DOCUMENT RESUME

ED 221 148 HE 015 513

AUTHOR Whitman, Neal; Weiss, Elaine
TITLE Faculty Evaluation: The Use of Explicit Criteria for Promotion, Retention, and Tenure. AAHE-ERIC/Higher Education Research Report No. 2, 1982.
INSTITUTION American Association for Higher Education, Washington, D.C.; ERIC Clearinghouse on Higher Education, Washington, D.C.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 82
CONTRACT 400-77-0073
NOTE 57p.
AVAILABLE FROM American Association for Higher Education, One Dupont Circle, Suite 600, Washington, DC 20036 ($5.00, members; $6.50, nonmembers).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Administrator Role; Case Studies; *College Faculty; Data Collection; *Employment Practices; *Evaluation Criteria; Faculty Development; *Faculty Evaluation; Higher Education; *Measurement Techniques; Peer Evaluation; Personnel Policy; Productivity; Standards; Student Evaluation of Teacher Performance; *Teacher Effectiveness; Teacher Role
IDENTIFIERS Oregon State System of Higher Education; Southern Regional Education Board; State University of New York

ABSTRACT

The use of explicit, written criteria to evaluate college and university faculty is examined. Four major issues concerning faculty evaluation are as follows: the desired outcomes of faculty evaluation; the functions of faculty activity that are to be evaluated; the criteria that should be used for each area evaluated; and the procedures for implementing the faculty evaluation program. In general, studies identify two major outcomes: personnel decisions made regarding promotion, retention, and tenure; and feedback to faculty leading to faculty improvement. The major areas to be evaluated are teaching, research, and service. For the most part, faculty evaluation programs attempt to increase objectivity through both qualitative and quantitative approaches. To achieve qualitative objectivity, criteria are developed to improve the quality of data collected from an individual evaluator. To achieve quantitative objectivity, data are collected from multiple data sources. A critical issue in faculty evaluation is determining how data are collected and reviewed. At some institutions, faculty are expected to provide evidence of their teaching, research, and service effectiveness. Since the 1970s, there has been a trend toward systematic, standardized data collection. The role played by administrators in faculty evaluation is addressed. Meta evaluations (i.e., evaluating the methods of evaluation) conducted by Oregon State System of Higher Education, the State University of New York Faculty Council of Community Colleges, and the Southern Regional Education Board are examined. Appended materials include a form for peer review of undergraduate teaching based on dossier materials, guidelines for use of results of the student instructional report, and a bibliography. (SW)

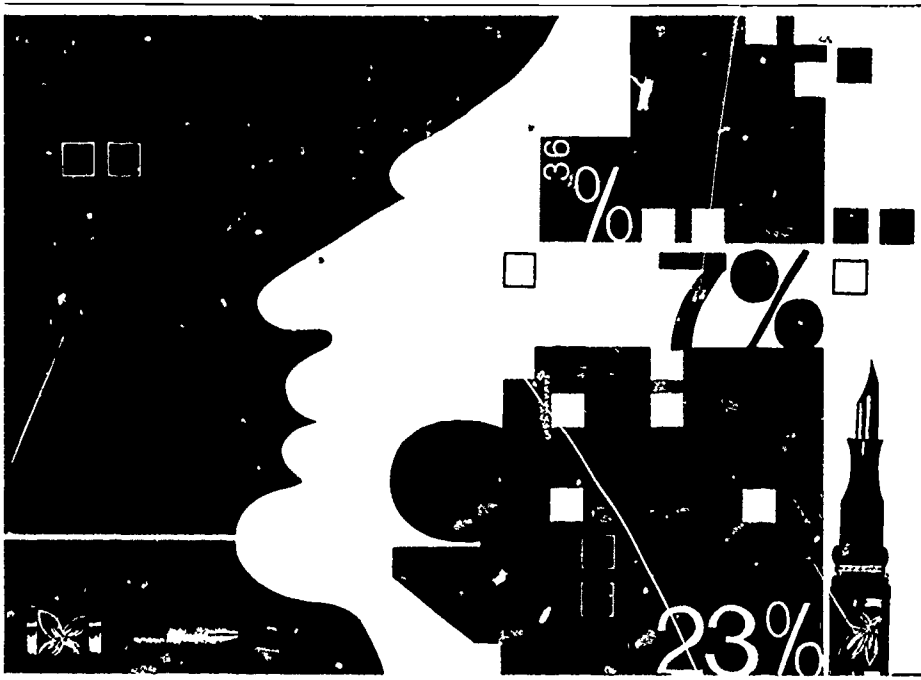
Report *1982*

2

Faculty Evaluation:

The Use of Explicit Criteria for Promotion, Retention, and Tenure

Neal Whitman
Elaine Weiss



ERIC®
AAHE

Faculty Evaluation: The Use of Explicit Criteria for Promotion, Retention, and Tenure

Neal Whitman and Elaine Weiss

AAHE-ERIC/Higher Education Research Report No. 2, 1982

Prepared by

ERIC®

Clearinghouse on Higher Education
The George Washington University

Published by

AAHE

American Association for Higher Education

Cite as:
Whitman, Neal and Weiss, Elaine. *Faculty Evaluation: The Use of Explicit Criteria for Promotion, Retention, and Tenure.* AAHE-ERIC/Higher Education Research Report No. 2, 1982. Washington, D.C.: American Association for Higher Education, 1982.

ERIC Clearinghouse on Higher Education
The George Washington University
One Dupont Circle, Suite 630
Washington, D.C. 20036

American Association for Higher Education
One Dupont Circle, Suite 600
Washington, D.C. 20036

This publication was prepared with funding from the National Institute of Education, U.S. Department of Education, under contract no. 400-77-0073. The opinions expressed in this report do not necessarily reflect the positions or policies of NIE or the Department.

Contents

Board of Readers

The following individuals critiqued and provided suggestions on manuscripts in the 1982 AAHE-ERIC/Higher Education Research Report series:

Vinod Chachra, Virginia Polytechnic Institute and State University
Randall Dahl, Kentucky Council on Higher Education
Kenneth Eble, University of Utah
Nathaniel Gage, Stanford University
Lyman Glenny, University of California at Berkeley
Harold Hodgkinson, National Training Laboratories
Arthur Levine, Carnegie Foundation for the Advancement of Teaching
Michael Marien, Future Study
James Mingle, Southern Regional Education Board
Wilbert McKeachie, University of Michigan
Kenneth Mortimer, Pennsylvania State University
Marvin Peterson, University of Michigan
Robert Scott, Indiana Commission for Higher Education

Foreword

As colleges and universities experience various financial pressures and face less-than-certain enrollment trends, there is an increased interest in the methods and criteria by which faculty performance is assessed. This interest has been made evident in the meta-evaluations conducted by systems of higher education, more systematic programs of faculty evaluation developed by individual institutions, and research studies carried out by evaluation specialists.

Prior to 1970, faculty evaluations were usually conducted in an informal fashion. In recent years, however, more formal evaluation methods have been developed and used on an increasingly widespread basis. The utilization of more systematic faculty evaluation methods has been manifested by the development and use of explicit written criteria.

In the coming decade, the use of formalized, explicit faculty evaluation will become more commonplace due to a low turnover in faculty. While institutional growth dwindles or stops, faculty retirement age is being extended, and institutions are finding themselves with a large proportion of tenured faculty. Thus, colleges and universities must devise systems to promote teaching and research excellence, at the same time as they respond to mounting financial pressures and changes in the education "marketplace." Furthermore, decisions in the nation's courts compel institutions to provide documentation to back up personnel decisions. In the history of American higher education there probably has been no time where the internal and external forces have come together so strongly to support a formalized system to measure the performance of faculty.

This Research Report by Neal Whitman, director of educational development, Department of Family and Community Medicine at the University of Utah School of Medicine, and Elaine Weiss, president of Educational Dimensions, Inc., Salt Lake City, examines the use of explicit, written criteria to evaluate college and university faculty. It also traces the use of evaluations in faculty development initiatives and promotion, retention, and tenure decisions. It will be of interest to academic administrators responsible for conducting faculty evaluations as well as to faculty who are the focus of such evaluations.

Jonathan D. Fife
Director
ERIC Clearinghouse on Higher Education
The George Washington University

Overview

One view of faculty evaluation until the 1970s was that factors other than academic merit influenced promotion, retention, and tenure (PRT) decisions, including the ability to get along and not make waves. A trend of the 1970s that has continued into the 1980s has been for colleges and universities to develop faculty evaluation programs that are more systematic and comprehensive than those of the past. A particular feature of many of these programs is the use of *written explicit* criteria to evaluate faculty.

One explanation for the attention to faculty evaluation that began in the 1970s was changes in the economics of higher education. During the 1960s, when many colleges and universities were expanding, it was all administrators could do to find and keep faculty. In the 1970s, when program retrenchment became a reality, declining enrollments and financial resources plus increasing costs of operation influenced both administrators and faculty to reconsider policies and procedures for making personnel decisions. Related factors that brought attention to faculty evaluation were faculty demands for a greater share in governance and state government demands for accountability.

A manifestation of the increased interest in faculty evaluation was the willingness of systems of higher education to conduct "meta evaluations," that is, to evaluate their methods of evaluation. Meta evaluations conducted during the 1970s in three different regions of the country included those of the Oregon State System of Higher Education, the State University of New York, and the Southern Regional Education Board. A striking feature of these meta evaluations was that common issues were identified.

One set of issues concerns the *purpose* of faculty evaluation. Although many faculty evaluation programs purport to help develop faculty as well as to provide data for PRT decisions, often the reality is that faculty development is paid only lip service. Many faculty believe that faculty development is important, but see faculty evaluation as a negative process unrelated to faculty improvement. Often administrators assume that faculty evaluation automatically leads to improvement because faculty will seek better evaluations. Unfortunately, it does not. Conditions necessary to make that connection include trust between administrators and faculty; faculty involvement in designing and implementing the evaluation program; and educational resources, such as consultation, to accompany evaluation results. Creating these conditions is desirable because it is efficient to use one system of data collection to support faculty evaluation and development programs.

A second set of issues concerns the *areas* to be evaluated. Traditionally, teaching, research, and service are the three areas evaluated; usually, however, little weight is given to service. On the other hand, teaching and research often are seen as competing obligations. A disturbing finding of some studies is that there is wide disagreement within institutions and, sometimes, even within academic departments, concerning the weights that are given to teaching, research, and service. A frustrating finding is that, although many administrators and faculty would like to give more

weight to teaching, the state of the art of evaluating teaching does not instill confidence in the reliability or validity of teacher evaluation.

A third set of issues concerns the *criteria and standards* used to evaluate faculty. The literature of the 1970s and early 1980s reveals strong concerns over the objectivity of faculty evaluation, especially in the area of teaching. One approach to promoting the goal of objectivity is qualitative, i.e., techniques are used to reduce the bias of an individual evaluator. A second approach is quantitative, i.e., data are collected from many sources so that the evaluation does not depend on a single person.

The most common strategy of the qualitative approach is to provide those doing the evaluating and those being evaluated with written explicit criteria and standards. The reasoning for providing evaluators with explicit criteria is that they will focus on the important elements of teaching, research, and service. The reasoning for providing those being evaluated with explicit criteria is that the rules of fair play dictate giving everyone an equal opportunity to succeed.

A controversial criterion in evaluating the area of teaching is student learning. Proponents believe that student learning is the ultimate evidence of effective teaching. Opponents argue that effective teaching and student learning are not necessarily associated. Certainly, much is yet to be learned about the relationship between teaching and learning. The effort to use student learning as one of may criteria to evaluate teaching will increase our understanding of this relationship.

The quantitative approach to objectivity requires using multiple sources of data to evaluate faculty. If there exists one conventional wisdom in the field of faculty evaluation it is that using multiple data sources is desirable. Students have been the most studied source of data. Student rating forms are commonly used to evaluate faculty, and many studies indicate that students constitute a reliable data base. Some bias has been found in student ratings, but not enough to invalidate them. The real validity issue is whether the items placed on student rating forms really characterize effective teaching.

Peer evaluation has been insufficiently studied and there is a lack of understanding of how colleagues are used and should be used. A controversial feature of peer review concerns the use of classroom visitation versus review of teaching materials. Some faculty perceive classroom visits as threatening, negative, and unreliable. Other faculty believe that reviewing teaching materials is too removed from the act of teaching.

The use of self-evaluation is even less studied than the use of peer review. Documentation by faculty of their teaching, research, and service activities seems to hold potential as an additional source of data. The use of teacher dossiers probably will become more common in faculty evaluation programs.

Administrators remain the principal actors in initiating, developing, and implementing faculty evaluation. In most cases, department heads and deans use available data rather than collect their own. A healthy trend would be increased involvement of faculty themselves. Faculty judgment

in developing meaningful criteria and standards is essential to adequate evaluation. A concern for the authors of this monograph is a "crisis of spirit" created by easy-to-measure criteria that do not reflect the important qualities of teaching, research, and service.

The fourth set of issues concerns the *administrative procedures* used to evaluate faculty. The trend is for institutions rather than individual faculty to become responsible for collecting data on faculty performance. Once data are collected, administrators also dominate the review process. However, faculty desire for shared governance and faculty unions' attempts to reduce the power of administrators presage an increase in faculty participation.

Another influence on administrative procedures is the court system. In general, courts have reinforced the requirement that institutions provide written criteria and procedures that guarantee due process. Because of the increased number of litigations initiated by faculty disappointed by personnel decisions, it is imperative that administrators keep up to date with legal requirements.

One noteworthy effort to improve faculty evaluation programs was carried out by the Southern Regional Education Board. According to the evaluation of their faculty evaluation project, the most important characteristics for improving faculty evaluation are active support and involvement of top-level administrators plus faculty involvement.

Although the economic factors that precipitated the examination of faculty evaluation may change, there is little likelihood that there will be a return to the informal methods of faculty evaluation that characterized the pre-1970 era. The concept of fair play, reinforced by the courts and by both administrators and faculty, dictates that institutions make clear the purposes of evaluation and the areas to be evaluated. In particular, for the 1980s, one can expect the use of written explicit criteria to be studied and more commonly used.

Background

In a summation of his ten years of research on the sociology of higher education, Lionel S. Lewis found that the body of evidence indicated that merit was a minor factor in academic advancement. In *Scaling the Ivory Tower*, he contends that factors other than academic performance influenced academic advancement, including the ability to get along and not make waves (1975).

Lewis's view of faculty advancement in the 1970s is supported by Prodgers in his plea for more systematic faculty evaluation (1980). According to his anecdotal account of how assistant professor Z was considered for promotion in the mid-1970s, the department chairman and academic vice president of the college met to discuss Z's past performance. They knew Z had dealt effectively with his departmental duties, had worked on occasion with business and industry, and had published two or three articles. Weighing Z's past performance against their vision of an ideal faculty member, they concluded that they "liked the cut of his jib" (1980, p. 1).

Looking back at these days, Prodgers noted in 1980 that, increasingly, personnel decisions are no longer based on the "cut of one's jib." Rather, the move is toward a systematized and standardized attempt to "measure" the quality of faculty performance (1980, p. 1).

In reviewing the literature, one finds justification that Prodgers is correct: One of the strongest trends in higher education in the 1970s, especially in the second half of the decade, was to examine how faculty were evaluated. Actually, perhaps "reexamination" would be a more accurate characterization because, although interest in faculty evaluation in the 1970s was unprecedented, it was not entirely new. In his study of faculty evaluation, Miller found that interest existed in the 1920s and '30s and again in the late '40s and early '50s (1974, p. 1). However, compared with these earlier periods, interest in faculty evaluation in the 1970s was considerable, particularly in contrast to the 1960s. Miller observed that the relatively low interest in the 1960s probably was "due to the wealth of higher education while expansions in programs and personnel sought to keep pace with growth in enrollment . . ." (1974, p. 1).

The fact is that, during the expansion years of the 1960s, American colleges and universities could "get by" with poorly defined evaluation procedures (Centra 1979). From the college administrator's point of view, the lack of well-defined evaluation procedures was not a problem. Rather, administrators were more concerned with finding and keeping faculty than with evaluating them (Centra 1979). In her study of community colleges during this period of growth, rapid enrollments, and new campuses, Mark noted, "It was all administrators could do to keep colleges fully staffed. The problem was not evaluation, but finding someone to hire" (1977, p. 1).

Also, from a faculty member's point of view, the lack of well-defined evaluation procedures was not a problem. As Mark pointed out, "Prior to the 1970s many, if not most evaluation systems were political, personal, subjective and chaotic—largely ignored by faculty so long as it did not interfere with their teaching or job security" (1977, p. 23).

In contrast to the complacency of the 1960s the use of informal approaches to faculty evaluation was questioned in the 1970s (Smith 1976). The reexamination of how faculty were evaluated was brought about primarily by economic changes. College administrators and faculty members became concerned with faculty evaluation when program retrenchment became a reality. Declining enrollments and financial resources plus increasing costs of operations influenced both administrators and faculty to reconsider the policies and procedures for promotion, retention, and tenure (PRT).

In a useful book designed to help administrators and faculty members develop and maintain systematic faculty evaluation, Miller (1974) linked the increased interest in faculty evaluation to three issues: finance, governance, and accountability.

> *Finance:* "Scarcity of resources means fewer new positions and some existing ones phased out. Making these difficult decisions requires a broad data base, and systematic faculty evaluation can serve as one data base" (p. 3).
> *Governance:* "The faculty is demanding a greater voice in institutional governance, particularly in matters of promotion and tenure. These critical questions must be decided on the soundest data base possible, including evidence of teaching effectiveness from student ratings" (p. 3).
> *Accountability:* "Precise accountability requires some systematic means of gathering, analyzing, and evaluating data, hence demands for improved methods of evaluating faculty performance can be expected—especially from state legislators" (p. 3).

Thus, the increased interest in faculty evaluation in the 1970s can be explained largely by economic factors, and the impetus for more systematic evaluation can be closely linked to the issues of finance, governance, and accountability. One manifestation of the great interest in faculty evaluation during the 1970s was the willingness of *systems* of higher education to assess their own faculty evaluation programs. Daniel L. Stufflebeam, the director of the Evaluation Center at Western Michigan University, recognized that, "Good evaluation requires that evaluation projects themselves be evaluated" (1978, p. 17). He uses the term "meta evaluation" to refer to evaluation of evaluations. Three meta evaluations conducted by higher education systems will be briefly reviewed here to demonstrate evidence of this trend toward self-assessment and to identify the major issues that will be addressed in this research report.

Three Meta Evaluations

Oregon State System of Higher Education

In 1973 the Teaching Research Division of the Oregon State System of Higher Education began a three-year study entitled, "Faculty Teaching: Models for Assessment of Quality." An impetus for the study, supported by a grant from the Fund for the Improvement of Postsecondary Education, was the recognition that, "Faculty evaluation is an ongoing process even if there are no systematic means for making the assessments" (Scott, Thorne, and Beaird 1977, p. 1).

Baseline data were collected in the 1973–74 academic year from four discipline areas in member institutions and from cross sections of institutions stratified by academic rank. Based on these results, a Faculty Perception Questionnaire was used in the 1975–76 academic year that asked faculty to rate 34 factors in terms of their influence in promoting faculty at their institutions, e.g., publications, student ratings, etc. "Coefficients of consensus" were derived for each factor based on the proportion of respondents from a given group who agreed or disagreed over the level of influence.

Based on the ratings of influence and coefficients of consensus, the 34 factors were organized into three clusters: *definitely influential, definitely uninfluential,* and *ambiguous.* The ambiguous cluster identified factors for which there was low consensus regarding their influence. For both college and university faculty, the ambiguous cluster, with 19 factors, was the largest. Thirteen factors were commonly ambiguous to both college and university faculty. An example was "innovative effort in teaching." Six factors were ambiguous to one group, but not the other. For example, "evidence of student learning in courses" was ambiguous to college faculty, but definitely uninfluential to university faculty; "supervision of theses" was ambiguous to university faculty, but definitely uninfluential to college faculty (Scott, Thorne, and Beaird 1977, pp. 10–14).

The research group hypothesized four circumstances that may explain the widespread uncertainty reflected by this high level of ambiguity. In their view, the primary source of uncertainty was that faculty were unaware of their institution's evaluation procedures or criteria. A second possible source was a lack of communication within specific departments. A third source was ambiguity of procedures, guidelines, and policies on specific campuses. Finally, the authors acknowledged the possibility that the questionnaire item could itself have been ambiguous. Nevertheless, they concluded, "The ambiguous group of factors is, without a doubt, unnecessarily large. The size of this cluster results in faculty who are trying to be all things to all people" (Scott, Thorne, and Beaird 1977, p. 18).

State University of New York Faculty Council of Community Colleges

During the summer of 1977 the faculty council sponsored a research project to study the theoretical foundations as well as the applied practices of faculty evaluation. According to the study's author, the project was prompted by several factors: economic, programmatic, administrative, educational, and political (Mark 1977).

Mark's review of theoretical models is noteworthy because a major criticism of faculty evaluation systems is the lack of substantive theory on which to base evaluation (Meeth 1976). Administrators whose faculty believe there is a lack of theory may find it helpful to become familiar with the major models reviewed by Mark. In general, all these models agree that any evaluation system needs to use a variety of data sources to effectively differentiate among faculty members. After studying the current practices in the SUNY system, Mark found that, "All segments of the community college ought to be involved in some way with the evaluation process, but with varying and weighted degrees, depending on the choice of faculty" (Mark 1977, p. 102).

To study the current practices, Mark surveyed 30 institutions and found that 14 had updated faculty evaluation systems that were written and perceived effective by the chief executive officer. Based on her in-depth study of four of these 14 institutions, Mark found that an adversary relationship characterized communications between administrators and faculty members. Instead, there needed to be "an atmosphere of cooperation to discuss what evaluative criteria to use and how to assess them" (1977, p. 108). She emphasized that, "However a program is evaluated, the key element must be establishing criteria" (1977, p. 111).

Southern Regional Education Board

In order to study faculty evaluation practices, the Southern Regional Education Board surveyed its 843 postsecondary institutions in 1975 and conducted numerous in-depth institutional case studies in 1976 and 1977. The economic stimulus for this effort was made clear in the report prepared by the SREB Task Force on Faculty Evaluation and Institutional Rewards: "Evaluating faculty performance for purposes of promotion, tenure and salary increases is of singular importance today because of leveling and declining student enrollments, lack of faculty mobility and increasing financial pressures on institutions" (Moomaw et al. 1977, p. 1).

The survey of all regional postsecondary institutions yielded 536 usable responses, a return rate of 63.6 percent. Institutions were chosen for in-depth case studies based on indication in their survey that they had a systematic approach to faculty evaluation. Case studies were developed from detailed interviews with presidents, deans, department heads, and faculty.

Based on the survey and case studies, the task force found that faculty evaluation tended to be more systematic at doctoral level institutions compared with master's and bachelor's level and two-year colleges. However, at all levels, many institutions were vague about precise criteria, standards, and evidence to be used. There was strong agreement at all types of institutions that instructional activity was the number one area of consideration in evaluating faculty. However, there was little evidence of well-developed procedures.

In addition, the task force found that administrators were both the main decision makers and the main sources of information for these eval-

uations. Few institutions used student data in reliable, consistent, or comparable ways to make personnel decisions, and even faculty colleague data often were collected only on an informal basis.

Four Major Issues

The three self-studies reported here were conducted in the mid-1970s in three different regions of the country: the Northwest, Northeast, and South. A striking feature of these studies is their identification of common issues. These issues will be organized according to the SREB task force framework (Moomaw et al. 1977) and will provide the basic structure for the remainder of this report.

Purpose: What are the desired outcomes of faculty evaluation? In general, studies identify two major outcomes: (a) personnel decisions made regarding promotion, retention, and tenure; and (b) feedback to faculty leading to faculty improvement. The SREB task force found that, although most faculty believe that faculty development and improvement should be the primary reason for faculty evaluation, few examples could be found of institutions using the results of evaluation for that purpose (Moomaw et al. 1977). In a similar vein, the SUNY faculty council study found that the goal of self-improvement, growth, and development received much lip service as a "supposedly" important function of the evaluation process; however, "in practice, there is little evidence that real and meaningful attention is paid to faculty who are in need of help" (Mark 1977, p. 98). This report will examine the purposes of faculty evaluation and under what conditions faculty evaluation can lead to faculty development as well as personnel decisions.

Areas: What functions of faculty activity are to be evaluated? Teaching, research, and service are commonly targeted. The Oregon study asked, "What weights are assigned . . . to teaching, scholarship, and service . . .? Where should a newly appointed faculty member place the majority of his energy if he is intent upon attaining a timely promotion?" (Scott, Thorne, and Beaird 1977, p. 1) The SREB task force observed that, although administrators say that teaching is the most important area of evaluation, procedures for evaluating instruction generally are poorly developed (Moomaw et al. 1977). This report will examine the weight given to areas of evaluation.

Criteria: For each area to be evaluated, what criteria should be used and how specific should the criteria be? For each criterion, what are the standards of attainment? In the Oregon study, Scott, Thorne, and Beaird found a need among all ranks of faculty to understand performance criteria and institutional expectations relative to each area of faculty functioning (1977). Finally, for each standard, what sources of data should be used to show evidence of attainment? The SREB task force found that data on which judgments are made were not gathered systematically or consistently (Moomaw et al. 1977). A major aim of this report will be to examine the trend toward written explicit criteria.

Procedures: What is the sequence of activities for implementing the faculty evaluation program? The SREB task force found that administrators usually initiate and carry out faculty evaluation practices with little faculty involvement (Moomaw et al. 1977). A lack of faculty involvement was noted in the SUNY faculty council study; Mark advised that "Faculty must be involved in the development of any process that is to affect their profes-

sional careers" (1977, p. 101). This report will examine the dominant role now played by administrators.

Purposes of Evaluation

The abundance of literature on faculty evaluation generally identifies two purposes: (1) developing and improving faculty and (2) providing information to make promotion, retention, and tenure (PRT) decisions. However, although this two-fold purpose of faculty evaluation is accepted in theory, whether both objectives are met is not so certain (Moomaw et al. 1977; Prodgers 1980). One critic of faculty evaluation concluded that current methods and practices do not serve well the development of faculty and the reward of excellence (Fincher 1980).

One problem with this two-fold purpose of faculty evaluation is the perceived inherent conflict between faculty development and PRT decision making. For example, Hawley argues that

> *if the purpose of the program is to improve the quality of instruction, faculty members will rightly feel sabotaged when the data are also used in making decisions about tenure and salaries. In the first case, evaluation can be seen as helpful; in the second case, it takes on an adversary tone (1977, p. 39).*

In theory, it makes sense that, if faculty are provided with feedback regarding their deficiencies, they will take action to remedy the deficiencies. However, to support Hawley's point of view, there is little evidence that faculty evaluation improves faculty performance (Rippey 1981).

Some educators contend that the apparent conflict between faculty development and evaluation is not inherent. Rather, the problem has been the failure to recognize that PRT decision making is not an end: It is a means to improve instruction and, hence, provide a better education for students (Rose 1976). In a comprehensive study of the relationship between faculty development and evaluation, Smith recognized that some college administrators and faculty members believe these two functions should be administered as separate programs. However, he argued that faculty development and evaluation should be combined into one program because they share a common goal, improvement of college teaching (Smith 1976).

The authors of this research report believe that the conflict between faculty development and evaluation is not inherent. Faculty evaluation can serve the dual purposes of faculty improvement and PRT decision making if it is accepted that both purposes share the long-range goal of improved instruction and student learning. Collecting two separate data bases for faculty development and faculty evaluation strikes us as inefficient and costly—unnecessarily so. However, the fact that there is little demonstration of faculty improvement resulting from evaluation must be addressed. The question is, under what conditions can evaluation lead to improvement? For example, student ratings alone are not likely to improve

instruction. However, according to a literature review conducted by Levinson and Menges, seven studies support the contention that a combination of student ratings and personal consultation (help with interpretation of ratings, suggestions for improving teaching skills, etc.) favors instructional improvement (1979). In addition, Levinson and Menges found that a combination of self-ratings and student ratings leads to improvement, especially when student ratings are less positive than self-ratings (1979). This condition also was identified by Rippey whose literature review did not overlap Levinson and Menges. An additional condition identified by Rippey was that evalution conducted early in a college course favored instructional improvement because it allows faculty adequate time to make modifications (1981).

At the University of Utah, Department of Family and Community Medicine, conditions of evaluation favorable to improvement deliberately were built into the evaluation of clinical teachers in their family practice teaching rounds. Student ratings were combined with educational consultation and were contrasted to faculty self-ratings; furthermore, data collection was begun early enough in the teaching rounds to give the instructor time to make changes in teaching style and strategies. In a study of this process, Whitman and Schwenk found that faculty evaluation led to faculty improvement (1982).

The debate over the purposes of faculty evaluation will continue into the 1980s. Although the need to make PRT decisions based on comprehensive and systematic evalution is undisputed, the rhetoric about "improvemennt of instruction" will continue, perhaps without resolution (Parramore 1979). Furthermore, there will be an exploration of other uses of faculty evaluation: providing information to students for course selection, allocating teaching resources, and research on teaching (Brandenburg, Braskamp, and Ory 1979; Rippey 1981).

Areas for Evaluation

In higher education, three areas of faculty performance usually identified for evaluation are teaching, research, and service. Service is considered a catch-all category and rarely attracts attention as a bone of contention. According to a survey of university department heads, public and community service is infrequently recognized and rewarded. Moreover, department heads did not believe that service should be a major factor in evaluating faculty (Centra 1977, p. 133).

On the other hand, teaching and research often are seen as competing obligations. Implicit in the teaching-research dichotomy is the widespread belief that many faculty relegate teaching to a second-class status because research is rewarded in the PRT process (Jauch 1976). In general, institutions vary in the weight each area exerts in making PRT decisions (Rippey 1981). In its "Statement on Teacher Evaluation," the AAUP declares that institutions should at least set forth specific expectations regarding teaching, research, and service (AAUP 1975). In fact, many institutions do so. For example, presidents of large universities, especially private ones

such as Stanford University and the University of Chicago, have made institutional statements favoring research over teaching, whereas small colleges emphasize teaching (Miller 1974).

The actual weight given to teaching versus research in a particular institution often is not clear to those who work there. In a study of faculty and their department heads at the University of Missouri-Columbia, Jauch found that two-thirds of the faculty members perceived publication to be more important than teaching most or all of the time. On the other hand, department heads were evenly divided on the question (1976, p. 9).

Is the publish or perish threat a real one? Lewis claims that except in a dozen or so prestigious institutions, the threat is an empty one. Yet, many faculty believe that it is difficult to achieve tenure without publications (1975). That view was supported by Jauch who found that, "Apparently, an individual can be promoted with a good publication record even though his adequacy as a teacher may be in doubt. A good teacher with a poor publication record is at somewhat of a disadvantage" (1976, p. 9). Other studies support the contention that many colleges and universities declare teaching to be a high priority, but award tenure and promotion largely on the basis of publication record (Seldin 1975; Knapper 1978).

A reason for not giving more weight to teaching is that it is difficult to evaluate. This view is typified by the statement, "If only you could give the promotions committee more data about the candidate's teaching we would be glad to use it" (Rippey 1981, p. 24).

One prominent critic of how teaching is evaluated is L. Richard Meeth, who entitled his overview to the *Change Report on Teaching: 2*, "The Stateless Art of Teaching Evaluation." He commented that,

> *Systematic, comprehensive, and valid evaluation of teaching has been an educational problem for many years. It continues to evade educators, although most administrators and legislators desire it as a meaningful way to determine rewards and sanctions for faculty, and most serious teachers seek it as a way of improving their performance and more closely relating what they do to what students learn. Most evaluation of teaching has resulted in unfair and inconclusive distinctions among teachers without establishing reliable or valid relationships between what teachers do and what students learn (1976, p. 3).*

Knapper found it ironic that, at a time when teaching has assumed greater importance from the point of view of the students and the community-at-large, it has not assumed a great importance from the point of view of faculty evaluation (1978). Although student pressures in the 1960s helped to stimulate an examination of teaching practices and brought about the use of student questionnaires to rate faculty teaching performance, serious examination of how to evaluate teaching did not follow until the 1970s. Specifically, attention was given to establishing criteria to measure faculty performance and standards to judge it.

Criteria and Standards

One manifestation of the interest in faculty evaluation in the 1970s was the development of explicit criteria. The need for specific and written criteria on which to evaluate faculty was heightened by the scarcity of financial resources. A typical comment was that losing a valuable faculty member or keeping an unproductive one were errors of such magnitude that it was essential that criteria used in these decisions be as fair and explicit as possible. In fact, for criteria to be fair, they had to be explicit (Grinnel and Kyte 1976, p. 44).

The specific attention to the development of explicit criteria also can be explained by additional factors that are related to the "no growth" environment. First, academic unions, which grew in number and strength in the 1970s, became concerned with spelling out the conditions under which faculty receive tenure (Ladd and Lipset 1973). Second, aspiring faculty who have been denied promotion have returned to the courts for redress. In general, the courts expect institutions to publish criteria for making PRT decisions (Centra 1979, p. 141).

In 1971, Wolff published a study of criteria used for faculty promotion in college and university speech departments. The study is noteworthy because it still reflected the growth period of the 1960s. Wolff mailed a faculty questionnaire on promotion to 200 speech department chairpersons of randomly selected colleges and universities. Based on a 58 percent response rate, she found that criteria for faculty promotion in order of importance were (1) teaching effectiveness, (2) academic degrees, (3) publication, (4) extracurricular speech activities, (5) research, (6) committee involvement with school development, and (7) scholarly activities (Wolff 1971). Emblematic of the primitive state of faculty evaluation at the time was the general nature of these criteria. In fact, these are not much more detailed than the three areas of faculty evaluation: research, teaching, and service.

Also, the ease of promotion during a period of growth was reflected by comments made by departments on the survey (Wolff 1971, p. 283):

- "Promotions can be granted to an individual who just stays around and does an adequate job."
- "Tenure is automatic unless teaching effectiveness or notorious conduct leads to uncontestable dismissal."
- "Our speech department has all top-ranking faculty."

If Wolff's study was emblematic of evaluation during the period of growth, the study published in 1972 by Schulman and Trudell symbolized the changing environment. In anticipation of a law passed by the California Assembly in 1971 requiring public school and community college teachers be evaluated at least once every two years (Senate Bill 696), the Innovations Committee of Los Angeles Pierce College studied guidelines for evaluation that would be acceptable to those affected by the new law (Schulman and Trudell 1972).

The committee surveyed the literature on evaluation (they found that more than 2,000 studies of teacher evaluation had been made since 1900), submitted a pilot questionnaire to instructors and administrators in the Los Angeles Community College District, and mailed a revised questionnaire to instructors and administrators within the 94 public community colleges in California. Based on responses from more than 60 percent of the questionnaires, representing about 70 percent of the community colleges, the committee found that criteria for evaluation of teaching were, perhaps, the most troublesome aspects of faculty evaluation (1972, p. 34).

For example, there was little agreement about how to measure teaching effectiveness with objectivity. The committee's recommendation was to admit the subjectivity of measuring teaching effectiveness and to select criteria that can be utilized in as nonsubjective a manner as possible. According to the committee, examples of specific criteria that described teaching effectiveness included (1) ability to relate to students, (2) ability to arouse interest, (3) friendliness, (4) empathy, and (5) knowledge of subject matter. To implement these criteria as objectively as possible, the committee suggested that classroom visits, if used, should be made by judges who are most competent to determine effectiveness, for example, department or division colleagues. Also, when students are given evaluation forms to complete on their instructors, they should be instructed as to the nature of their task and cautioned against emotional judgments, pro or con (Schulman and Trudell 1972).

The problem with subjectivity also was addressed in an evaluation plan implemented at New River Community College in Dublin, Virginia (McCarter 1974). In describing the New River program, McCarter stated:

> *Frequently, an expressed goal of instructional evaluation is to achieve objectivity during the process. That this is to any degree possible is at least a doubtful proposition. Even so, it need not deter a school or college from attempting a creditable faculty evaluative system (1974, p. 32).*

To promote the goal of objectivity, McCarter recommended the use of *collective* judgments by students, peers, and supervisors. By using the judgments of many, the subjectivity of the individual would be minimized (1974).

The suggestion in the California study that evaluators should strive to be as nonsubjective as possible and the suggestion in the New River plan that the collection of judgments be as comprehensive as possible represent two approaches to dealing with subjectivity in evaluating instruction. House characterizes these approaches to objectivity as qualitative versus quantitative. The qualitative sense of objectivity refers to the quality of an observation regardless of the number of people making it. Being objective means that the observation is factual, but being subjective means that is biased. The quantitative sense of objectivity refers to the number of people making the observation. One person's opinion is regarded as subjective; whereas, objectivity is achieved through the experience of many

observers (House 1980). For the most part, faculty evaluation programs attempt to increase objectivity through both qualitative and quantitative approaches: To achieve *qualitative* objectivity, criteria are developed to improve the quality of data collected from an individual evaluator. To achieve *quantitative* objectivity, data are collected from multiple data sources.

Qualitative objectivity. Many people can provide data about faculty performance: students, colleagues, administrators, and faculty themselves. The major approach to improving the quality of data from these persons is to provide them with criteria that are specific and written. The rationale is that by providing specific behaviors, features, measures, or indicators to be examined in the areas of teaching, research, and service, the persons doing the evaluating will know what to assess. By providing evaluators with criteria, it is hoped that evaluation will be fair, relevant, and appropriately focused.

This approach has been reinforced by court decisions. For example, in the case of Harkless V. Sweeny Independent School District of Sweeny, Texas (11 FEP 1005, 1075), it was pointed out that objectivity in faculty evaluation could be achieved by adhering to the following three guidelines (Balch 1980):

> *1. The language in the evaluation instrument used to describe each characteristic to be measured must be composed of words which are reasonably precise and uniform in meaning.*
> *2. There must be a fairly specific standard of measurement to guide the evaluator in ascribing a particular value to a particular characteristic.*
> *3. There must be a reasonably well-defined system for assigning relative weight to the characteristics measured (p. 4).*

The problem with this approach is that researchers disagree as to whether there is a well-defined set of criteria for judging faculty performance (Tuckman and Hagemann 1976). Whereas Johnson and Stafford claim that the faculty reward structure is determined by rational criteria (1974), others argue that no acceptable criteria have been developed (Batista 1976) or that administrators and faculty are using different sets of criteria (Meany and Ruetz 1972).

The difficulty of agreeing on criteria was cited in a study of faculty evaluation in Ph.D. graduate departments of sociology. According to Gaston, Lantz, and Snyder, "The role of all criteria for promotion (publication, good teaching, and service) remains unclear in the actual promotional decision" (1975, p. 242). In a study of written criteria used by graduate schools of social work during the 1974–75 academic year, Grinnel and Kyte found that, although most schools had carefully defined *procedures* to evaluate faculty, they lacked specific, objective *criteria* on which to base these evaluations. However, as evidence of a trend toward the use

of written criteria, 64 out of 72 responding schools reported they were either in the process of working on new written criteria or anticipated doing so. (1976, p. 44).

A major difficulty in developing criteria is in the domain of judging the quality of work. For example, in the area of service, it is relatively easy to document participation. However, merely being involved in public or community service is not a sufficient indicator of effectiveness (Centra 1979). Similarly, in the area of research, it is relatively easy to establish a number of required publications, however it is difficult to produce explicit criteria to judge the quality of published works (Gaston, Lantz, and Snyder 1975).

There are examples of faculty on promotion committees critically evaluating the quality of published work, but the criteria used were not preestablished. For example, a dean's ad hoc tenure review committee at Pennsylvania State University, upon denying tenure for a faculty member, read the person's published work and found two studies "deficient in design, methods, implementation, and, in the case of one, in conclusions" (Balch 1980, p. 8).

The need for specific criteria in the area of teaching was recognized by Spencer, Crow, and Glass, who reported the work conducted by an ad hoc committee of the Cornell University Medical College Department of Psychiatry during the 1977–78 academic year. The authors found two major difficulties with evaluating teaching: (1) the absence of a single, concrete end product such as the published results of research and (2) problems with reliability and validity that seem to accompany any attempt to measure teaching effectiveness (1979). Others also cite the difficulties colleges and universities have with developing criteria in the area of teaching (Meeth 1976; Miller 1974).

One technique to develop criteria for teaching is to design evaluation forms that list the items teachers, students, and administrators deem important. Berk (1979) and Wotruba and Wright (1975) present easy-to-follow methodologies to design such a form, and Fenker describes how a teacher evaluation form was designed at Texas Christian University (1975). In addition, Arreola describes how an instrument developed at Michigan State University was adapted by Florida State University (1973).

According to a literature review conducted by Dwyer in 1973, teacher evaluation forms were being used extensively by colleges and universities in the United States as a means to evaluate teaching effectiveness. However, he found that what was lacking was evidence that the characteristics listed on these forms made any real difference in the achievement of educational objectives by students (1973). This criticism of rating forms was supported by Meeth:

> *If we better understood how students learn, we might better understand how and what teachers ought to teach. The many lists of teaching activities prepared over the years . . . cannot be ranked very conclusively from most to least important in terms of producing learning (1976, p. 3).*

In order to clarify the types of criteria that could be developed to more conclusively evaluate teaching, Meeth (1976) adapted Thorndike's categories of criteria to teaching effectiveness. "Immediate" criteria are the lists of teaching behaviors that people believe are related to teaching effectiveness, e.g., lecture style was conversational, audio-visual aids were reinforcing, etc. Although these are better than nothing, Meeth complains that immediate criteria are furthest from learning outcomes and are a long way from relating teaching to learning.

Closer to learning outcomes are "intermediate" criteria, which describe the process of teaching:

- *Students were motivated to learn*
- *The structure of the learning-experience was determined by the goals of the experience.*
- *The content was well ordered, comprehensive, and appropriate to the abilities of the learners.*
- *Rewards and sanctions were appropriate to the goals of the learning experience.*
- *Goals and/or outcomes were clearly specified.*
- *Evaluation criteria, standards, and methodologies were clear and appropriate to the goals of the experience.*
- *Methodology was appropriate to the goals of the experience and the abilities of the learners. (Meeth 1976, p. 4)*

Closest to learning outcomes are "ultimate" criteria, which describe what students learned:

- *The students learned what the instructor was trying to teach in cognitive, affective, and/or psychomotor development in rate and/or absolute achievement.*
- *Students retained what was learned.*
- *Teacher goals and/or outcomes for the learning experience were met.*
- *Student goals and/or outcomes for the learning experience were met. (Meeth 1976, p. 4)*

Some educators do not favor using student achievement (ultimate criteria) as a means to measure teacher effectiveness because of differences in the difficulty of instructional objectives, difficulty with measuring some instructional objectives, and the potential abuse by instructors who "teach to the test" (McCarter 1974, p. 32). Proponents of using student achievement argue that, when only the process of teaching is measured (intermediate criteria), only half the evaluation process is really accomplished (Mark 1977, p. 104).

Including student learning as one of the criteria to be used in evaluating teaching is one of the most controversial issues in the field of faculty evaluation. Murray acknowledged that to say that the best teacher is one whose students learn the most has intuitive appeal. However, he warns

that, although it is easy to agree with a statement like that, it is almost impossible to put it into action (Murray 1979). The difficulty of using student learning as a criterion to measure teaching effectiveness also was acknowledged by the Special Interest Group on Instructional Evaluation at the 1977 Annual Meeting of the American Educational Research Association, which rejected its use to evaluate the effectiveness of instruction or instructors. (Darr 1977).

Nevertheless, including student learning as one of the criteria to be used in evaluating teaching makes sense to the authors of this report. The rationale described by Martin for assessing teaching methodologies aptly justifies why it is necessary to seek out how to use student learning as a criterion of teaching effectiveness.

> *So the teacher chooses—subject matter, points of emphasis within the discipline, in other words,* what *will be taught; the teacher chooses the methodology of this inquiry, its strategy and tactics, in other words,* how *to proceed; the teacher chooses the timing, the sequences, the specific chronology of events, in other words,* when *things will come together to form the basis for choice; and, finally, the teacher chooses the gut questions,* why? *and* so what? *These are the questions that figure in the conclusions and inferences for action.*
>
> *The teacher chooses and the teacher acts, and, working with the student, helps the student develop a capacity for choice and action. Our commitment to this skill, to this service, needs to be kept in mind as we assess the methodologies of the teaching profession (1981, p. 60).*

Quantitative objectivity. Achieving qualitative objectivity has been discussed in terms of providing explicit criteria to those who evaluate faculty. Because of the difficulties with developing valid criteria and doubts over the reliability of individual evaluators, multiple sources of data often are used. Using multiple data sources constitutes a *quantitative* approach to the problem of objectivity. The Southern Regional Education Board, among others, has recognized the need for this quantitative approach:

> *A system for evaluation [should] include provisions for collecting data from many sources and recommendations from multiple participants, since decisions made even in the most carefully conceived systems of evaluation will still largely depend upon a collection of subjective statements (Mooman et al. 1977, p. 7).*

In fact, if there is one conventional wisdom in the field of faculty evaluation, it is that multiple sources of data are preferrable to one source (Batista 1976; Centra 1979; Darr 1977; Goldschmid 1978; O'Hanlon and Mortensen 1980). Opinions differ over how to use them.

Students. There has been a considerable amount of research interest and effort regarding the use of student rating forms since the first formal form

(the Purdue Rating Scale of Instruction) was published in 1926 (Darr 1977). One reason for studying the use of student ratings is that this method of evaluation is used more frequently than any other. According to one survey, approximately 68 percent of universities in North America use student ratings (Bejar 1975). The rationale for using student ratings is that, since it is difficult to attribute student learning to the skills of teachers, the next best thing is to ask students to rate characteristics of teachers that one would logically expect to be determinants of student learning (Murray 1979). For the most part, faculty members believe that student ratings should be used as one of several sources of information in making PRT decisions (Goldenstein and Anderson 1977).

One focus of research regarding student ratings has been their reliability. To what extent are ratings consistent or dependable for a given teacher? One way of looking at reliability is to study inter-item consistency, i.e., if six items on a rating form are supposed to measure the same *aspect* of teaching, is there a high average correlation among the six items? In general, studies of inter-item consistency demonstrate high average correlation coefficients. In other words, if students rate a teacher high on one item, they usually will rate him or her high on other items intended to measure the same characteristic (Murray 1979, p. 9).

Another way of looking at reliability is to study inter-rater consistency, i.e., do students agree with one another in the ratings they give a teacher? In general, inter-rater reliability is high, particularly when there are 15 students or more. With less then 15 students, inter-rater reliability drops off considerably, and, with less than 10 students, it is probably unwise to use student ratings (Centra 1973).

A third way of looking at reliability is to study test-retest consistency, i.e., are ratings similar at two points in the same course or same type of course? In general, test-retest reliability is high. Teachers who receive a high rating in the middle of a course are likely to receive a high rating at the end of the course. Likewise, teachers who receive a high rating in a course are likely to receive a high rating when teaching the same or a similar course again (Murray 1979, p. 12).

On the other hand, in general, there is low reliability across different types of courses. For example, ratings for teaching a large introductory lecture course and for teaching an upper-class seminar may be unrelated. One implication of the low correlation of ratings across different types of courses is that student ratings used in PRT decisions require a good sampling from different types of courses because a high or low rating in one type of course cannot be used reliably to judge a faculty member's teaching skills (Murray 1979, p. 13).

Although most faculty favor inclusion of student ratings for faculty evaluation, some question the extent to which extraneous course, student, and instructor characteristics influence such ratings (Brandenburg, Braskamp, and Ory 1979). In a review of studies Murray (1979) found:

- Students in larger classes gave lower ratings.

- Teachers who assign low grades tend to receive lower ratings.
- Classes that meet at mid-day tend to receive lower ratings.
- Ratings on days when attendance is very high or very low tend to be high.

According to Murray's analysis of the studies of student bias, bias factors, although statistically significant; are not large enough to single-handedly invalidate student ratings as a measure of teaching effectiveness (1979, p. 24).

In addition to student bias, another objection to using student ratings in PRT decisions is that they are not valid measures of teaching effectiveness (Brandenburg, Braskamp, and Ory 1979). In other words, some teachers with high student ratings may actually be associated with low levels of student learning, and other teachers with low student ratings may actually be associated with high levels of students learning! A study often cited to support the invalidity of student ratings is known as the "Dr. Fox" study. In this study, an actor was trained to lecture *charismatically but nonsubstantively* on a topic he knew nothing about, "Mathematical Game Theory as Applied to Physician Education." The actor, introduced as Dr. Myron L. Fox to a group of psychiatrists, psychologists, and social workers, had been coached to use double talk, non sequitors, and contradictory statements. In general, those who attended the live lecture and those who viewed a videotape of the lecture rated Dr. Fox as a good teacher: He seemed interested in the subject; he used enough examples to clarify his material; he presented the material in a well organized form; he stimulated thinking, etc. (Naftulin, Ware, and Donnelly 1973).

One conclusion of this study is that students, even if they are professional educators, can be "seduced" into thinking that a teacher is good. This "Dr. Fox effect" has been replicated in a series of similar studies (Ware and Williams 1975; Kane and Schorow 1977; Ramagli and Greenwood 1980). The counter-argument to the "Dr. Fox effect" is that, although student ratings are not sensitive to content differences under some circumstances, such as a single-episode guest speaker, real teachers in real classrooms cannot fool students into thinking they have learned when they have not.

The fact that student ratings can be biased by irrelevant factors and that student ratings are not consistently correlated to actual learning highlights the need for multiple data sources. In recommending why student ratings should be used in conjunction with other measures, Sheehan pointed out that

> *Administrators should make use of this information without forgetting that classificatory errors can result because of the imperfect validity of the ratings. Until instrumentation is improved, the strategy of administrators should be one of collecting as much information from as many sources as possible (1975, p. 697).*

Faculty colleagues. At most colleges and universities, colleagues play an important role in evaluating faculty. In particular, faculty on promotion committees make judgments regarding the quantity and quality of research and service. In addition, faculty are asked to rate each other's teaching based on classroom visits and review of teaching materials (Centra 1979). However, in contrast to the literature on student ratings of faculty, there seems to be a dearth of literature on the use of colleague evaluation (Batista 1976; Darr 1977). Most studies of colleague evaluation compare their ratings of teaching to those made by students or administrators. For example, Blackburn and Clark studied the colleague, student, and administrator ratings of teacher effectiveness for 45 faculty members in a midwest college and found that the ratings were significantly correlated (1975, p. 247). Similar results had been documented by Murray (1972).

For the most part, the method of colleague evaluation is similar to that used for student evaluation: Faculty are rated on items deemed important to good teaching. In fact, Nadeau (1977) suggested the students and faculty use the same rating forms. Hildebrand, Wilson, and Dienst (1971) described how faculty developed their own rating form.

Although less is known about colleague evaluation than student evaluation, it is suspected that there are problems with the reliability of peer evaluators. Often, colleagues base their judgments about the quality of teaching, research, and service on overall impressions rather than direct observation. Furthermore, these impressions may be biased by departmental jealousies and rivalries (Batista 1976, p. 261). The importance of getting along and not making waves was one of Lewis's major themes in *Scaling the Ivory Tower* (1975), and the presence of bias was underscored by Mark in her SUNY case study: "Personal biases are present and must be understood and not allowed to influence the evaluation of a colleague whose style, philosophy and manner of presentation differs from the evaluators'" (1977, p. 102).

In addition to problems with reliability, colleague evaluation is compromised by the same problems with validity experienced with student evaluation: Popularity with students and peers is not necessarily related to good teaching, and high ratings are not necessarily associated with learning outcomes. The main problem perceived by Batista is that research in the area of colleague evaluation has not been dealt with systematically. He has two recommendations for research: (1) to develop adequate instruments and (2) to study interaction between the characteristics of the evaluator and the characteristics of the person being evaluated (Batista 1976, p. 264).

The lack of understanding of how colleague evaluation works and should work was also cited by French-Lazovik (1981), who warned that, because considerable progress was made during the 1970s to improve the quality of student data, college administrators may believe that other data on teaching effectiveness are not needed. In one of the most detailed descriptions of how colleague evaluation should work, she points out that faculty

peer review is an essential data base because faculty peers are uniquely qualified to judge the *substance* of teaching.

The approach recommended by French-Lazovik is for faculty to maintain dossiers on their teaching, research, and service. For example, a dossier on teaching should include a brief and objective description of each course taught, its objectives, enrollment, credit hours, etc. In a valuable contribution to the field of colleague evaluation, French-Lazovik has designed a form (see Appendix A) for peer evaluators to use when studying faculty dossiers.

The thrust of French-Lazovik's position is that there is a need to evaluate aspects of teaching that can only be judged by other faculty. Referring back to Meeth's adaptation of Thorndike's categories of criteria, perhaps students are a suitable data base to evaluate immediate criteria, i.e., to assess teaching behaviors that are believed to be related to teaching effectiveness. On the other hand, perhaps faculty peers are a suitable data base to evaluate intermediate criteria, i.e., to assess the process of teaching.

The need to make better use of peer evaluation was emphasized by Batista (1976), who pointed out that colleagues are in better position to evaluate certain faculty behaviors than are students or administrators. Teacher behaviors that Batista contends cannot be validly evaluated by students or administrators include:

1. *Up-to-date knowledge of subject matter.*
2. *Quality of research.*
3. *Quality of publications and papers.*
4. *Knowledge of what must be taught.*
5. *Knowledge and application of the most appropriate or most adequate methodology for teaching specific content areas.*
6. *Knowledge and application of adequate evaluative techniques for the objectives of his/her course(s).*
7. *Professional behavior according to current ethical standards.*
8. *Institutional and community services.*
9. *Personal and professional attributes.*
10. *Attitude toward and commitment to colleagues, students, and the institution (p. 269).*

Thus, although there is a need to know more about how to reliably use peer review, a more systematic utilization of colleague evaluations ultimately will provide a more valid evaluation of faculty.

Self-evaluation. In comparison with student ratings, there has been little study of peer evaluation. However, there has been even less written about self-evaluation (Darr 1977). One approach to self-evaluation is for faculty to rate themselves on written scales similar or identical to those used by students. According to Blackburn and Clark (1975), there is a low correlation between student ratings and self-ratings; in general, faculty rate themselves higher. Consequently self-ratings rarely are used for PRT de-

cision making. However, self-ratings are recommended for the purpose of improvement. For example, Whitman recommends using a discrepancy between student ratings and self-ratings as a kernel of a problem for the teacher to solve, leading to nonrandom attempts to improve instruction (1981).

In addition to completing rating forms, another approach to self-evaluation is for faculty to describe their academic efforts. For example, with regard to teaching efforts, self-evaluation would include a description of the faculty member's approaches to teaching, problems with teaching, and efforts to improve. We recommend that "effort to improve teaching" be included as a criterion of teaching effectiveness. Thus, documentation of how faculty assess their own needs and implement plans to meet these needs could be used as a source of data in evaluating faculty. For example, as a "principle of sound evaluation " O'Hanlon and Mortensen suggest that:

> *The total evaluation of faculty members should include consideration of what they are doing for their own development, including attendance at workshops, redevelopment of teaching materials, trying new approaches, and seeking help from colleagues and instructional consultants. These considerations should include how the teacher is profiting from evaluations received from students and others (1980, p. 666).*

In addition to documenting improvement of teaching, we recommend documenting research and service improvement. We justify this data source on the basis that, by definition, a good faculty member is one who seeks to improve performance of teaching, research, and service at any current level of performance.

Administrative evaluation. Virtually all faculty evaluations conducted for the purpose of PRT decision making use evaluation by administrators, since department heads and deans are usually involved in making personnel decisions. However, rather than generating their own data, administrators tend to evaluate faculty based on student and colleague data sources already collected (Darr 1977). When it is possible to judge from research studies, ratings by administrators tend to be the same as ratings by colleagues (Ericksen and Kulik 1974, p. 3).

Because of the time involved in becoming familiar with an individual faculty member's teaching, research, and service efforts, it is unlikely that administrators will personally evaluate faculty except in small institutions (Darr 1977; O'Hanlon and Mortensen 1980). Thus, most attention to administrative evaluation is placed on *how* administrators use available data sources. For example, at Franklin and Marshall, a small liberal arts college in Lancaster, Pennsylvania, a faculty member's department head and dean evaluate teaching by reviewing (1) evaluations from all students in all the teacher's courses, (2) exit interviews with department seniors, (3) "grapevine" feedback from students, (4) course syllabi, and, some-

times, (5) observations of classroom teaching. Based on these data sources, the department head and dean independently rate the faculty member on an ordinal scale:

- *Rate a 0 on this criterion if, on the basis of the evidence, it can be said that the faculty member was below average on all measures and counts.*
- *Rate a 1 on this criterion if, on the basis of the evidence, it can be said that the faculty member was below average, taking all counts and measures as a whole even though on some measures or counts he or she may have been above average.*
- *Rate a 2 on this criterion if, on the basis of the evidence, it can be said that the faculty member was average, taking all counts and measures as a whole.*
- *Rate a 3 on this criterion if, on the basis of the evidence, it can be said that the faculty member was above average, taking all counts and measures as a whole.*
- *Rate a 4 on this criterion if, on the basis of the evidence, it can be said that the faculty member was above average on all counts and measures.*
- *Rate a 5 on this criterion if, on the basis of the evidence, it can be said that the faculty member was clearly excellent on all counts and measures (Michalak and Friedrich 1981, pp. 586–87).*

After the department head and dean confer and reconcile any differences, the ratings are reported to the faculty member, who can appeal a rating to the dean. The dean makes the final decision.

Their approach to evaluating a faculty member's scholarship is similar. The department head and dean independently rate scholarship using another ordinal scale:

- *Rate a 0 on this criterion if the faculty member has, during the past year, (1) had no publications and (2) had no systematic program of research and study.*
- *Rate a 1 on this criterion if the faculty member has, during the past year, (1) published a book review or its equivalent or (2) pursued a systematic program of research and study leading toward further publication or the presentation of a new course.*
- *Rate a 2 on this criterion if the faculty member has, during the past year, displayed activity in scholarship by having (1) published an article or equivalent series of book reviews or subsidized studies and (2) pursued a systematic program of research and study leading toward further publication or the presentation of a new course.*
- *Rate a 3 on this criterion if the faculty member has, during the past year, displayed good scholarship by having (1) published one or two high-quality articles or edited an anthology or book of readings and (2) pursued a systematic program of research and study leading toward publication or the presentation of a new course.*
- *Rate a 4 on this criterion if the faculty member has, during the past*

year, displayed excellence in scholarship by having (1) pursued a systematic research and study program leading toward further publication or the presentation of a new course and (2) published a book or equivalent of articles and/or monographs, or the equivalent in the fine arts; or (instead of 2) (3) devised a set of procedures or syllabus that could be expected to affect the teaching of the discipline in first-rate colleges and universities.

- *Rate a 5 on this criterion if the faculty member has, during the past year, displayed outstanding excellence in scholarship by having (1) pursued a systematic research and study program leading toward further publication or the presentation of a new course and (2) authored a high-quality book or an equivalent set of articles and/or monographs, or the equivalent in the fine arts; or (instead of 2) (3) devised a set of procedures or syllabus that can be expected to substantially change the teaching of the discipline in first-rate colleges and universities. (Michalak and Friedrich 1981, pp. 584–85)*

Michalak and Friedrich admit the subjectivity of the measures. However, they contend that the use of an original rating scale plus multiple data sources enhance reliability and validity. Clearly, their approach attempts to promote objectivity by both qualitative and quantitative means. In other words, qualitative objectivity is enhanced by a technique (the ordinal rating scale) to improve evaluations conducted by individual department heads and deans; quantitative objectivity is enhanced by using multiple sources of data rather than relying on observations of a single party. Although the procedures used at Franklin and Marshall reflect the administrative routines of that institution and may be difficult to replicate elsewhere, we recommend the technique of providing administrators with descriptive scales and using multiple raters.

Having considered the purposes of evaluation (faculty improvement and PRT decision making), the areas to be evaluated (teaching, research, and service) and criteria (explicit and written) to assess these areas of performance, the fourth major issue to be addressed concerns *procedure*, i.e., the sequence of activities for implementing a faculty evaluation program.

Administrative Procedures

A critical issue in faculty evaluation is determining how data are collected and reviewed. One approach is for individual faculty to bear the "burden of proof." In other words, at some institutions, faculty are expected to provide evidence of their teaching, research, and service effectiveness. Prior to the 1970s, this was the predominant approach to data collection. Its major disadvantage is that data collection is not standardized and often what is evaluated is not faculty performance in the areas of teaching, research, and service, but rather *faculty skill in collecting and presenting data in a favorable light*. Since the 1970s, there has been a trend toward systematic, standard data collection. Increasingly, institutions are spelling out what data faculty should collect regarding their own performance,

often in the form of teachers' dossiers. Moreover, institutions are spelling out what data the institution will routinely collect regarding faculty performance. The major advantage is that faculty members know in advance what their responsibilities are (Centra 1979).

The shift in responsiblity from those who are going to be evaluated to those who are going to be doing the evaluating is reflected in Eble's recommendations to department heads with respect to PRT decisions:

> *1. Keep careful records of what each member of a department does as a teacher, quarter by quarter, year by year.*
> *2. At tenure time or time of other important reviews, reduce these data to an easily grasped form and place them in the hands of everyone involved in the review.*
> *3. Most of all, say a great deal about teaching* before *tenure time. If you wait to speak, it will always be too late (1978, p. 30).*

On almost every aspect of promotion, retention, and tenure, institutional policies and practices vary. Although some colleges and universities may still use informal procedures to make PRT decisions, the trend toward more systematic evaluation has included a formalization of faculty personnel policies and procedures.

In the academic profession, as in other professions, members of the profession make personnel decisions. In most colleges and universities, it is senior faculty who help make PRT decisions; although, in some cases even junior faculty are represented in the process. While practices vary, often faculty committees make recommendations to administrative officers, e.g., department heads, deans, academic vice presidents, and presidents. (Commission on Academic Tenure 1973).

Once the data are collected, the review process, for the most part, is dominated by administrators. Given the importance of evaluation decisions to individual faculty, Moomaw was surprised that faculty rarely play a substantial role in the functioning of faculty evaluation programs. The Southern Regional Education Board survey of assignment of principal evaluation responsibility for decisions on salary, promotion, and tenure demonstrated conclusively that the academic dean and department chairperson are the two most important persons. The department chairperson was more important in doctoral and master's legree institutions and the academic dean was more important at bachelor's degree and two-year institutions (Moomaw et al. 1977).

There are two views regarding the dominant role played by administrators in evaluating faculty. According to one view, evaluating faculty for making PRT decisions is an appropriate responsibility for administrators. Defending the view that PRT decision making is an administrative responsibility, the dean of one New England college of education commented, "There is no set of data or procedure for gathering it and no review process that can substitute for our [deans'] professional judgment" (Philippi 1979, p. 9).

Philippi acknowledges the difficulty of evaluating faculty, but does not back away from the responsibility:

> *In a university the academic administrator is a double agent, serving as an agent of the very faculty he is supposed to evaluate as the agent of the university . . . If an academic administrator is incapable of consistent, sound judgment, termination of that administrator would alone improve personnel practices in the institution (1979, p. 10).*

The view that administrators are responsible for evaluating faculty does not exclude faculty from participating in developing and evaluating the process. In fact, faculty involvement in developing the evaluation program and critiquing it is explicity recommended by Mark (1977).

According to a second view of the role played by administrators, administrators should have less control. Notably, academic unions generally try to reduce or eliminate the power of administrators to reward faculty. For example, unions have sought to have new appointments defined as "probationary," which implies a claim to permanency for faculty who can demonstrate that they can handle the job (Ladd and Lipset 1973, p. 72). In general, where collective bargaining exists, due process for faculty being evaluated is spelled out, and faculty committees are playing an increased role in recommending personnel decisions and listening to appeals. Nevertheless, administrators continue to play dominant roles in the process.

The need for due process is highlighted by the decisions of courts. In a review of court cases and their implications (which should be read by all administrators and faculty concerned with faculty evaluation), Balch documents that many courts have hesitated to take a role in the decision-making process of faculty evaluation. For the most part, courts defer to the expertise of administrators and faculty to evaluate faculty. However, she notes a rise in the number of cases brought to court and the willingness of courts to become more involved than they used to be. Balch attributes the rise in the number of cases brought to court to the financial retrenchment in higher education. As relocation becomes more difficult for faculty, the willingness to fight negative personnel decisions through legal channels becomes more attractive (1980).

The willingness of courts to become more involved can be attributed to the notion of "state action" in private institutions. In other words, some courts view private colleges and universities that receive large amounts of federal and state funding as public institutions. Consequently, the Fourteenth Amendment, which provides that the state shall not deprive any person of life, liberty, or property without due process of law, may apply to private as well as public institutions. Thus, at a minimum, colleges and universities should comply with due process in making PRT decisions, i.e., provide faculty with proper notice and an opportunity for a fair hearing. Also, although courts show little or no interest in the specific criteria included or evaluation methods used, they do expect criteria and methods to be published (Centra 1979).

As a result of her review of legal actions, Balch recommended that *administrators* should:

1. become knowledgeable concerning the ever-changing legal obligations and rights of public/private institutions toward faculty evaluation.

2. be certain of his/her role as an administrator in carrying out the evaluation policies and practices of the institution.

3. make certain that the evaluation process of each faculty member is job-related.

4. be sure that the faculty evaluations are not discriminatory in intent, application, or results.

5. be certain that evaluation forms contain precise and uniform language and should be statistically valid.

6. guarantee that evaluators at the institution are trained in how to use and analyze evaluation instruments.

7. provide for the performance evaluation process to include the appropriate variety of represented groups: students, faculty peers within and without the department, and administrators.

8. insist that performance evaluation procedures be conducted in entirety before making any changes in personnel decisions.

9. inform faculty members in writing of the results of their performance evaluation.

10. see that their institution develops thorough written policies pertaining to the use of faculty evaluation, procedures for administration, and various rules which may govern any decisions rendered.

11. make certain that these policies are communicated to all newly hired faculty members before *they sign their first contract so that both parties fully understand the entire evaluation process.*

12. provide for consistency of standards and procedures of faculty evaluation.

13. work for improved administrative-faculty communications to keep evaluation procedures "above board."

14.create a sense of fairness in facing evaluation problems.

15. not take rash action to mere "hear-say" of other faculty members.

16. check on current insurance policies for maximum coverage permitted by law (for possible cases of administrative liability in evaluation).

17. employ legal counsel who has a good knowledge of the institution, its organizational structure, policies, and goals.

18. have this legal counsel keep the administrative staff and faculty, as well as students, current on their rights in the entire evaluation picture (1980, pp. 38–39).

In addition, Balch recommends that *faculty* should:

1. be aware of and get to know the next highest person in administrative authority. When a problem arises, contact this person first.

2. try to have a third neutral party present when dealing with a "hot" issue with either students or administration.

3. keep and maintain written memoranda of conferences and/or telephone conversations.

4. be aware that nothing can be assumed to be confidential if told to a student or a colleague.

5. keep up to date written personal records of all academic accomplishments, service, and honors.

6. utilize the student evaluation process of the institution and for added strength, design self-evaluation forms for classes to respond to other types of questions.

7. keep all records from the time of hiring (contracts, faculty handbook, catalogs, etc.) in a chronological file.

8. not attack verbally and publicly the department chairperson, dean, or president of the institution.

9. keep current about new laws, rules, regulations, or policies which might affect the teaching position.

10. try to remain out of the "losing" categories such as "immoral, behaviorally undesirable, or incompetent" (1980, pp. 37–38).

There are pressures on colleges and universities to develop administrative procedures acceptable to faculty and their union representatives and to implement a faculty evaluation program consistent with the requirements of due process. These pressures point to as much faculty involvement as possible in designing, implementing, and critiquing the evaluation process. In response to these pressures, the Southern Regional Education Board initiated its faculty evaluation project in 1977 to help member institutions design, revise, or critique their faculty evaluation programs. This regional approach is worth noting because the major reasons for its success may be applicable elsewhere.

SREB faculty evaluation project. SREB's faculty evaluation project was an 18-month project to help 30 participating institutions promote the principles of comprehensive, systematic faculty evaluation. A stimulus for the project had been SREB's 1975 survey and the 1975–76 case studies, which showed evidence that, in general, faculty evaluation was not comprehensive and systematic. In the fall of 1977, the project staff conducted two regional conferences to discuss the SREB findings and to encourage member institutions to apply to be among the 30 colleges and universities to develop new or revised faculty evaluation programs with the assistance of SREB resources. Fifty-six institutions applied for the 30 positions. Selections were based on diversity of type of institution and reflected various levels of sophistication and types of practice. The aim of the project was to help improve faculty evaluation on a regional level:

Central to the project's rationale was the belief that institutions could benefit from collectively addressing the same issues and using similar

change strategies under a regional umbrella which included periodic group experiences and access to similar regional resources while working on appropriate local approaches (O'Connell and Smartt 1979, p. 2).

To implement a regional approach, the SREB formed a task force on faculty evaluation, which reviewed staff findings, produced recommendations for developing a new or revised evaluation program, and served as an advisory committee to monitor progress throughout the project. At each campus, an institutional team of at least two faculty members and one academic administrator was formed. Institutional team members attended three workshops at six-month intervals. After each workshop, one of the workshop leaders visited each campus to consult with the institutional team there. In addition, project staff kept in contact with institutional team members.

The SREB faculty evaluation project was evaluated by a three-member team: Jon F. Wergin of Virginia Commonwealth University, Al Smith of the University of Florida, and George E. Rolle of the Southern Association of Colleges and Schools. On a rotating basis, two members of the evaluation team observed each of the three semi-annual workships and used an evaluation form to assess the effectiveness of these workshops. Also, after each workshop, one member of each institutional team was interviewed. In addition, following each visit by a consultant to one of the 30 campuses, the consultant and the institutional team members completed an evaluation form. Finally, each evaluation team member visited five institutions and reviewed the portfolio of five institutions.

According to the evaluation team, the 30 participating institutions could be organized into three categories:

1. Fifteen institutions had set a goal of developing a new comprehensive faculty evaluation system from scratch. The evaluation team found that five accomplished their goals in full, i.e., a new system had been developed, field-tested, approved, and readied for full implementation. Four had developed a new system that was currently being field-tested; four had developed parts of a new system such as a student evaluation form; and two had not progressed much beyond preliminary data collection such as faculty surveys and interviews.

2. Nine institutions had set a goal of modifying or "fine tuning" their current system, e.g., revising the student rating form or tying faculty evaluation close to faculty development. The evaluation team found that eight of these institutions had made significant progress. In the one school that had not made significant progress, poor communication and a low level of trust between the faculty and the administration seemed to be the deterrent to implementing revisions in their faculty evaluation program.

3. Six institutions had set a goal of reviewing and assessing the status quo and improving communication about faculty evaluation. These tended to be large institutions with existing formal systems of faculty evaluation. The evaluation team found that the project had little observable impact at only one of those schools (Wergin, Smith, and Rolle 1979, pp. 7–11).

The evaluation team concluded:

In summary, then, with a few exceptions, the institutional teams [had] made significant progress toward accomplishing their original goals. This progress has perhaps been most impressive in those colleges in the first group who started from "ground zero" . . . Further, there [were] major successes in both of the other two groups as well. Overall, across the 30 project institutions, observable progress toward goal accomplishment [was] visible and observable in all but four (Wergin, Smith, and Rolle 1979, pp. 8–9).

In addition to the importance of the SREB faculty evaluation project as an emblem of the growing interest in faculty evaluation during the 1970s, the project is important because the major reasons responsible for progress in the participating institutions are applicable to colleges and universities elsewhere. According to the evaluation team, seven characteristics in descending order of importance were:

1. active support and involvement of top-level administrators;
2. faculty involvement throughout the project;
3. faculty trust in administration;
4. faculty dissatisfaction with the status quo;
5. historical acceptance of faculty evaluation;
6. presence of an institutional statement covering the philosophy and uses of evaluation; and
7. degree of centralized institutional decision making (Wergin, Smith, and Rolle 1979).

For college faculty and administrators who wish to improve faculty evaluation at their campuses, these characteristics provide a template for assessing conduciveness to change. Also, in the absence of these characteristics, agents of change are provided with organizational goals to aim for when preparing for change in faculty evaluation.

Summary and Conclusions

A trend that began in the 1970s and continued into the 1980s has been to examine how faculty are evaluated. Manifestations of this trend have been meta evaluations conducted by systems of higher education, more systematic programs of faculty evaluation developed by individual institutions, and research studies carried out by evaluation specialists.

Purposes of Evaluation

A critical review of this trend indicates that, currently, faculty evaluation does not serve well the dual purposes of making personnel (promotion-retention-tenure) decisions and helping faculty improve. An examination of faculty evaluation systems indicates that often making personnel decisions is more readily served than helping faculty improve. For many faculty, evaluation of their performance is threatening and the ends of evaluation are perceived as punitive. This view is reinforced in institutions where there is a low level of trust between the administration and the faculty. Also, negative attitudes of faculty toward evaluation can be expected where faculty have not played important roles in the initiation or development of the faculty evaluation program.

In some institutions, administrators naively believe that faculty development flows naturally from faculty evaluation. It is assumed that, if faculty are provided with evaluation data, they will seek to improve their performance. However, there is little evidence that this is automatically true. Studies indicate that evaluation can lead to development under certain conditions, e.g., if educational consultation accompanies evaluation. Ironically, in some institutions where educational resources are available, faculty developers intentionally disassociate themselves from the faculty evaluation program because of its negative image.

An advantage of linking faculty development to evaluation is the efficiency of one system of data collection. Unfortunately, at the present time, many administrators and faculty see that the purpose of faculty evaluation is to make personnel decisions and pay lip service to the purpose of faculty improvement. This is unlikely to change in the near future. However, it will change in colleges and universities that use development and improvement as a criterion in their evaluation of faculty. In other words, faculty evaluation will serve the dual purposes of making personnel decisions and developing faculty where it is rewarding in the PRT process for faculty to demonstrate evidence of development and improvement.

Areas for Evaluation

The major areas to be evaluated are teaching, research, and service. The trend in faculty evaluation is to debate the weight of teaching versus research. In most colleges and universities service is considered in a distant third place, although this may not be the case in institutions that have a historic tradition of public service.

Although some research-oriented universities stress the importance of research over teaching, most colleges and universities purport to stress teaching. However, many faculty believe that the emphasis on teaching

is lip service and that research is given more weight in evaluation. In some cases, this perception is correct. In these institutions, often administrators lament the difficulty of evaluating teaching and would like to give it more weight "if only" it could be measured adequately.

Faculty will be less threatened by evaluation when (1) the weight given to teaching, research, and service is made explicit and (2) discrepancies between purported and actual weights are diminished. Evaluating what is believed to be the important areas of performance rather than easy-to-measure areas will go a long way toward increasing confidence in the evaluation process. This will require advances in the state of the art of developing criteria and standards.

Criteria and Standards

With the general trend toward more systematic and comprehensive faculty evaluation, efforts have been made to improve the *objectivity* of evaluation. The *qualitative* approach to objectivity emphasizes improving the quality of data collected from any single source, namely by providing written explicit criteria and standards of performance. The *quantitative* approach to objectivity emphasizes collecting data from multiple sources, e.g., students, peers, self, and administrators.

Because of the impetus to give more weight to teaching, criteria development has focused mostly on clarifying just what are the attributes of effective teaching for those doing the evaluating as well as for those being evaluated. Thus far, the trend is to design forms that ask students and peers to evaluate an instructor on elements considered demonstrative of effective teaching.

One unresolved issue is the use of student learning as evidence of effective teaching. To resolve this issue, much more will have to known about the relationship between teaching and learning. Although more is being learned from the experimental work of cognitive psychologists, the efforts of colleges and universities to include student learning as one of many data sources also will increase our understanding of student learning as a criterion of teaching effectiveness.

Another unresolved issue concerns peer review of classroom teaching versus review of teacher dossiers. Some faculty consider classroom visitation an infringement of rights, a negative action. Others believe that teacher dossiers are too removed from the act of teaching. Unfortunately, relatively little has been done to study the area of colleague evaluation. The state of the art is primitive regarding how to use peer review in technically adequate, useful, efficient, and ethical ways.

In the view of the authors, the development of written, explicit criteria and standards has produced a tremendous conflict in values, which we have chosen to call a "crisis in spirit." On one hand, there is the value of fair play. By making criteria and standards used to evaluate faculty *explicit*, faculty know what is expected of them. Being explicit protects faculty against race and sex discrimination as well as against other arbitrary judgments.

On the other hand, there is the value of faculty motivated by intrinsic rather than extrinsic reasons. Explicit criteria encourage faculty to do things for the sake of evaluation. A potential abuse is that faculty will meet criteria, but not with quality or the desired spirit of action. For example, suppose that one criterion of effective teaching is "the instructor provides students with an up-to-date bibliography." A teacher who is intrinsically motivated to conduct courses may naturally be familiar with new contributions to the literature and will update the bibliography as a matter of course. In this case, we can imagine a teacher who critically reads the literature and thoughtfully adds to and subtracts from the bibliography with student needs in mind. Before the development of explicit criteria, his or her maintenance of an up-to-date bibliography might even have been cited post facto as evidence of good teaching at the time of tenure review.

With the advent of explicit criteria, one could now imagine a faculty member adding new citations to the bibliography without having read the new materials. In addition, older citations could be dropped without weighing which of the old sources deserve to be maintained on the bibliography. In fact, one could even imagine an extrinsically motivated teacher preparing an *annotated* bibliography based on information provided on book jacket covers and journal article abstracts! Meeting criteria for the sake of evaluation could produce what we have chosen to call a "crisis in spirit."

The Commission on Academic Tenure in Higher Education (jointly sponsored by the American Association of University Professors and the Association of American Colleges) anticipated the same problem in its final report:

> *Evaluation too often stresses quantity rather than quality. Review committees are impressed by the number of publications rather than by their significance. Extrinsic signs such as the general reputation of journals or publishers are often substituted for a positive assessment of the work itself. Nontenured members of faculties, believing that largely quantitative tests of publication prevail, lose confidence in the evaluation process and are often prompted to undertake quick projects that will expand their bibliographies, rather than to work on more difficult or more long-term problems (1973, p. 39).*

Precedent for this crisis can be found in the movement toward instructional objectives. In the 1960s, when instructional objectives became popular, proponents argued that instructional objectives promoted fairness because students would know what to study. The argument was similar to the one used on behalf of explicit criteria for faculty evaluation: If students and teachers are informed in advance of what is the basis of their evaluation, everyone will have equal opportunity to succeed.

However, the experiences of some faculty with instructional objectives are that it is difficult to write objectives for high levels of learning, e.g.,

analyzing versus knowing, low-level learning tends to trivialize learning, e.g., "the student will be able to define . . . recognize . . . identify;" and students who study only to meet objectives do not go beyond the objectives. Similarly, it is possible that faculty evaluation will be based on criteria that are the easiest to measure and faculty will meet standards in a procedural rather than substantive fashion. The challenge in faculty evaluation is to promote fairness by using explicit criteria *and* promote quality with high standards.

Administrative Procedures

A review of faculty evaluation reveals that, in many colleges and universities, faculty involvement in initiating, developing, and implementing evaluation systems is low. The willingness of administrators to take responsibility for routine data collection (e.g., course ratings by students) reinforces the notion that evaluation is something done to faculty rather than by faculty. More faculty involvement in designing and evaluating systems of faculty evaluation can be expected in colleges and universities that use colleague ratings and self-ratings. Preparation of teaching dossiers further involves faculty in the evaluation process. Involvement of faculty is desirable because faculty judgments are needed to produce meaningful criteria and standards.

Although the economic factors that stimulated an examination of faculty evaluation may change, the authors predict no return to the "good old days" when one was promoted because the department head and dean "liked the cut of his jib." The requirement of courts that institutions of higher learning provide written explicit criteria and due process, the expectation of faculty and faculty unions for shared governance, and the support by administrators and faculty for fair personnel decisions all point to a continuation of the trend toward examining how faculty are evaluated and developing more systematic, comprehensive systems in the 1980s.

Appendix A

Suggested Form for Peer Review of Undergraduate Teaching Based on Dossier Materials

Dossier Materials	Suggested Focus in Examining Dossier Materials

1. What is the quality of materials used in teaching?

- Course outline
- Syllabus
- Reading list
- Test used
- Study guide
- Description of non-print materials
- Hand-outs
- Problem sets
- Assignments

- Are these materials current?
- Do they represent the best work in the field?
- Are they adequate and appropriate to course goals?
- Do they represent superficial or thorough coverage of course content?

Peer Reviewer's Rating: Low ___ | ___ | ___ | ___ | ___ | ___ | ___ Very High

Comments _____

_____

_____

2. What kind of intellectual tasks were set by the teacher for the students (or did the teacher succeed in getting students to set for themselves), and how did the students perform?

- Copies of graded examinations
- Examples of graded research papers
- Examples of teacher's feedback to students on written work
- Grade distribution
- Descriptions of student performances, e.g., class presentation, etc.
- Examples of completed assignments

- What was the level of intellectual performance achieved by the students?
- What kind of work was given an A? a B? a C?
- Did the students learn what the department curriculum expected for this course?
- How adequately do the tests or assignments represent the kinds of student performance specified in the course objectives?

Peer Reviewer's Rating: Low ___ | ___ | ___ | ___ | ___ | ___ | ___ Very High

Comments _____

_____

_____

3. How knowledgeable is this faculty member in subjects taught?

- Evidence in teaching materials
- Record of attendance at regional or national meetings
- Record of colloquia or lectures given
- Has the instructor kept in thoughtful contact with developments in his or her field?
- Is there evidence of acquaintance with the ideas and findings of other scholars?
  (This question addresses the scholarship necessary to good teaching. It is *not* concerned with scholarly research publication.)

Peer Reviewer's Rating Low ___|___|___|___|___|___|___ Very High

Comments _____

_____

_____

4. Has this faculty member assumed responsibilities related to the department's or university's teaching mission?

- Record of service on department curriculum committee, honors program, advising board of teaching support service, special committees (e.g., to examine grading policies, admission standards, etc.)
- Description of activities in supervising graduate students learning to teach.
- Evidence of design of new courses.
- Has he or she become a departmental or college citizen in regard to teaching responsibilities?
- Does this faculty member recognize problems that hinder good teaching and does he or she take a responsible part in trying to solve them?
- Is the involvement of the faculty member appropriate to his or her academic level? (e.g., assistant professors may sometimes become overinvolved to the detriment of their scholarly and teaching activities.)

Peer Reviewer's Rating: Low ___|___|___|___|___|___|___ Very High

Comments _____

_____

_____

5. *To what extent is this faculty member trying to achieve excellence in teaching?*

- Factual statement of what activities the faculty member has engaged in to improve his or her teaching.
- Examples of questionnaires used for formative purposes.
- Examples of changes made on the basis of feedback.
- Has he or she sought feedback about teaching quality, explored alternative teaching methods, made changes to increase student learning?
- Has he or she sought aid in trying new teaching ideas?
- Has he or she developed special teaching materials or participated in cooperative efforts aimed at upgrading teaching quality?

Peer Reviewer's Rating: Low ___ | ___ | ___ | ___ | ___ | ___ | ___ Very High

Comments _____

_____

_____

Peer Reviewer's Signature _____

Date _____

Appendix B

Guidelines for Use of Results of the Student Instructional Report
The following guidelines were developed by Educational Testing Service staff and college and university representatives to assist institutions in the appropriate use of student ratings of faculty. Although the guidelines are based primarily on the use of the Student Instructional Report (SIR), they have a value beyond their association with the use of this particular instrument.

The Student Instructional Report (SIR) typically and appropriately is used for instructional improvement; for tenure, promotion, or salary decisions; and by students for course selection. These guidelines provide information to teachers, administrators, and students who use SIR in any of these ways.* Each guideline, unless otherwise indicated, is appropriate for all three uses.

It is important that faculty members and administrators understand clearly how the results of student evaluations will be used, who will have access to any results, and how their use relates to local contractual arrangements or institutional policies.

These guideline recommendations were based on a series of studies with the Student Instructional Report and other research with similar instruments. A committee of SIR users, ETS staff, and researchers met to review and discuss the guidelines. The final list represents the experience and knowledge of this group.

1. Use multiple sources of information. For whatever purpose the results may be used, it is critical to keep in mind that student instructional ratings represent only one source of information about teaching performance. Other information about teaching, in addition to student opinion, also should be included. In particular, SIR should not be used as the sole basis for evaluating teaching effectiveness.

2. Use multiple sets of ratings. A pattern of ratings over time is the best estimate of instructor effectiveness as seen by students. Ratings from only one course or from one term may not fairly represent a teacher's performance (although, for course improvement, ratings from a single course can be useful.) For personnel decisions, it is essential to examine rating trends or patterns over time (see additional comments in number 4 regarding possible course bias):

3. Obtain a sufficient number of student raters. The reliability of the SIR items depends on having a sufficient number of students responding in order to reduce the effects of a few divergent raters.

---

*Although there may be other uses of SIR results, these guidelines address the three most frequent ones.

Currently, reports are not printed for a class with fewer than five students. Reports based on responses from fewer than 10 students are flagged with an asterisk and users are advised to interpret them with caution. When fewer than 10 students respond to any individual item, the same caution applies.

The proportion of a class that rates an instructor also is important. If over a third are absent or choose not to respond, the results may not be representative of the class.* (The reliabilities of all SIR item means are listed and discussed in *SIR Report Number 3.)*

4. Take into account course characteristics. A few course characteristics appear to affect ratings and should be taken into account by reference to appropriate comparative data or in other ways. Small classes (that is, under 15) often receive more favorable ratings than larger classes, perhaps deservedly, since they often provide a better learning environment. Courses required by the college that are not part of a student's major or minor field tend to receive somewhat lower ratings than other courses. Ratings also may differ because of the subject field of the course. For each of these characteristics, the differences may not be large, but together they can be significant.

5. Rely more on global ratings than on other items for personnel decisions. Overall ratings of the teacher or the course (items 39 and 38) tend to correlate higher with student learning scores in a course than do other items or factors in SIR. Decision makers, therefore, should focus initially on the overall evaluation items. Other items and factors in SIR, which are useful for diagnosing teacher or course strengths and weaknesses, are important for improvement purposes and for interpreting the overall ratings in personnel decisions. These items tend to reflect different teaching styles and therefore should not be summed or averaged to provide a total score. (*SIR Report Number 4* presents data on the relationship between student ratings and learning scores.)

6. Supplement diagnostic information for teaching improvement. SIR results help to diagnose teachers' strengths and weaknesses. Although studies have shown that some teachers can improve after receiving SIR results, others may not know how to change their instruction. Instructional development services and resources can help teachers who want to do something about these weaknesses. It is appropriate to use SIR results in instructional counseling and to direct teachers to resources for instructional improvement.

---

*On the SIR report itself, item means are not computed when more than 50 percent of the students either omit an item or mark it not applicable, and factor scores are not computed when there is a high (50 percent) omit or not applicable rate in one or more of the items in the factor.

7. Use comparative data. Since student ratings typically tend to be favorable, comparative data (both national and appropriate local data) provide a context within which teachers and others can interpret individual reports. In making comparisons, it is important to look at the distribution of students' responses in each class as well as at means and deciles, and not to overinterpret small differences. Differences of less than 10 percentile points on any item or factor generally are not critical, and SIR data are presented only at 10 percentile intervals. In most cases differences of at least 20 percentile points are needed to be significant relative to the national comparative data.

Users of the Student Instructional Report are reminded that the national data are comparative rather than normative, and the tendency toward high ratings may work to the disadvantage of some instructors. Institutions may wish to supplement the national data with local normative data that are developed over time. (The *SIR Comparative Guide* includes a full discussion of the composition of the national data.)

8. Employ standardized procedures for administering the forms in each class. When the results will be used in personnel decisions, it is critical to employ standardized administrative procedures. Each institution will want to develop its own method. One possibility is to have a student, another faculty member, or someone other than the teacher involved distribute, collect, and place the questionnaires in a sealed envelope. (Mailing the forms to students usually results in a poor response rate.) The teacher should not be present during the process, which probably will take less than 15 minutes of class time. The timing, preferably during the last week or two of class, also should be standard; it probably is best to give results to instructors after grades for the course have been reported.

Additional Suggestions

1. For additional diagnostic information, use the optional items and written comments. Use of optional items can make the SIR adaptable to a wider range of courses. Up to 10 additional and locally written items can be added to SIR in Section IV (items 40–49). These might be course specific, provided by the individual teacher or department, or they might be drawn from the "Suggested Supplementary Items" list included in *Instructor's Guide to Using SIR*. Information from these items and from the comments written by students in response to the last part of SIR (for example, How can the course, or the way it was taught, be improved?) provide additional helpful information to teachers. Faculty members and others who receive this information should keep in mind, however, that it may not be possible or desirable to satisfy all students' complaints or wishes.

2. Teachers should be encouraged to supplement their instructional ratings. This is especially important in personnel decisions or in student use of SIR results for course selection. Teachers should be encouraged and

given the opportunity to describe what they were trying to accomplish in the course and how their methods fit those objectives, or to discuss circumstances they feel may have affected the evaluations. What may seem like poor ratings on a particular aspect of a course may be due, for example, to the teacher's attempt at a new or different approach to the course.

3. Carry out local studies, if possible. It also may be desirable for an institution to supplement SIR research findings with local studies.

4. Do not overuse the forms. If ratings are used in every course every term, students can get bored and may respond haphazardly or not at all. Faculty members may resent the lost class time and also may pay less attention to the results. For these reasons an institution may wish to monitor the frequency of use of student evaluation. Strike a balance between the need for external evaluation and the need to experiment freely in instruction.

Bibliography

The ERIC Clearinghouse on Higher Education abstracts and indexes the current literature on higher education for the National Institute of Education's monthly bibliographic journal *Resources in Education*. Many of these publications are available through the ERIC Document Reproduction Service (EDRS). Ordering number and price for publications cited in this bibliography that are available from EDRS have been included at the end of the citation. To order a publication, write to EDRS, P.O. Box 190, Arlington, Virginia 22210. When ordering, please specify the document number. Documents are available as noted in microfiche (MF) and paper copy (PC).

Anderson, Scarvia B.; Ball, Samuel; and Murphy, Richard T. *Encyclopedia of Educational Evaluation*. San Francisco: Jossey-Bass, 1977.

Arreola, Raoul A. "A Cross-Institutional Factor Structure Replication of the Michigan State University SIRS Faculty Evaluation Model." *College Student Journal* 7 (1973):38–42.

Atkin, J. Myron. "Behavioral Objectives in Curriculum Design: A Cautionary Note." *Science Teacher* 35 (May 1968):27–30.

Baleh, Pamela M. "Faculty Evaluation in Higher Education: A Review of Court Cases and Implications for the 1980's." Photocopied, 1980. ED 187 285. MF-$1.11; PC-$5.14.

Baird, Leonard. "Common and Uncommon Models for Evaluating Teaching." In *Evaluating Learning and Teaching*, edited by C. Robert Pace. New Directions for Higher Education, No. 4. San Francisco: Jossey-Bass, 1973.

Batista, Enrique E. "The Place of Colleague Evaluation in the Appraisal of College Teaching: A Review of the Literature." *Research in Higher Education* 4 (1976):257–71.

Bejar, Isaac I. "A Survey of Selected Administrative Practices Supporting Student Evaluation of Instructional Programs." *Research in Higher Education* 3 (1975):77–86.

Berk, Ronald A. "The Construction of Rating Instruments for Faculty Evaluation." *Journal of Higher Education* 50 (September/October 1979):650–69.

Blackburn, Robert T., and Clark, Mary Jo. "An Assessment of Faculty Performance: Some Correlates Between Administrator, Colleague, Student and Self-Ratings." *Sociology of Education* 48 (Spring 1975):242–56.

Bland, Carol; Reineke, Robert A.; Welch, Wayne W.; and Shahady, Edward J. "Effectiveness of Faculty Development Workshops in Family Medicine." *Journal of Family Practice* 9 (September 1979):453–58.

Bloom, Benjamin S.; Englehart, Max D.; Furst, Edmond J.; Hill, Walker H.; and Krathwohl, David. *Taxonomy of Education Objectives: Handbook I. Cognitive Domain*. New York: David McKay Co., 1956.

Boyd, James E., and Schientinger, E. F. *Faculty Evaluation Procedures in Southern Colleges and Universities*. Atlanta: Southern Regional Education Board, 1976. ED 121 153. MF-$1.11; PC-$6.29.

Brandenburg, Dale C.; Braskamp, Larry A.; and Ory, John C. "Considerations for an Evaluation Program of Instructional Quality." *CEDR Quarterly* 12 (Winter 1979):8–12.

Brown, David Lile. "Faculty Ratings and Student Grades: A University-wide Multiple Regression Analysis." *Journal of Educational Psychology* 68 (October 1976):573–78.

Centra, John A. "Reliability of Student Instructional Report Items." SIR Report No. 3. Princeton, N.J.: Educational Testing Service, 1973.

———. *How Universities Evaluate Faculty Performance: A Study of Department Heads*. Graduate Record Examination Board Research Report No. 75-5bR. Prin-

ceton, N.J.: Educational Testing Service, 1977. ED 157 445. MF-$1.11; PC-$5.14.
———. *Determining Faculty Effectiveness*. San Francisco, Jossey-Bass, 1979.
Cohen, Arthur M., Lombardi, John; and Brawer, Florence B. *College Responses to Community Demands*. San Francisco: Jossey-Bass, 1975.
Commission on Academic Tenure. *Faculty Tenure*. San Francisco, Jossey-Bass, 1973.
Conine, Tali A. "Faculty Self-Appraisal." *Journal of Allied Health* 10 (February 1981):49–52.
Council of the American Association of University Professors. "Statement on Teaching Evaluation." *AAUP Bulletin* (Summer 1975):200–202.
Darr, Ralph F., Jr. "Evaluation of College Teaching: State of the Art, 1977." Paper presented at the Ohio Academy of Science Psychology Division, April 1977, at Columbus, Ohio. ED 162 559. MF-$1.11; PC-$5.14.
Dressel, Paul L. *Handbook of Academic Evaluation*. San Francisco, Jossey-Bass, 1976.
Dunn, Rita, and Dunn, Kenneth. *Administrator's Guide to New Programs for Faculty Management and Evaluation*. West Nyack, N.Y.: Parker Publishing Company, 1977.
Dwyer, Francis. "Selected Criteria for Evaluating Teacher Effectiveness." *Improving College and University Teaching* 21 (January 1973):51–52.
Eble, Kenneth. "What to Say about Teaching at Tenure Time." *ADE* [Association of Departments of English] *Bulletin* 57 (May 1978):30–33
Erickson, Glenn R., and Erickson, Bette L. "Improving College Teaching: An Evaluation of a Teaching Consultation Procedure." *Journal of Higher Education* 50 (September/October 1979):670–83.
Ericksen, Stanford C., and Kulik, James A. "Evaluation of Teaching." *Memo to the Faculty* (Center for Research in Learning and Teaching, University of Michigan) 53 (March 1974):1–6.
Fenker, Richard M. *The Evaluation of University Faculty and Administrators* 46 (November/December 1975):665–87.
Fincher, Cameron. "Linking Faculty Evaluation and Faculty Rewards in the University." *Issues in Higher Education*, No. 16. Atlanta: Southern Regional Education Board, 1980. ED 187 275. MF-$1.11; PC-$3.49
French-Lazovik, Grace. "Peer Review: Documentary Evidence in the Evaluation of Teaching." In *Handbook of Teacher Evaluation*, edited by J. Millman. Beverly Hills: Sage Publications, 1981
Gaston, Jerry; Lantz, Herman R.; and Snyder, Charles R. "Publication Criterion for Promotion in Ph.D. Graduate Departments." *The American Sociologist* 10 (November 1975):239–42.
Genova, William J.; Madhoff, Marjorie K.; Chin, Robert; and Thomas, George B. *Mutual Benefit Evaluation of Faculty and Administrators in Higher Education*. Cambridge, Mass.: Ballinger Publishing Company, 1976.
Ginther, John R. "A Radical Look at Behavioral Objectives." Paper presented at American Educational Research Association Annual Meeting, April 1972 in Chicago.
Glasman, Naftaly S. "Evaluation of Instructors in Higher Education." *Journal of Higher Education* 47 (May/June 1976):309–26.
Goldenstein, R. J., and Anderson, R. C. "Attitudes of Faculty toward Teaching." *Improving College and University Teaching* 25 (Spring 1977): 110–11.
Goldschmid, M. L. "The Evaluation and Improvement of Teaching in Higher Education." *Higher Education* 7 (1978):221–45.
Grasha, Anthony F. *Assessing and Developing Faculty Performance: Principles and Models*. Cincinnati: Communication and Association Associates, 1971.

Greenwood, Gordon E., and Ramalgi, Howard. "Alternatives to Student Ratings of College Teaching." *Journal of Higher Education* 51 (November/December 1980):673–84.

Grinnel, Richard M., and Kyte, Nancy S. "Measuring Faculty Competence: A Model." *Journal of Education for Social Work* 12 (Fall 1976):44–50.

Gronlund, Norman E. *Stating Objectives for Classroom Teaching*. New York: Macmillan Publishing Co., 1978.

Guba, Egon G., and Lincoln, Yvonna S. *Effective Evaluation*. San Francisco: Jossey-Bass, 1981.

Hawley, Robert C. "Faculty Evaluation—Some Common Pitfalls." *Independent School* (May 1977):39–40.

Hildebrand, Milton; Wilson, Robert C.; and Dienst, Evelyn R. *Evaluating University Teaching*. Berkeley: Center for Research and Development in Higher Education, University of California at Berkeley, 1971. ED 057 748. MF-$1.11; PC-$6.29.

Hind, Robert R.; Dornbusch, Sanford M.; and Scott, W. Richard. "A Theory of Evaluation Applied to a University Faculty." *Sociology of Education* 47 (Winter 1974):114–28.

House, Ernest R. *Evaluating With Validity*. Beverly Hills: Sage Publications, 1980.

Irby, David, and Rakestraw Phillip. "Evaluating Clinical Teaching in Medicine." *Journal of Medical Education* 56 (March 1981):181–86.

Jauch, Lawrence R. "Relationships of Research and Teachings: Implications for Faculty Evaluation." *Research in Higher Education* 5 (August 1976):1–13.

Johnson, G. E., and Stafford, Frank P. "Lifetime Earnings in a Professional Labor Market: Academic Economists." *Journal of Political Economy* 82 (1974):549–69.

Kane, Robert L., and Schorow, Mitchell. "Rating the Lecturer: A Study of How Different Disciplines Used Comparative Criteria." *Journal of Community Health* 2 (Summer 1977):278–86.

Kelly, Janet, and Woiwode, Daniel. "Faculty Evaluation by Residents in a Family Medicine Residency Program." *The Journal of Family Practice* 4 (April 1977):693–95.

Ketefian, Shake. "A Paradigm for Faculty Evaluation." *Nursing Outlook* 25 (November 1977):718–20

Knapper, Christopher K. "Evaluation and Teaching: Beyond Lip Service." Paper presented at International Conference on Improving University Teaching, July 1978, in Aachen, Canada. ED 165 662. MF-$1.11; PC-$3.49.

Knox, Alan B. *Adult Development and Learning*. San Francisco: Jossey-Bass, 1978.

Ladd, Everett Carll, Jr., and Lipset, Seymour Martin. *Professors, Unions, and American Higher Education*. Berkeley: Carnegie Foundation for the Advancement of Teaching, 1973.

Levinson, Judith L., and Menges, Robert J. *Improving College Teaching: A Critical Review of Research*. Evanston, Ill.: Center for the Teaching Professions, Northwestern University, December 1979.

Lewis, Lionel S. *Scaling the Ivory Tower*. Baltimore: Johns Hopkins University Press, 1975.

Mager, Robert. *Preparing Instructional Objectives*. Belmont, Calif.: Fearon Publishers, 1962.

Mark, Sandra Fay. "Faculty Evaluation Systems: A Research Study of Selected Community Colleges in New York State." Photocopied. New York: Faculty Council of Community Colleges, State University of New York, 1977. ED 158 809. MF-$1.11; PC-$12.48.

Marques, Todd E.; Lane, David M.; and Dorfman, Peter W. "Toward the Development of a System for Instructional Evaluation: Is There Consensus Regarding

What Constitutes Effective Teaching?" *Journal of Educational Psychology* 71 (June 1979):840–49.

Martin, Warren Bryan. "Introduction to Section 2. Through Teaching, the Students Learn." In *Perspectives on Teaching and Learning*, edited by Warren Bryan Martin. New Directions for Teaching and Learning, No. 7. San Francisco: Jossey-Bass, 1981.

McCarter, W. Ronald. "Making the Most of Subjectivity in Faculty Evaluation." *American Vocational Journal* 49 (January 1974):32–33.

McKeachie, Wilbert J. "Student Ratings of Faculty: A Reprise." *Academe* (October 1979):384–97.

Meany, John O., and Ruetz, Frank J. "A Probe into Faculty Evaluation." *Educational Record* 53 (Fall 1972):300–307.

Meeth, L. Richard "The Stateless Art of Teaching Evaluation." *Change Report on Teaching* 2 (July 1976):3–5.

Melton, Reginald F. "Resolution of Conflicting Claims Concerning the Effect of Behavioral Objectives on Student Learning." *Review of Educational Research* 48 (Spring 1978):291–302.

Michalak, Stanley J., Jr., and Friedrich, Robert J. "Research Productivity and Teaching Effectiveness at a Small Liberal Arts College." *Journal of Higher Education* 52 (November/December 1981):578–97.

Miller, Richard I. *Evaluating Faculty Performance*. San Francisco: Jossey-Bass, 1972.

_____. *Developing Programs for Faculty Evaluation*. San Francisco: Jossey-Bass, 1974.

_____. *The Assessment of College Performance*. San Francisco: Jossey-Bass, 1980.

Milton, Ohmer, et al. *On College Teaching*. San Francisco: Jossey-Bass, 1980.

Moomaw, W. Edmund; O'Connell, William R., Jr.; Schietinger, E. F.; and Smartt, Steven H. *Faculty Evaluation for Improved Learning*. Atlanta: Southern Regional Education Board, 1977. ED 149 683. MF-$1.11; PC-$6.29.

Murray, Harry G. "The Validity of Student Ratings of Teaching Ability." Paper presented at Canadian Psychological Association Meeting, 1972, in Montreal.

_____. "Student Evaluation of University Teaching: Uses and Abuses." Vancouver: Colloquium presented at the University of British Columbia, March 1979. ED 189 920. MF-$1.11; PC-$5.14.

Nadeau, Gilles G. "A Plan for Evaluation for Promotion, Tenure, and Assignment." Paper presented at American Educational Research Association Annual Meeting, April 1977, in New York.

Naftulin, Donald H.; Ware, John E.; and Donnelly, Frank A. "The Doctor Fox Lecture: A Paradigm of Educational Seduction." *Journal of Medical Education* 48 (July 1973):630–35.

North, Joan, and Scholl, Stephen. *Revising a Faculty Evaluation System: A Workbook for Decision Makers*. Washington, D.C.: Small College Consortium, 1978. ED 165 634. MF-$1.11; PC-$6.29.

O'Connell, William R., Jr., and Smartt, Steven H. *Improving Faculty Evaluation: A Trial in Strategy*. Atlanta: Southern Regional Education Board, 1979. ED 180 395. MF-$1.11; PC-$5.14.

O'Hanlon, James, and Mortensen, Lynn. "Making Teacher Evaluation Work." *Journal of Higher Education* 51 (November/December 1980):664–72.

Parramore, Barbara H. "Evaluation of University Faculty." *CEDR Quarterly* 12 (Winter 1979):3–7.

Philippi, Harlan A. "A Dean's Response to a Review of Procedures and Policies Governing Appointment, Promotion, and Tenure in New England Institutions of Higher Education." Photocopied, 1979. ED 173 132. MF-$1.11; PC-$3.49.

Pohlman, John T. "A Description of Effective College Teaching in Five Disciplines as Measured by Student Ratings." *Research in Higher Education* 4 (1976):335–46.

Popham, W. James. "Instructional Objectives: 1960–1970." Paper presented at Eighth Annual Convention of the National Society for Programmed Instruction, May 7, 1970, in Anaheim, Calif.

———. *Educational Evaluation*. Englewood Cliffs, N.J.: Prentice Hall, 1975.

Prodgers, Stephen B. "Toward Systematic Faculty Evaluation." *Regional Spotlight* (Southern Regional Education Board) 13 (January 1980). ED 181 833. MF-$1.11; PC-$3.49.

Ramagli, Howard J., Jr., and Greenwood, Gordon E. "The Dr. Fox Effect: A Paired Lecture Comparison of Lecturer Expressiveness and Lecture Content." Paper presented at American Educational Research Association Annual Meeting, April 1980, in Boston. ED 187 179. MF-$1.11; PC-$5.14.

Rippey, Robert M. *The Evaluation of Teaching in Medical Schools*. New York: Springer Publishing Company, 1981.

Rose, Clare. "Faculty Evaluation in an Accountable World: How Do You Do It?" Paper presented to National Conference of American Association of Higher Education, March 1976, in Chicago. ED 144 442. MF-$1.11. PC-$3.49.

Rutman, Leonard. *Planning Useful Evaluation*. Beverly Hills: Sage Publications, 1980.

Scott, Craig; Thorne, Gaylord L.; and Beaird, James H. "Factors Influencing Professorial Assessment." Paper presented at American Educational Research Association Annual Meeting, April 1977, in New York. ED 141 398. MF-$1.11; PC-$5.14.

Schulman, Benson R., and Trudell, James W. "California's Guidelines for Teacher Evaluation." *Community and Junior College Journal* 43 (February 1972):32–34.

Scriven, Michael. *The Methodology of Evaluation*. AERA Monograph Series in Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Seldin, Peter. *How Colleges Evaluate Professors*. New York: Blythe-Pennington, Ltd., 1975.

———. *Successful Faculty Evaluation Programs: A Practical Guide to Improve Faculty Performance and Promotion/Tenure Decisions*. Crugers, N.Y.: Conventry Press, 1980.

Sheehan, Daniel S. "On the Invalidity of Student Ratings for Administrative Personnel Decisions." *Journal of Higher Education* 46 (November/December 1975):687–700.

Siegel, Laurence. "Individual Task Profiles for Faculty Evaluation." *Engineering Education* 70 (December 1979):245–49.

Smith, Albert B. *Faculty Development and Evaluation in Higher Education*. AAHE-ERIC/Higher Education Research Report No. 8, 1976. Washington, D.C.: American Association for Higher Education, 1976. ED 132 891. MF-$1.11; PC-$8.81.

Smith, Bardwell L. et al. *The Tenure Debate*. San Francisco: Jossey-Bass, 1973.

Spencer, James H.; Crow, John F.; and Glass, Richard. "Development of Department Promotion Guidelines." *Journal of Medical Education* 54 (June 1979):477–83.

Stufflebeam, Daniel L. "Meta Evaluation: An Overview." *Evaluation and the Health Professions* 1 (Spring 1978):17–43.

Tuckman, Howard P., and Hagemann, Robert P. "An Analysis of the Reward Structure in Two Disciplines." *Journal of Higher Education* 47 (July/August 1976):447–64.

Walker, Noojin. "Faculty Judgments of Teacher Effectiveness." *Community and Junio College Research Quarterly* 3 (April-June 1979):231–37.

Ware, John E., Jr., and Williams, Reed G. "The Dr. Fox Effect. A Study of Lecturer Effectiveness and Ratings of Instruction." *Journal of Medical Education* 50 (February 1975):149–55.

Wergin, Jon F.; Smith, Al; and Rolle, George E. "An Empirical Analysis of Faculty Evaluation Strategies. Implications for Organizational Change." Paper presented at American Educational Research Association Annual Meeting, April 1980 in Boston.

Whitman, Neal. "Evaluations Can Isolate Teaching from Learning." *Instructional Innovator* 26 (March 1981):35–36.

_____. "A Guide to Clinical Performance Testing." Idea Paper No. 7. Manhattan, Kan.: Center for Faculty Evaluation and Development, Kansas State University, 1982.

Whitman, Neal, and Schwenk, Thomas. "Faculty Evaluation as a Means to Improvement: A Practical Approach to Faculty Development." *Journal of Family Practice*, in press.

Wolff, Florence I. "A Survey of Evaluative Criteria for Faculty Promotion in College and University Speech Departments." *Speech Teacher* 20 (April 1971):281–83.

Wotruba, Thomas R., and Wright, Penny L. "How to Develop a Teacher-Rating Instrument." *Journal of Higher Education* 46 (November/December 1975):653–63.

AAHE-ERIC Research Reports

Ten monographs in the AAHE-ERIC Research Report series are published each year, available individually or by subscription. Subscription to 10 issues (beginning with date of subscription) is $35 for members of AAHE, $50 for nonmembers; add $5 for subscriptions outside the U.S.

Prices for single copies are shown below. *Add 15% postage and handling charge for all orders under $15.* Orders under $15 must be prepaid. Bulk discounts are available on orders of 25 or more of a single title. Order from Publications Department, American Association for Higher Education, One Dupont Circle, Suite 600, Washington, D.C. 20036; 202/293-6440. Write or phone for a complete list of Research Reports and other AAHE publications.

1982 Research Reports—*AAHE members, $5 each; nonmembers $6.50 each; plus 15% postage/handling.*

1. Rating College Teaching: Criterion Validity Studies of Student Evaluation-of-Instruction Instruments
   Sidney E. Benton
2. Faculty Evaluation: The Use of Explicit Criteria for Promotion, Retention, and Tenure
   Neal Whitman and Elaine Weiss

1981 Research Reports—*AAHE members, $4 each; nonmembers, $5.50 each; plus 15% postage/handling.*

1. Minority Access to Higher Education
   Jean L. Preer
2. Institutional Advancement Strategies in Hard Times
   Michael D. Richards and Gerald R. Sherratt
3. Functional Literacy in the College Setting
   Richard C. Richardson Jr., Kathryn J. Martens, and Elizabeth C. Fisk
4. Indices of Quality in the Undergraduate Experience
   George D. Kuh
5. Marketing in Higher Education
   Stanley M. Grabowski
6. Computer Literacy in Higher Education
   Francis E. Masat
7. Financial Analysis for Academic Units
   Donald L. Walters
8. Assessing the Impact of Faculty Collective Bargaining
   J. Victor Baldridge, Frank R. Kemerer and Associates
9. Strategic Planning, Management, and Decision Making
   Robert G. Cope
10. Organizational Communication and Higher Education
    Robert D. Gratz and Philip J. Salem

1980 Research Reports—*AAHE members, $3 each; nonmembers, $4 each; plus 15% postage/handling.*

1. Federal Influence on Higher Education Curricula
   William V. Mayville
2. Program Evaluation
   Charles E. Feasley
3. Liberal Education in Transition
   Clifton F. Conrad and Jean C. Wyer
4. Adult Development: Implications for Higher Education
   Rita Preszler Weathersby and Jill Mattuck Tarule
5. A Question of Quality: The Higher Education Ratings Game
   Judith K. Lawrence and Kenneth C. Green
6. Accreditation: History, Process, and Problems
   Fred F. Harcleroad
7. Politics of Higher Education
   Edward R. Hines and Leif S. Hartmark
8. Student Retention Strategies
   Oscar T. Lenning, Ken Sauer, and Philip E. Beal
9. The Financing of Public Higher Education: Low Tuition, Student Aid, and the Federal Government
   Jacob Stampen
10. University Reform: An International Review
    Philip G. Altbach