is not just incorrect, it is a disgrace. First, the visit itself alters the teaching, so that the visitor is not looking at a representative sample. This defect is exacerbated by preannouncing the visit. Second, the number of visits is too small to be an accurate sample from which to generalize, even if it were a random sample. Third, the visitors are typically not devoid of independent personal prejudices in favor of or against the teacher, arising from the fact that visitors are normally administrators or colleagues of the teacher and in their other role are involved in adversary proceedings with them, alliances with them, and so on. Fourth, nothing that could be observed in the classroom (apart from the most bizarre special cases) can be used as a basis for an inference to any conclusion about the merit of the teaching. That this is so follows inexorably from the results of the enormous number of studies on style research that have been done and summarized on various occasions (see Centra, 1979). These result in the conclusion that no style indicators can be said to correlate reliably with short- or long-term learning by the students across the whole range of subjects, levels, students, and circumstances. (Anything less could scarcely be defended for personnel evaluation.) Fifth, regardless of the fact that no observations of teaching style can legitimately be used as a basis for inference about the merit of the teaching, the visitor normally believes the contrary. This is often because visitors have their own preferences as to a certain style or have many years of experience in teaching this same type of course or student. Consequently, they believe that not doing it their way, or perhaps in one or two other ways that they approve, is doing it badly. These prejudices are without foundation, to the best of our knowledge, and should not be allowed to come into the evaluation of teaching in any normal case. Among the hypothetically possible exceptions are the possibility that what the teacher is saying is known to be false by a visitor and is so grossly false as to constitute an impossible vehicle for teaching the truth, that the visitor observes racist or sexist or other immoral practices by the teacher, or that the visitor observes a total lack of classroom discipline to a degree that cannot possibly be reconciled with the continuance of a learning process of any kind. Since none of these events has ever been recorded on any of the classroom visiting sheets of

the many thousands that I have either inspected directly or of which I have seen summaries, this cannot be taken seriously as a reason for making classroom visits part of teacher evaluation. It is scarcely surprising that the research studies (see Centra, 1979, pp. 74-76) show that colleague reports are not even mutually consistent; hence they are unusable as evidence.

Ultimately, the problem about the visitor is the lack of similarity between the visitor's thought processes and those of the student. They are generally separated by several decades in their learning maturity; thus, they may have substantially different vocabularies and cognitive repertoires, and they certainly lack many cultural similarities. Because of this, the visitor's empathic impressions are not likely to be a good indicator of how much learning is going on in the heads of the students (quite apart from the corrections that have to be applied to this first approximation to the key process in good teaching). Since the secondary indicators in teaching style (i.e., everything else you can see, leaving the empathy aside) turn out to be invalid as indicators of teaching effectiveness, this leaves the visitor—or should leave the visitor—at a loose end.

In spite of these staggering objections, the method of classroom visitation is of course the universal method whereby teachers in the elementary and high school are evaluated and is—depressingly enough—being quite steadily implemented at the postsecondary level, on the grounds that it represents an improvement over past practices. No variation of this kind of observation is of any value either; visits by experts are no better, and visits by peers are no better, for purposes of personnel evaluation. There are no valid indicators to be seen, no matter who looks. Visits by a consultant are defensible in the effort to provide *help to improve* teaching if it is already known that the teaching is very unsuccessful, because the consultant may be able to suggest some options. However, the (costly) classroom visit is usually unnecessary; a tape recording, or even a verbal report by the teacher, along with the student evaluation forms, is more than enough basis for consulting recommendations in most cases.

Course Content

Teaching is usually evaluated without any serious attempt at evaluating the quality of the content of the course. This is one side of the "methods-madness" that has made schools of education a laughingstock in the intellectual community, and increasingly in the total community, for many years. From what we have said about the definition of good teaching, it is clear that one cannot make one's judgment of teaching merit entirely on the basis of the content of *what is learned by* the students, but to do so is much better than making the judgment on the basis of merely inspecting *what is presented to* students. Both should be considered, but with the emphasis upon student

performance. Here is the one place where peer evaluation of a limited kind is appropriate—evaluation of materials (both texts and student work) not process—and even here it is better to use people from another institution but in the same subject-matter area, eliminating costs by trading services in kind with that institution. Such an arrangement, with its attendant social pressure, tends to do more to improve standards than in-house evaluation of materials does anyway, apart from increased validity.

Teaching Processes

If one half of methods-madness is not looking at content, the other half is looking too hard at process. This not only involves the obvious traps of substituting style preferences for validated merit criteria but it also often involves making the formulation of behavioral objectives (or lesson plans) an end in itself. Yet the very best teachers (e.g., Socrates), on any criteria we can justify plausibly often eschew a predefinition of coverage because of the advantages of picking up on what turns out to be of current or great interest or difficulty to the particular students being taught, the "targets-of-opportunity" approach. Excesses are possible here, too; neither approach should be unbridled or totally excluded.

Enrollments in Further Courses

Another popularly acclaimed, though infrequently employed, measure of teaching success is checking on the relative number of enrollments in further courses in the same subject matter by "graduates" of the instructor being evaluated. This is an unethical indicator, and its use cannot be countenanced by central administration because it is nonuniversalizable. That is, one can only score points on this dimension by stealing, buying, or seducing students from other departments. It is exactly like the process of getting more work out of the students by having them give up their homework for other courses. However, reducing dropouts—the other end of the effects spectrum—is a highly legitimate criterion of merit, if the education missed can be shown to be a truly serious loss.

Student Questionnaires

Most student questionnaires are improper bases for faculty evaluation, despite the great potential of the approach, either because they involve ratings on style or ratings on nonuniversalizable indexes or because of the way in which the data from them are synthesized. If only mean ratings are used, then the important case of the instructors who are tremendously successful with a subgroup of the class, perhaps the best students, is overlooked. Although their mean score may be no different from those of other

instructors who receive weak ratings from everybody in the class, their teaching potential is obviously different, and a sensitive administration should be making some provisions to tap such a promising source of inspiration in a more appropriate situation than the type of class that resulted in a low overall score. Or the consultant may be able to work out a way to generalize from this promising basis to success with the rest of the students. Similarly, averaging the means from courses taught at different levels may be unfair to teachers who are good at one grade level and not at another or good with introductory but not with graduate courses.

Alumni Surveys

More honored in principle than in practice, alumni surveys are essentially useless for normal formative and summative evaluation of teachers though they can sometimes be used for course or program evaluation. They have extremely low response rates; they relate to the ancestors of present performance and, hence, may exaggerate or underestimate the merit of the only performance that should be used for current personnel decisions, namely, current performance; they involve a perspective as to what will be valuable that may no longer apply to new graduates; and the attributions of causation involved are suspect. These reasons do not exclude some use of alumni surveys in selecting Distinguished (Elderly) Teacher Awards.

Overkill

Evaluation done every term on every course is likely to produce hostility from both faculty and students and stereotyped responses or low response rates, especially if a long form is used. It is also unnecessarily expensive, takes up too much class time, and is unlikely to achieve the best results because it allows no chance for experimentation, for example with new texts or approaches. Of course, undersampling (once every five years) and letting the instructor choose which classes are to be evaluated are worse alternatives.

HOW TO EVALUATE TEACHING

Student Questionnaires

The student questionnaire should be a key component in the evaluation process from about grade six. The piece of paper itself is only part of the story. Preparing students as evaluators is another part, and the methods for administering the questionnaire are a third. These methods must be proof against complaints about the possibility of selective return and prompting. A

good straightforward approach is to have assistants from the central administration staff (a cheaper fallback system is using secretaries) take the questionnaires out to each class, have the instructor leave the room for a few minutes, provide a brief explanation of the process and how the results are to be used (possibly in writing), encourage questions, and pass out questionnaires to be filled out by every member of the class. One should get a 99% return rate from those present, and one should worry if it is less than 75% of those who complete the course.

At the college level, the date for distribution should be announced in advance if one wants most of those enrolled for the class to attend, since knowing that the questionnaire will be distributed on that date is sometimes an incentive to attendance. This does weaken one's defenses against certain types of preparation of the students by the professor, but this can be controlled by asking about it on the form so it is a less worrying problem than low attendance rates. "Absentee ballots" are not a good idea, logistically or for inferential purposes. One can get that 99% response rate and tolerance from the faculty for taking time out of their classes, if the whole process takes less than five minutes. This is the first of three interlocking reasons for using a very short questionnaire.

The second is the cost, which goes up quickly with length of printing, collating, keypunching or scanning, synthesizing, reporting, and interpreting. The third is that all the apparent justifications for using a long questionnaire are unsound. The usual questionnaires with 20 to 70 items are fishing expeditions of an entirely improper kind. Unless it is demonstrable that some possible answer to each particular question indicates merit or demerit in teaching, then the question has no place on the personnel evaluation questionnaire. Detailed questions may be used in a questionnaire designed by the instructor with the assistance of a teaching support consultant, or vice versa, when the instructor wants to get feedback on some specific effort or style venture of personal concern. But no such style choices can be legitimated for purposes of personnel evaluation, and no such questions should ever be on a form seen by anybody except the instructors and the consultants of their choice. If placed on a form along with legitimate questions, they will be likely to bias the response of somebody who favors or disfavors the style that they uncover, and such biases are illegitimate. For this reason they would be certain to affect seriously the legitimacy of any personnel decisions that were appealed in court, for example, but the ethical point is more serious than that.

Since one cannot help but be worried by the contamination of instructor ratings by irrelevant personality factors and by factors connected with like or dislike of the subject matter or course content, one should make some effort to syphon off those considerations. A simple way is to stress the contrast

between these considerations in the verbal and written introduction to the basic judgmental question on the questionnaire, that is, the difference between liking or disliking the instructor as a person and thinking well of him or her as an instructor and between liking or disliking the course content and thinking that the instructor did or did not do a good job of teaching it. Another approach would consist in asking specifically for an expression of liking (1) for the instructor as a person and (2) for the subject matter, and *then* asking for (3) an evaluation of the job done by the instructor in teaching this course. The content rating (2) could then be torn off the form and sent to the relevant administrator for *course* evaluation; the personal evaluation (1) could be torn off and thrown away or given to the instructor if requested, and one would then use the rest of the form (3) for personnel evaluation.

Special thought must be given to the cases in which the instructor *is* responsible for course content (e.g., graduate level specialty courses) and cases where attitude toward subject matter is important (e.g., in K-3 reading instruction). A promising approach is to divide the "attitude toward content" question into two parts: "general attitude" and "attitude improved because of this teacher's approach to it," and add the latter rating to the instructor's file. It is also possible that the latter will be picked up on the holistic rating of goodness of teaching. We need some empirical research to find out whether this is true; meanwhile, the safest approach is to add the extra question.

The crucial question is a simple request for an overall judgment of the merit of the instructor as an instructor, and this question is enough. There are some other possible questions on such a form of a kind not often encountered. They should come before the requests for an overall rating, since they tend to depress it (and probably increase its validity). Technically this is desirable, because the main problem with the usual results is that the scores are too high to allow adequate headroom within which exceptionally good teaching can distinguish itself. These other possible questions are of two overlapping but loosely distinguishable kinds. The first concerns matters that can be described as ethical or professional obligations of the instructor and can be phrased in either a positive or a negative way or alternately. They may concern such matters as the match between the preannounced content of the course and the actual content, the match between the content of the course and the content of the tests or assignments, the extent to which the reasons for grades were explicitly and adequately justified, the extent to which the possibility of appealing a grade or disciplinary action was explained, the reliability with which scheduled class and conference hours were met, the use of racist or sexist or other bigoted remarks or materials (in and out of class), and so on. This is the so-called Black Marks list (when phrased negatively), and the extent to which it reminds instructors, as well as students, of the minimum professional obligations of the instructor is surprising. These are obligations that can be discharged rather easily and that one constantly finds are an underlying cause of dissatisfaction and bad holistic ratings.

The second kind of legitimate question consists of a listing of the components of instruction for independent "microassessment," that is, "please rate the merit of the catalog description/text/quizzes/grading system/class handouts/sections/labs/fieldwork, etc." Probably the best general-purpose form includes these component-evaluation microquestions, the overall macroquestion, and (optionally) a request for suggested improvements in the form and in the process of administering it. But one can vastly reduce the effort by first using only the 1-3 macroevaluation question(s) and going to component analysis only when the instructor requests it or is seriously and regularly below the mean for comparable courses. At that point, when remedial action is appropriate, the components evaluation is an obvious source of help. When a crucial personnel decision has to be made, the "professionality" questionnaire can provide a useful and valid supplement to information from the other two. (The two can be rather easily combined, at some cost in length that is unimportant for occasional use.)

Data from these forms should be accompanied in the file by comments on them from the instructor. Note that in all of the preceding only student evaluations are used. Students are in the best position to judge, and the evidence suggests they are at least quite good judges (i.e., there is a strong positive correlation of ratings with actual learning), especially if some care is given to preparing them for the rating task and making its value and use clear. Below the sixth grade, student evaluations require more preparation of the class and perhaps can best be done in group discussions. Where possible, ratings by paraprofessionals should be used, either as a supplement to or (in K-4) a substitute for student ratings. Failing these, no substitution should be made of ratings by administrators.

Quality and Professionality

The quality and professionality of content and process must be given careful consideration. The three key quality dimensions are currency, correctness, and comprehensiveness. These are picked up by peer or expert review. For professionality, the concern is with the implementation of a just and helpful teaching process. Ratings for both are made from a sample of (1) the materials provided, (2) the texts required and recommended, (3) the exams, (4) the term paper or project topics, (5) student performances, (6) the instructor's performance in giving appropriate grades for student work, and (7) the instructor's performance in justifying the grades to the students (via written material) and providing other helpful feedback, for example, comments on projects or term papers. The workload involved in this kind of evaluation is quite modest, contrary to appearances, because only a (matrix)

sample is required. Other sources of data on these issues are the students' evaluations and the instructor's self-report. Major discrepancies among the data sources require further investigation.

Some traps exist. The text cannot gain points for the instructor unless the instructor chooses it; even then, since it may be little used, it can only account for a few points, perhaps 4 out of 20 or 30. The grades given to students might be all A's or all F's and justified, or they might be "on the curve" and unjustified, either because the performances in general deserved better or worse or because of particular inconsistencies (i.e., inequity). One needs to see the performances, the grades, and the justifications.

Although peers must be used for the evaluation of content, they are usually and culpably not competent to evaluate all that falls under this category, notably what we have here called professionality. This is the technical competence of the instructor in the small area of validated pedagogy, for example, in constructing examination questions that are unambiguous, involve adequate coverage of the intended target area, are not overcued, and so on. This kind of quality is rather easily identified by somebody with good skills in test construction, and such a person should routinely review samples of the tests of those instructors that are up for (or a year from) particularly important personnel decisions like tenure. In the K-12 situation, this is only sometimes a key skill, but it should be checked whenever relevant.

Another point at which expert evaluation of professionality has some relevance concerns the way in which essays are graded or marked. There are professionally required standards here, with which virtually no faculty member at universities has the slightest familiarity; K-12 teachers are often better trained. As a remedy for this situation, when it is unsatisfactory, administrators should request that, in talking about self-improvement, the instructor fill out, as a normal part of the process, a form indicating how papers are in fact graded. The kind of grading that is legitimate involves at least the following requirements: the papers are graded "blind"; the papers are graded question by question, not paper by paper, to avoid the known large halo effect that results from having read a good or bad first question by the same author just before reading the second question; the exams are shuffled after each question is graded so that different students stand the brunt of one's fatigue (or initial optimism) for different questions; and when one has graded all the Nth questions in a given set of papers, one must then go back and grade the first 6 to 10 papers again in order to see whether one's grading standards have slipped, upward or downward. With multiple-choice exams, many of the preceding requirements are otiose, but the requirement for technical proficiency in constructing them becomes much more important and is something beyond the competence of most academic instructors at the moment.

Actual Learning Gains

Learning gains are useful in evaluating teaching only if one can, in advance, specify and justify the standards or comparisons that will be used. By themselves, they have no legitimate interpretation except, perhaps, when they are zero or negative. In the case of elementary schools that randomly assign pupils to teachers, or for multiply taught sections for an introductory college course, comparisons between sections can be extremely useful. In cases in which there are national norms, they can sometimes be useful. And, for purposes of instructional improvement, one can run a comparison against one's own previous performance and discover whether slight or large changes in approach or text turn out to have significant results.

The main aim of the comparisons is to try to determine what could reasonably have been taught to the students. Although it is always fairly straightforward to find out what was actually taught, that is evaluatively useless without some sense of what was possible and what was remarkable. If allocation of students between the afternoon sections of a large introductory class is random, then one can rather easily discover something quite important about what can be done, if several instructors are involved and prepared to experiment. Patterns of low or high performance by a particular instructor that extend across a couple of years or more become very significant. Shorter periods of study are likely to yield results that are due merely to unrecognized idiosyncracies of the particular classroom, time, groups present, and so on.

One must first realize that only highly salient performances deserve any attention at all and that there may be none of these. No rankings should be squeezed out of this kind of data but if someone consistently produces half the gain scores of five others, after warnings and help, it is time to fire him or her. (The 8th Circuit Court of Appeals in 1979 upheld a superintendent's right to do this.) Second, it is nice that just where student evaluations are least plausible (K-5), the measurement of learning gains is easiest. Third, this method also works on a posttest only basis, with random assignment, but of course the tests must be (1) secure, (2) reasonably comprehensive, and (3) professionally constructed. Fourth, superteachers do sometimes emerge from this process and should be given the opportunity to work as the teaching improvement consultants, with plenty of recognition and reward. They deserve every bit of it. Fifth, using this approach on sections of introductory courses at the college level gives a wonderful cross-check on the validity of student evaluations, at least at the first-year level.

Professional Development Dossier

Although the results of self-evaluation and self-improvement efforts should show up eventually in improved performance on the above scales, it

is worth including them directly in a dossier since doing so encourages faculty efforts at professional development and hence speeds up the improvement. This procedure also improves the validity of decisions that have to be made under a time constraint. A professional development dossier should include a rationale for each course's or year's methods and coverage (where these are the instructor's option); evidence about professional development activity such as readings, courses, workshops, and consulting aimed to improve teaching; and, most important, a description of planned and performed experiments on the individual instructor's own teaching approach. A basic list of courses taught, enrollments, and grades; of committees (instructional, departmental, school and district); and of other formalized service should be included here, preferably as hard-copy output from the institutional data base. These four crucial types of data can be ramified usefully and without significant extra cost; for example, feedback from the student should be supplemented by feedback from teaching assistants or teaching aides when these are employed and when confidentiality of the responses can be preserved (as is usually the case).

For purposes of faculty self-development (only), the student questionnaire can be expanded to include an open-ended question in which there is a call for identifying particularly good and particularly bad features of the instruction. One should keep the volume of these responses down when teaching over 500 students a year, because it is so hard to simplify large numbers of these responses, by requesting that they be provided only when either the top or the bottom two scoring categories are used (i.e., only when an A or a D or an F rating of the instructor is given by the student). This also improves their utility, because it avoids the problem of, for example, interpreting favorable responses with long lists of criticisms attached. Responding to what are perceived as salient features is more sensible than trying to balance out large numbers of free-form responses.

Exit Interviews

In addition to the student ratings of the instructor during the class, another type of student rating is exceptionally valuable, at least when done sporadically, and that is the "exit" interview done at graduation. It falls midway between the in-class rating and the alumni rating and is vastly better than the alumni rating because: the rate of return can be around 100% instead of around or below 30%, memories are more reliable at the earlier time, it refers to more recent editions of the courses and the instructor, and causal inferences are more likely to be valid at that point. Exit interviews or questionnaires have sometimes turned up really interesting results, identifying instructors who are thought of as truly outstanding by a quite disproportionately large proportion of the students, but who do not show up in any other search. The "interview" may consist in no more than an additional short form to fill out when applying for the degree or certificate.

The approach outlined here does a great deal to protect faculty from two sources of injustice that are particularly pervasive in current systems. The first is the kind of injustice that makes it impossible for the teacher of a generally unpopular but absolutely essential prerequisite course, for example, calculus for architect students, to score really well compared to the usual norms. The other kind of injustice relates to the individuals who put an enormous amount of work into developing a course, keeping it up to date, and making it as effective a teaching device as they possibly can, at the expense of lots of happy friendly humanistic interchanges and socializing. When an evaluation system is used that does not pay a great deal of attention to the content of the course and that throws in irrelevant requests that ask students to rate the course on "touchy-feely" dimensions, these efforts are largely unrewarded. (Touchy-feely has a place somewhere sometimes, probably as a secondary advantage after the bases are covered; if we are careless in setting up a teacher evaluation system, we reverse these priorities.)

LOGISTICAL CONSIDERATIONS ABOUT STUDENT QUESTIONNAIRES

Much of the preceding refers to summative evaluation for personnel decisions. For that purpose, it is quite important to give out the class questionnaires, if used, at about the same time in the term for all instructors. What is the best time? The faculty at the University of San Francisco argued that the time from the 10th through the 12th weeks of class (semester system) is about right, not too near to the beginning to give a reading on the basis of inadequate experience and not too near to the end so as to interfere with the intensive review period for the examinations and the occasional drop in attendance while students work at home. Getting all the classes visited within this period is logistically feasible for a relatively small campus but would involve substantial difficulties in a large one. However, validity considerations suggest that administration of the questionnaire after final grades have been issued is preferable, and this procedure also increases headroom on the distribution.

How often should the evaluation process be run? Every second or third course given, or term of teaching (K-6), stratified by upper and lower division and graduate categories (college), is a good compromise. Moreover, it cuts the personnel and time requirements for distributing and collect-

ing the forms by one-half or two-thirds and gives faculty a chance to do something about an unsatisfactory readout before the next evaluation is upon them. Student ratings are only one measure and an obtrusive one. Unobtrusive measures, such as reading a sample of student papers, assignment lists, and so on, could be used more often, though it seems unnecessary. More effectively, a matrix sampling across teaching assignments and evaluation components might be used.

On the issue of preserving confidentiality and the integrity of data collection, one could, especially in the precollege situation where consultant help is scarce, turn to secretaries to distribute and collect questionnaires, but there is still a credibility problem about that. It is much better not to use graduate students or indeed any students from the same campus, for similar reasons. The system of having each class elect a student who will do the job and take the materials over to the central administrative offices has worked quite well at some institutions, but it does not provide a person who can answer questions and exhibit authority and inside knowledge about the details and importance of the process and its results. Under no circumstances can one go to mail ballots, or "drop it in the nearest collection box" ballots, because the response rate deteriorates seriously. It is hard to justify validity with return rates under 75%, especially if they are variable between classes. One must select the best procedure in terms of the accessible resources at a particular site.

Although it is fairly easy to use optically scannable cards, this usually gets one into the business of providing the appropriate marking pencils, which can be a problem. It is better to use straightforward forms and have them keypunched; this approach, rather surprisingly, often turns out to be less expensive, even with an error check. The computer program for combining the data on the types of questions that I have been talking about is extremely simple and can provide a variety of interesting readouts. After investigating variations between small and large classes, required and optional classes, first-year and fourth-year classes, and differences between fields, one often gets a pleasant surprise, finding that most of these differences are not sufficiently important to be worth reporting separately, although one can run them through the machine for a check every time the system is operating. In general, however, the principle is always the same; teachers should only be compared to teachers with comparable tasks.

One should also have the computer automatically flag performances that are a half (or a full) standard deviation above the mean and those that are one (or two) standard deviations below the mean, not because a standard deviation means something in terms of traditional significance with the kind of skewed distribution one gets here, but because it is a convenient and quite appropriate flag. If one does not use a set of questions on components or

professionality before the holistic question, one will probably have to set the upward flag at half or two-thirds of a standard deviation in order to get any success stories at all, but the top or bottom 15%, 10%, and so on should not be used because the system does not make a ranking meaningful. The standard deviation is just a convenient statistical measure that works well here; there may be better ones, for example, semiinterquartile range. It may well be the case that no teachers are two standard deviations down on any occasion, and none may be consistently down by as much as one. Such a situation is quite desirable, if the overall performance is about where it should be. The system described does not require that there be any losers, unlike a ranking system in which there must always be a bottom 10%, and there can still be winners.

The general issue of records-keeping deserves a mention here. Great universities and small school districts or private schools alike often have no record of what courses an instructor has taught over the years before, for example, a promotion review and, hence, no enrollment or grading data. School districts often have venerable central office records, defined by the board as the only official record and from which adverse comments disappear, hence no longer "exist." College administrators (and not just college faculty) are known to have falsified student responses in order to impugn or favor a particular individual. It is quite easy to avoid all these abuses with a little care: Faculty and administrators should regard this as a necessary duty for the needed fresh start toward satisfactory faculty evaluation.

Certain by-products of the questionnaire kind of approach are of some interest to administrators, especially at the college level. For example, one can get a very good readout on the actual teaching load of the faculty, in terms of numbers of classes and numbers of live bodies; it turns out that there is sometimes a startling number of "phantom" classes, that is, classes that are not in fact meeting although they have not been cancelled, usually because there weren't enough people to justify a scheduled class meeting but the instructor didn't want to convey to the department chair or dean the fact that the assigned teaching load had partly evaporated. A comparison of the number of grades awarded with the number present at the time of the official questionnaire will give an attendance ratio, which is also something that bears watching for faculty members who are trying to improve their performance. The number present for the evaluation should not be printed out on the summary sheets without a reference to the number absent, if the central administration's records or the registrar's computer can produce a current figure on that. Printing out some index of bimodal distributions is also easy but should, in general, be reserved for the situation in which someone is looking for help. Which brings us back to the "development connection."

HOW TO USE STUDENT EVALUATIONS IN A FACULTY DEVELOPMENT PROCESS

Most of the preceding refers fairly specifically to summative evaluation; it is time to say something more specific about formative evaluation, partly because the summative system described above is often thought to be "hostile" to formative evaluation. In general, there is no need at all for the latter to involve the kind of rigorous supervision of questionnaire administration so far described. Nor need formative evaluation be done at any particular scheduled time, except when a rather careful evaluation of some new procedure is involved. The aim is simply to get useful suggestions for improvement; which leads to very different procedures. In my classes I have frequently used the following rather eccentric approach; although collegiate in specifics, much of it can be transferred to K-12. Evaluation forms are distributed to everybody who comes into the class the moment they walk in the door on the first or second or third day, while the tourists are still shopping around. These forms request the student to turn them in by putting them on the front desk or by giving them to another student to turn in, if they decide not to continue with the course after even one or two sessions. One wants to hear from them whether there was something misleading about the advertising, something needs-irrelevant about the content, or something offputting about the early presentations, and this is the best possible way to pick it up. Perhaps, for example, there was something about the way one outlined the proposed examination and assignment process that seemed particularly formidable. One will not pick these objections up from people that left because of them and hence are not there toward the end when the usual forms are distributed. Indeed, the usual process is heavily biased in favor of the instructor because of the highly self-selected population that stays around. One loses much of the critical and negative feedback, and I personally find that I learn a good deal from it, though I do not pretend that I like reading all of it.

With respect to students who stay around, one next requests that all of them should feel free to submit anonymous suggestions and criticism at any time throughout the term. One needs to make very careful and specific arrangements about this, to preserve their anonymity and to encourage them to see that one is going to do something with the results. One might issue feedback forms and put a collection box in the back of the room, perhaps passing it around once a week. Immediately after the midsemester test is returned, one requests that a full component analysis form, preferably including a Black Marks list, be submitted. A student volunteer may help with synthesis. Now one has the chance to show that this input really is valued, by discussing it and taking whatever steps seem best to improve the class in the light of it. This is also a very good time to demonstrate that some of the criticisms are contradictory and, hence, that one cannot satisfy everybody.

Next, one applies to the faculty senate or the appropriate administrator for permission to give an early final. The final is given during the last class (or two classes) and is quickly corrected so that it can be returned at the time originally scheduled for the official final in finals week. Attendance at that session is required, as it would be for a final, the penalty for absence being an incomplete in the course and possibly a lower grade. At that final session, the exam questions are reviewed in front of the class, an answer key with "model answers" is handed out, a sheet of actual but unsatisfactory answers that illustrate common mistakes is handed out with the appropriate commentary on it, an opportunity is provided to raise objections to the grade with the instructor and perhaps with teaching assistants, and then the students are required to hand in a final evaluation form on the class.

Now one finds out whether one's attempted remedies worked with those who requested them. The students who had been optimistically supposing that they were going to get an A and, in the light of this, had given the instructor quite a good grade in the 10th, 11th, or 12th week are now suddenly confronted with reality; unsurprisingly, their evaluation of the instructor is often affected. However, one cannot expect them to provide as accurate a judgment of the course before receiving a grade as after. Moreover this arrangement—apart from improving the evidential basis of the student evaluations and, probably, the validity of the ratings—makes the exams part of the learning process, not an arrow shot into the air and falling to earth one knows not where except for a grade. This scenario illustrates the point that improvement and effort should be directed by the most accurate evaluation that can be combined with good teaching practice, whether or not the institution uses the best system. Credibility here can be a trade-off for validity, since one does not have to persuade anyone except oneself. That is a major feature of formative evaluation.

In the course of formative evaluation, it is entirely appropriate to ask questions of the students about style, when one is striving to achieve a particular style. But one's rationale for such striving needs very careful examination. It is even more appropriate to ask some microquestions that identify possible problem areas, such as the text, handouts, quizzes, midterm, assignments, grading process, treatment of questioners, availability in office and after class, and final exams. There are no style assumptions lying behind the belief that it is better to perform well than poorly in each of these areas. Whereas it is certainly possible to have all instructors ask everybody about these matters on the standard form, that is, to build them into the general form, such a procedure is not ideal; the consequent increase in the load on the computer and/or staff is not only costly but also the process is

less effective because the rigid time for collection of that data does not provide one with the opportunity to see whether improvements can be implemented that work well with the class that registered the complaints. A one-shot feedback is never as useful formatively as the two in the above model. A reasonable compromise that makes the extra load worthwhile for the institution consists in instructors using the "long form" (with the micro-questions, and so on) twice, once after the midterm test or assignment, for the instructor's own edification and to identify needed changes, and once at the official time, with only the holistic responses going into the official record.

Now we come to the sequence of external events related to formative evaluation. If an instructor is getting bad ratings overall on a short form that do not seem to be explicable in terms of, for example, the peculiarity that the course is required and has unpopular content, then the next step is to recommend the use of the components analysis ("long") form that I have just mentioned. Professional consulting provided at this point will often suggest a number of ways to improve performance. Only if all of that fails should one even consider going to an analysis of style and a discussion of alternative possible approaches to that, because the results of style research make this a last resort. The consultant has several nonstyle possibilities to worry about first, the professionality considerations, such as the giving of unusually low grades, which may precipitate a suggestion that the instructor alter a particular teaching process because it is unprofessional.

It seems probable that most people who have the requisite content knowledge are capable of becoming quite good instructors, but it is not clear that many of them do so. It is possible or likely that some faculty—including those with tenure—neither could nor will become or remain reasonable instructors. Hence, an evaluation system must, first, identify any unsatisfactory instructors, second, try to upgrade performance, and, third, provide evidence to remove them either from the faculty or from the instructional faculty if improvement does not occur.

Tenure is not an insurmountable obstacle to removal for nonperformance at an appropriate level, and tenure should be phased out anyway. We are now moving away from the legal model of the "master-slave" relationship in colleges and toward the legal model of "just cause" for dismissal, as a result of the affirmative action legislation and the gradual development of an orientation toward collective bargaining. Therefore, previous fears about violations of academic freedom, which were the most important basis for tenure, are somewhat less of a worry and can certainly be taken care of through protective contracts, whether formal or informal. The residual infamy of tenure, due to the number of people who are "burned out" (or wrongly tenured) but kept on by it, stands as a proclamation of lack of responsibility that is becoming increasingly prominent as the hard times for

education develop. We cannot afford to continue that way, and we cannot move any other way without a rock-solid process for evaluating teaching and for improving it, which must be the court of first resort. But one must not forget that *valid summative-type evaluation is the essential basis for recommending and detecting improvement.*

INTEGRATING THE RESULTS OF FACULTY EVALUATIONS

Whereas the results of formative feedback go to the instructor alone or to some consultant chosen by the instructor, the summative evaluations from students also go to the responsible first-level administrator. They must be combined, on appropriate occasions, with the information in the files, direct measurement of learning gains (if available) and appropriate comparisons, ratings by teaching assistants, the quality ratings by topic experts of the content and by process experts of the examining and certain other professional processes, and a pick-up of out-of-class contributions to teaching, for example, individual tutoring, science club talks, forums, campus newspaper articles, or curriculum revision work. In addition, considerations of worth must be brought in for the appropriate decisions. The integrative process is a very tricky one, and all the preceding work may be wasted if bias can sneak in, as it usually can and does. It is not possible to give a complete outline of what should happen at this point without excessive length, but the ideal toward which one should strive is clear enough: The integrative process should be one in which the relative importance of the components is clearly expressed in advance, so that the calculation of an overall score/grade is automatic, given the grades on each component, that is, a simple weighting and summing procedure. In this respect it should be the same as the procedure used at the college level for combining the results of the teaching evaluation with the results of research and service evaluations. Evaluating teachers involves more than evaluating their teaching, though we have focused on that, as the hardest part. Extreme rigidity in the integrative process is far better (although not necessary) than allowing the department chair or the dean or the principal to do a seat-of-the-pants synthesis of the data at this point, a major way for bias to come in.

There are, of course, a variety of situations in which one does not want to use equal weighting for all these components, but it should be a matter of complete openness to all involved as to how the actual weights are determined and what they are. Where it is possible to arrange for individual variations in the "contract," this should be done at the time of appointment, subject to later revision, and in fact usually can be done.

This is an incomplete version of what is only a fairly complicated procedure. It is a perfectly comprehensible procedure and can readily be made a

rather equitable one. One of the results to be avoided if at all possible is the pressure to go with different evaluative approaches by different departments or schools. It may be a useful political procedure to start off with that as a possibility in order to get people on board who cannot admit that their teaching process in a lab or a clinic has anything in common with the teaching process in a philosophy seminar or a basic statistics course. But the real truth of the matter is that the pluralistic view is without intrinsic merit. All essentials of the procedures recommended here work perfectly well in a kindergarten or on a university campus, for any course from finger-painting, through School of Music courses, to molecular biology. One may want to add an extra question or two here, drop one or two there, especially from the formative evaluation procedures and perhaps even from the Black Marks list, but these are very minor changes. At a given site, all of the questions should be on all of the forms that are generally distributed, with a "not applicable" option for the students. The backup use of other forms, with reference to style, where necessary or desired, may certainly be made as idiosyncratic as the instructor pleases, and schools or departments may have a preferred version of these. But those forms should not get into the personnel review process even at the departmental level, because of the legal, ethical, and scientific hazards.

KEY ETHICAL CONSTRAINTS ON FACULTY EVALUATIONS

The underlying reason for taking such a firm line about excluding style-related questions and answers and observations from personnel evaluation procedures is not merely the absence of any scientific evidence that connects particular approaches to teaching with successful results; it is a much graver matter. It affects formative as well as summative evaluation, which is why it has been left to the end. The error in racism, or sexism, does not lie in the empirical falsehood of the claim, for example, that the crime rate among blacks is higher than it is among whites, or that management success with male subordinates is much less likely for women than for men; both these and many other similar generalizations are probable or true. However, their truth fails to justify the appeal to them in order to make a personnel decision adverse to a black or female candidate, not because it happens to be illegal or unethical, but because it is scientifically indefensible. Generalizations like this are very much less reliable as predictors than inferences made from the track records of the individual candidates. In addition, it is unquestionable that any use of such generalizations would lead to self-fulfilling and socially undesirable results, so on that pragmatic as well as ethical ground they must also be abhorred. One does not need the additional ground in order to see

that, no matter what the results of research on teaching styles ever turns out to be, one will never be able to use those results for making decisions about individual instructors because to do so would be a case of guilt by association. One may well ask what the point of research on teaching styles is, if it can never be used in this way. The best answer is that it can increase our repertoire of possibly useful last-resort options. The guilt-by-association point is simply this. Even if high correlations between certain teaching styles and good learning did exist, there would be individuals who used a different style and were successful, since correlations are only statistical. One could not take adverse action against someone who shared the characteristics that tend to identify the unsuccessful group, since that person might be one of the exceptions. One can only use data that refer specifically to that individual's performance: the learning gains of that person's students, the student evaluations of that person, and so on. Not only does this avoid the ethical errors but also the evidential force of such data far overpowers any loose generalization about teaching styles; it is, in fact, exactly the kind of data on which the style generalizations are based. Hence, going to it in the actual case preempts the generalization. The student and content quality judgments are based on this exact case; the generalizations are based only on cases something like this case. Given the generally quite good validity of the student evaluations, when they are discrepant with style desiderata (if any) the student ratings would have to be treated as better indicators. In their absence, and the absence of other instructor-specific data, the case is undecidable. Personnel decisions require personal data.

The preceding result affects formative evaluation too. Because of the costs in time and psychic energy, teachers should not be subjected to any pressure to modify teaching style except as a last resort, after the failure of components and professional evaluations to turn up remedies. Hence, no style questions should appear on standard required questionnaires.

Ethics and, increasingly, the law require certain other steps when possibly unfavorable personnel action is contemplated. These include:

(1) a chance to review the evidence and react to it
(2) a chance to scrutinize the chain of argument from the evidence to the unfavorable conclusion
(3) advance warning that provides time for improvement and a clear description of what degree of improvement will be satisfactory (This should not be taken to mean that the administration must provide a sure-fire remedy—because that is not always possible—but only a clear definition of what would constitute acceptable performance.)
(4) the above events to be recorded as having occurred on specified dates, preferably with the log signed off by both parties (This is the "audit trail" requirement.)

(5) since age (till 70) will be disallowed as a criterion, we urgently need a sound evaluation system that can be used throughout the tenured years, or else no dotard can be dropped before 70 (and perhaps not then, in the future). Often a reduction in load, and salary, and a change of title is the kindest move, but it must be done using a system applied to everyone regardless of age, or it will be disallowed, and rightly so. "Applied" means enforced, not pantomimed, on younger tenured teachers.

A final point on the ethical side. It is not ethical, and it may not even be legal, to deny to the students who generate the key evaluation judgments in the process recommended above the opportunity to see the summarized results. One might certainly argue against it with respect to some detailed written-in remarks, because those cannot properly be used without balancing them against all the rest, something that students may not be in a position to do; but the summarized holistic responses, and their distribution, cannot properly be withheld from the students. In addition to questions of propriety, there is the thoroughly unattractive possibility that the students, frustrated in their attempt to get these results, will set up their own evaluation system and the administration will then either have to cooperate with it, which increases the chance of questionnaire overload with a reduction in response rate and validity as well as a loss of class time, or take the indefensible position that students are not allowed to poll their peers in order to evaluate instructors.

Some compromises will be necessary in negotiating a joint effort, because the kind of evaluation of instructors that students are interested in is as different from the kind described as formative is from summative. Students are interested in such questions as the assignment mode and load, the cost of the textbooks and whether they are really necessary, and certain teaching style variables, which can well be an appropriate basis for a student with a certain learning style to use in selecting in or out of a particular class. But compromises of this kind should and can be made in the interest of a campus-community approach. Student morale and cooperation are not well-served by leaving them out of this process, and, in the long run, faculty evaluation and faculty improvement suffers from not having the best input from students.

None of the above entails an obligation to publish the results. As Centra has pointed out to me, making them available might be enough. However, publishing may produce more substantial improvement by faculty and more learning by students.

CONCLUSION

Personnel evaluation is, in general, a field that requires some care and attention, and it is usually done with amazing incompetence, as one can

over by studying the forms used by the White House, by large
s, or by the military. If it is to be improved, presumably the
should be the source of the leadership. These institutions, after
t of the little relevant research and interpretation. In the key
e evaluation of teachers at the K-12 as well as at the college level,
hown a disgraceful disinclination to work on it, and in recent
n that indifference has partly evaporated because of heat from
ere remains a depressing failure to do it in a defensible way.
remarks in this and the opening paragraph of this article will
ciently irritating to lead to improvement. It is easy to refute them:
nly produce a single school district or college in the country
her evaluation procedures avoid the gross invalidity and injustice
from the dozen sources of error discussed here. Or even if that
n done, one need only show that there are good excuses for not
affirm that neither response is possible and that faculty, students,
ers deserve better.

REFERENCES

. *Determining Faculty Effectiveness*, San Francisco: Jossey-Bass, 1979.
et al. "Do student ratings involve guilt by association?" *Evaluation Notes*, 1980
nd 1981 (March).

Wichita State University

INTER-DEPARTMENTAL CORRESPONDENCE

ATTACHMENT C
To Agenda 11/8/82

Document 10

To Bill Mathis, University Senate President Date November 3, 1982

From Jim Clark, Chair, Faculty Welfare Committee

Subject Faculty Survey on Salaries and Fringe Benefits

Attached are some very brief summary statistics on the results of the fringe benefit survey (450 usable surveys were returned). More detailed statistics will be available at the next Senate meeting, or may be obtained from me before the meeting (stop by 109 Clinton or call x3220).

Sample survey filed following this Summary

Responses to Salary vs. Fringe Benefits Question

"Out of a 10% increase, how much do you want in salary and how much in fringe benefits?"

Option	% Choosing
10% salary, 0% fringe benefits	10.6%
8% salary, 2% fringe benefits	20.9%
6% salary, 4% fringe benefits	30.0%
4% salary, 6% fringe benefits	19.2%
2% salary, 8% fringe benefits	8.9%
0% salary, 10% fringe benefits	10.3%

Some Significant Correlations

1. Those choosing higher percentages of fringe benefits tend to be older and higher ranking faculty.

2. Compared to younger faculty, older faculty give higher rankings to increasing TIAA/CREF and early retirement benefits, while younger faculty give higher rankings to higher salaries, dental insurance, and free faculty tuition at Regents' institutions.

SUMMARY STATISTICS

Item	Evaluation (1-5)		Ranking (1-15)	
	Median	Mode	Median	Mode
1. Higher salaries	1.3	1	1.7	1
2. Increase TIAA/CREF	1.4	1	2.6	2
3. Dental insurance	2.3	2	4.6	4
4. Optical insurance	3.0	3	7.6	8
5. Occupational therapy	4.1	4	11.7	13
6. Psych. insurance	3.9	5	11.1	14
7. Family med. insurance*	2.2	1	5.3	3
8. Liability insurance	3.1	3	8.8	7
9. Prepaid legal services	3.6	4	10.7	12
10. Early retirement benefits	2.3	2	6.9	6
11. Free faculty tuition	3.3	5	9.8	13
12. Free family tuition*	2.9	5	9.1	14
13. Ticket discounts	4.2	5	13.0	15
14. Enrichment	2.5	2	7.3	7
15. Menu approach*	2.5	1	6.0	1

* Summary statistics may be misleading due to multimodal distributions.

Wichita State University

Distributed to entire Faculty

INTER-DEPARTMENTAL CORRESPONDENCE

To WSU Faculty Date October 7, 1982

From Faculty Welfare Committee

Subject Survey on Fringe Benefits

For the first time in many years the Kansas Legislature has appointed a committee to take a serious look at improving fringe benefits for state employees. In response, the Council of Senate Presidents is trying to develop a unified position for all the Regents' universities on the relative importance to faculty of fringe benefits improvements compared to salary increases, and on the relative importance of different possible fringe benefit improvements. Your answers on this survey will determine WSU's position on these questions. Taking a few minutes to think about and answer (on the response form) the survey questions will help make the survey as representative as possible.

I. The first part of the survey asks for information about you, using the IDENTIFICATION block on the form. Please use the codes below to fill in this part of the survey. (If you would prefer not to give this information, please go on and answer the rest of the survey anyway.)

A. Faculty Rank (in your primary role)

1. Professor
2. Associate professor
3. Assistant professor
4. Instructor
5. Administrator
6. Other

B. College/division affiliation

1. College of Business Administration
2. College of Education
3. College of Engineering
4. College of Fine Arts
5. College of Health Related Professions
6. Fairmount College
7. Academic Services
8. Central Administration
9. Other

C. Years at WSU

1. Less than 3
2. 3 - 6
3. 7 - 9
4. 10 - 19
5. 20 - 29
6. 30 or more

D. Years of Age

1. under 20
2. 20 - 29
3. 30 - 39
4. 40 - 49
5. 50 - 59
6. 60 - 69
7. 70 and over

Appointment type

1. Tenured
2. Probationary
3. Continuing
4. Temporary
5. Other

II. The second part of the survey asks you to indicate your relative preference for salary increases compared to fringe benefit improvement, using the GENERAL CODE J block on the form. Assume you have a dollar amount equal to 10% of your salary to receive as a salary increase and/or fringe benefits increase. (We should mention that, since fringe benefits are not subject to income-based taxes, the state could deliver more benefits to you per dollar spent by buying you more fringe benefits than by adding to your salary.) Which one of the following combinations would you most prefer?

1. all 10% in salary, no increase in fringe benefits
2. 8% in salary, 2% in fringe benefits
3. 6% in salary, 4% in fringe benefits
4. 4% in salary, 6% in fringe benefits
5. 2% in salary, 8% in fringe benefits
6. no increase in salary, all 10% in fringe benefits

III. The third part of the survey asks you to give both your ranking of the importance of salary increases and various fringe benefits and your absolute preferences for these items. In the left column of the answer sheet, rank the items from 1 to 15, with 1 as most important and 15 as least important (no ties, please - give each item a different rank). Then evaluate the importance of each item individually to you, on a scale of 1 = "extremely important to me" to 5 = "of no importance to me."

When you have finished, please return the answer form unfolded to the Faculty Welfare Committee, Box 78. If you have any questions, please call your representative on the Faculty Welfare Committee, or call Jim Clark (committee chair) at Ext. 3220. Thank you for your assistance in our efforts.

Faculty Welfare Committee

Russ Adkins	Academic Services
Lloyd Benningfield	Administration
Jim Clark	Business
Dennis Ingrisano	Education
Don Killian	Liberal Arts and Sciences
Susan Kruger	Health Related Professions
Ray Olivero	Fine Arts
Mel Snyder	Engineering

UNIVERSITY SENATE

WICHITA STATE UNIVERSITY

Minutes of the meeting of November 8, 1982, (Vol. XIX, No. 5).

Members Present: Aagaard, Alexander, Billings, Breazeale, Brewer, Brinkman, Carmody, Chaffee, Childs, Clark, Crown, Dreifort, Duell, Egbert, Gosman, Graham, Greenberg, Harmon, Hunt, James, Janeksela, Kruger, Lee, Mathis, McCabe, Meisch, Menhusen, Milbrandt, Millett, Myers, Nelson, Olson, Rozzelle, Schoenhofer, Schrag, Sojka, Stevens, Terrell, Thomann, Throckmorton, Tilford, Triplett, Wilkerson, Wineke, Zoller.
Members Absent: Ahlberg, Davis, Fox, May, McCollum, McLeod, Rhatigan, Soles, Tanner, Thibault, Wilhelm.
Guests: Virgil Pangburn, Ben Rogers, Milt Myers, Paula Rhoads, Lorraine Kee, Bob Wherritt, Catherine Yeotis, Mary Lou Goodyear, Sid Rodenberg.

I. CALL TO ORDER

President Mathis called the meeting to order at 3:35 p.m.

II. INFORMAL PROPOSALS AND STATEMENTS

Senator McCabe moved a suspension of the rules for the purpose of allowing the reporter from the Sunflower to tape-record informal statements and announcements. Senator Hunt seconded.

VOTE

The motion to suspend the rules was defeated on a standing vote, 14 - 18.

Vice President Breazeale read a press release issued by President Ahlberg earlier that day announcing the resignation of Mr. Ted Bredehoft as Athletic Director and the appointment of C. Russell Wentworth as Interim Athletic Director.

III. APPROVAL OF MINUTES

The minutes of the October 25, 1982 meeting of the Senate were approved as distributed.

IV. NEW BUSINESS

Statement on Professional Ethics

The Senate received the revised version of the Statement on Professional Ethics which had been passed at the September 27th meeting of the Senate. Senator Wineke, chair of the editing committee, briefly summarized the changes made in the document.

Senator Zoller asked whether the document was intended to cover the activities of Teaching Assistants. Senator Clark replied that the Welfare Committee had intended that the sections dealing with teaching responsibilities were intended to apply to anyone with a teaching function at the University.

SUSPENSION OF RULES

President Mathis suggested suspending the rules to allow change in the order of agenda items for Senate consideration. There was no objection.

Summary of Survey on Fringe Benefits

Senator Clark, chair of the Welfare Committee, presented the committee's report of the Survey on Fringe Benefits (Attachment C). He announced that he had prepared twenty-five copies of a more complete report, with an item-by-item analysis of responses to the questionnaire, available for any Senator who wished to inspect it. He observed that the results of the

survey were not particularly surprising, although responses to a question regarding the preferred distribution (salary + fringe benefits) of a hypothetical 10% increase were more balanced than he had anticipated.

Senator Gosman asked if there was any chance that the results of this and similar surveys conducted at other Regents Institutions could be correlated and made to frame a consensus on the issue before requests were sent to the State. President Mathis stated that he has talked to all but two Senate Presidents at the Regents Institutions and that their responses have not been uniform. He also noted that methods of soliciting faculty preferences have differed among the institutions and stated that he did not know how existing differences could be reconciled. He announced that the issue would be discussed at the upcoming meeting of Senate Presidents.

Report on Teaching Evaluation

Dr. Ben Rogers, chair of the ad hoc Committee on Teaching Evaluation, presented that committee's report (Attachment B). In his remarks he noted that the committee had initially been charged to make a study of student evaluation of teaching performance. In reviewing literature on the subject, the committee found a fairly consistent pattern of evidence and felt that it would be in a position to make positive recommendations respecting the development and implementation of a program of evaluation. Accordingly, the committee requested and received an enlargement of its charge.

Dr. Rogers stated that the committee wished to educate the faculty as much as possible regarding the process of student evaluation. Thus it included with its report a survey of evaluation procedures currently in use on this campus, as well as the copies of several useful articles on the subject. After directing the Senate's attention to a minor error, he moved adoption of the report.

Senator Graham expressed praise for the committee's effort and asked if the committee members might be identified. President Mathis then identified John Belt, Charles Burdsal, Ben Rogers (chair), Robert Egbert, Ron Winters, Catherine Yeotis, Douglass Lee, and Tim Dickenson as committee members.

MOTION

Senator Dreifort moved a division of the question so that the report proper and the recommendations might be considered separately. Senator Brinkman seconded.

Senator Terrell objected that a statement on page 15 of the report, regarding evaluation procedures in the College of Business Administration, was either erroneous or misleading. That not being the question immediately before the Senate, his remarks were ruled out of order.

VOTE

The motion to divide the question passed.

Senator Terrell took issue with the section of the report (page 15) in which procedures in the College of Business Administration are discussed. He pointed out that the College faculty did agree on the adoption of an experimental system of student evaluations on the condition that faculty members participating would not be identified, that the survey would be conducted once every five years, and that the adoption of this system would not set a precedent leading to the imposition of a mandatory system of student evaluations. He also pointed out that faculty in Business Administration had also approved a resolution that acknowledged the legitimacy

of a variety of teaching styles and the right of faculty members to regulate teaching activities on the same basis that they regulate other professional activities. Dr. Rogers stated that the committee's report was not intended to contradict the points made by Senator Terrell. He pointed out that in none of the student evaluations conducted on campus are faculty members publicly identified. Senator Terrell responded that in the project conducted by Business Administration faculty did not even receive results of evaluations conducted in their own classes. Dr. Rogers asked if it were not true that, once all results had been gathered to form a data base for the College, individual faculty members were given the results from their classes and allowed to use them as they saw fit. Senator Terrell replied that that was the case, but that he would like to see the distinction made explicit in the report. Dr. Rogers offered to do that.

Senator Clark asked if there is any role for anecdotal evidence in summative evaluations. Dr. Rogers replied that the role of such evidence is not widely discussed in the literature. He pointed out that administrators regularly receive, and remember, this kind of information, but there is a problem with soliciting and gathering it systematically. He also observed that there is strong interest in peer evaluation, but procedures are lacking.

Senator Zoller asked if the report and recommendations, if approved by the Senate, would be sent to the General Faculty. President Mathis surmised that it probably would be.

Senator Billings asked who is mandated to administer this program. Dr. Rogers replied that adoption of a questionnaire would be mandated and that it would be used, first, to establish a data base for measurement of faculty performance. Its further use--for instance, in personnel reviews--is not mandated. Dr. Rogers stated, however, that a questionnaire of the kind proposed would provide a faculty member with a good means of soliciting student evaluations of his teaching performance; faculty are under pressure to provide evidence of teaching competence and should have a reliable means of doing so at their disposal.

Senator Gosman observed that alternative methods of evaluating teaching have been tried in the past and have not generally succeeded. He expressed concern that, in all likelihood, student evaluation will become the only method of teacher evaluation. Dr. Rogers pointed out that the ad hoc committee did not consider the matter of alternative methods of evaluation but restricted itself, as charged, to the issue of student evaluation.

Senator Greenberg likened teaching evaluation to intelligence testing, observing that the latter process has not been notably successful in measuring intelligence and that intelligence testing has been perverted to political uses. He expressed fear that the same development could occur in teaching evaluation. Dr. Rogers replied that the committee is aware that there has been misuse of evaluations on this campus and elsewhere; that is why the committee wanted to create the best, most reliable possible questionnaire to evaluate teaching effectiveness. He also observed that the adoption of a university-wide system of evaluation would allow students to register their opinion in reviews of faculty for tenure and promotion.

Senator Dreifort, noting that the report states that there is no direct correlation between teaching evaluation and teaching improvement, asked if the committee had any particular feeling on that question. Dr. Rogers replied that on certain types of evaluations--for instance, formative evaluations conducted during a semester--there was some correlation; that is generally not the case with summative evaluations.

Senator Billings objected to the use of questionnaires, observing that most teachers, in testing their students, do not rely on objective examinations. It seems inconsistent to use objective questionnaires in evaluation of faculty. She argued that the burden of evaluation should be placed on chairmen and administrators and that we ought not to overvalue questionnaires if we do adopt them.

Senator Terrell acknowledged that the committee had received a difficult task, for there is a great deal of literature on the subject in the various field journals. He asked how extensive the committee's research was. Dr. Rogers replied that the committee did not exhaust the research and that it relied primarily on the articles cited in the report because they seemed to be highly reliable surveys. Senator Terrell objected that there are other points of view than those in the articles cited and that it would be useful to have those presented.

Senator Milbrandt and, after him, Senator Zoller raised questions regarding the usefulness and fairness of a standardized questionnaire to be used by faculty in a wide variety of disciplines. Dr. Rogers replied that standards of performance in various disciplines could be determined once data bases were established.

MOTION Senator Dreifort moved the previous question. Senator Thomann seconded.

VOTE The motion to move the previous question passed.

VOTE The motion to accept the Report of the ad hoc Committee on Teaching Evaluation passed.

Senator Sojka asked if one standardized questionnaire could be used for both formative and summative evaluations. Dr. Rogers replied that it would depend on the nature of the questionnaire. On this campus, the LASTIC and IDEA forms are used both ways. He noted, however, that students have objected to the length of these forms; on the other hand, a short form may lack reliability.

Senator Thomann asked if, on recommendation #1, we are mandating that every class be evaluated at least once. Dr. Rogers replied that the committee is not ready to recommend that at this time. Senator Thomann observed that, to establish a reliable, unbiased data base, we would almost have to require universal compliance.

Senator Greenberg stated that such a recommendation should be approved by the General Faculty.

Senator Dreifort observed that teaching evaluations, while purportedly not mandatory, are in fact mandatory for any faculty member whose performance is reviewed. He also objected that teaching evaluations fail to include, and therefore account for, such influential factors as the physical conditions under which faculty must teach and students learn.

Senator Gosman stated that recommendation #1 raises a number of unanswered questions and suggested that the Senate is being asked "to buy a pig in a poke."

Senator Zoller observed that the report does not include an estimate of the anticipated costs of a university-wide evaluation program.

Senator Nelson stated that if the University has decided to make teaching performance an important factor in tenure, promotion, and salary reviews, it is up to us to decide how to measure it; anecdotal and other types of hearsay evidence are unreliable. She also stated that the faculty, not administrators, should take responsibility for evaluating teaching performance.

Senator Thomann observed that the Senate, in approving the recommendations, would not be approving a particular questionnaire; rather, it would be appointing a committee to develop one that would eventually have to be approved.

Senator Terrell expressed several concerns about the recommendations, based on his belief that evaluations are best conducted by experts. Noting that there is virtually no evidence of teaching performance other than student evaluations in current tenure and promotion dossiers, he observed that the faculty may already have defaulted on its responsibility to ensure that teaching performance be subject to expert evaluation.

Senator Greenberg reiterated his concern about the possible political uses of student evaluations. Senator Billings renewed her appeal for the development of an alternative to the presently proposed system.

MOTION — Senator Thomann moved adjournment. Senator Zoller seconded.

Senator Greenberg asked, as a point of information, what would happen to the question before the Senate. President Mathis replied that it would go automatically on the agenda for the next Senate meeting.

VOTE ADJOURNMENT — The motion to adjourn passed and the meeting adjourned at 5:08 p.m.

Donald Wineke, Senate Secretary
Lucille Brodie, Recording Secretary