The Evaluation of Teaching Effectiveness

Frederick T. Bail

Peter Dunn-Rankin

The purpose of this article is to stimulate discussion and use of various measures of teaching effectiveness in college instruction. Evaluation of teaching effectiveness will be considered within four contexts: (1) perceived importance, (2) current status of evaluation, (3) dimensions of evaluation, and (4) purposes of evaluation.

Importance of Teacher Effectiveness

A questionnaire on the importance of teacher performance was distributed to faculty members in colleges and universities throughout the United States in March, 1969, by the American Council on Education. Of the 60,000 faculty members who responded, 67% of those at universities and 86% of those at 4-year institutions agreed with the statement, "Teaching effectiveness, not publications, should be the primary criterion for promotion of faculty" (Pitts, 1970, p. 3). Opinion at the University of Hawaii is much the same. In May, 1971, a sample of faculty members from the University was asked to react to goals in higher education and to attach priorities to them. Of those polled, 56% agreed there should be greater emphasis on teaching performance than on publication, 22% disagreed, and 22% were neutral. Over half of the respondents felt that greater emphasis on teaching performance was a "high priority" issue (Marsh, 1971).

Astin and Lee (1967) surveyed over 1100 academic deans to ascertain their opinion on the relative importance of various criteria for promotion and tenure decisions. "Classroom teaching" was labeled a major factor by the greatest number (96%). The next most important factor ("personal attributes") was far behind (57%). Less than half of the deans responding felt that either "research" or "publication" was a major factor.

Current Status of Evaluation of Teacher Effectiveness

Given the stated value of teaching effectiveness, it is interesting to analyze the frequency with which different criteria are actually used to measure teaching effectiveness. In Table 1 it is seen that "informal student opinion" (41%) is used much more frequently as a criterion than "systematic student ratings" (12%). "Chairman evaluation" (85%) and "dean evaluation" (82%) are accorded places of extreme importance despite untested assumptions about their validity.

As Astin and Lee have concluded, ". . . the professor's scholarly research and publication—not information based on classroom visits, systematic student ratings, student performance on examinations, and similar sources—are currently the primary considerations in evaluating his teaching ability" (p. 298). It can be seen, therefore, that specific strategies for measuring teaching ef-

Source of Information	Use in all or most Departments (%)
Chairman evaluation	85.1
Dean evaluation	82.3
Colleagues' opinions	48.9
Scholarly research and publications	43.8
Informal student opinions	41.2
Grade distributions	28.0
Course syllabus and examinations	26.4
Committee evaluation	25.1
Student examination performance	19.6
Self-evaluation or report	16.3
Classroom visits	14.0
Systematic student ratings	12.4
Enrollment in elective courses	11.0
Long-term follow-up of students	10.2
Alumni opinions	9.9

*Adapted from Astin and Lee (1967), p. 298.

Table 1
Frequency of Use of Various Sources of Information in the Evaluation of Teaching Effectiveness*

fectiveness need to be examined. Before this examination, however, the definition of "teaching effectiveness" must be clarified.

There is a vast amount of semantic confusion surrounding the term "teacher effectiveness," in large part due to lack of objective measures for defining the criterion. Whether student ratings, chairman evaluations, or some other method is used to appraise teaching effectiveness, the individual department or institution must specify more precisely the teacher and student behaviors and attitudes that they identify as being the outcome of effective teaching. This is not meant to imply that all criteria must be tightly-worded behavioral objectives, though the accuracy of measurement will depend in part on the precision with which the desired outcomes are defined.

Rosenshine (1968) reduced the concept of "teaching effectiveness" to the more precise "ability to explain." In an experiment utilizing both student ratings and performance tests of this ability, he found that student ratings of the clarity of the lesson and teacher skill in presenting the lesson were significantly related to adjusted pupil achievement scores. Another study found that the "learning of facts" was related to ratings of instructor clarity, expressiveness, and lecturing habits, while "gains in comprehension" were related to instructor personality dimensions of energy, flamboyance, and permissiveness (Solomon, Rosenberg, and Bezdek, 1964).

One problem in attempting to specify criteria of effective teaching relates to Thorndike's distinctions among immediate, intermediate, and ultimate criteria. Gustad (1967) notes that the few criteria we do use are *immediate* criteria. Suppose one teaches principles of behavior modification to prospective teachers. Although the *ultimate* criterion of teaching success might be effective application of such principles in the classroom and the *intermediate* criterion an understanding of the rationale underlying behavior modification, the *immediate* criterion is typically a score on a test of knowledge of those principles. If current measures of faculty effectiveness do not correlate very well with *immediate* criteria, it is little wonder that faculty are especially bothered by questions concerning the relationship between measures of effectiveness and the *ultimate* criterion.

Dimensions of Evaluation

Although this analysis seems discouraging, ignoring assessment difficulties will only serve to intensify the problem. An attempt will be made in this section to outline three *dimensions* of evaluation: (1) methods of evaluation, (2) types of rating scales, and (3) factors underlying evaluation of teaching effectiveness.

Evaluation Methods

To evaluate teacher effectiveness, it is first necessary to review the range of methods that are currently available. This particular analysis of possible evaluation methods is drawn in part from discussions by Eble (1970) and Dressel (1959). The major evaluation *methods* discussed are as follows: examining course materials, measuring the attainment of objectives, and rating teacher effectiveness.

Course Materials in the Evaluation of Teaching

Examining course materials is a valuable but usually overlooked evaluation method. There is little beyond the brief (and often outdated or inaccurate) "blurb" in college catalogues to describe a course and its objectives to the prospective student or, for that matter, the fellow faculty member. Sometimes students publish course descriptions, but these rarely describe course objectives.

A syllabus is an important part of the "course materials" because, ideally, it provides a clear statement of the course objectives. Disagreements concerning course objectives among students or faculty are often mislabeled as disagreements about course effectiveness because the objectives are not specified. What a faculty member knows about the objectives of other courses in the department is typically inferred from informal information derived from students. He thus may make assumptions that are not valid. Whereas a precise statement of objectives costs a little extra time initially, it organizes the course more clearly for instructor and student alike. It reveals inter-course omissions and repetitions within a department. It clarifies expectations of students and the self-expectations of the instructor. Finally, by stating the objectives as clearly as possible, it allows more accurate evaluation of the attainment of those objectives.

Inspection of the extent to which texts, adjunct teaching material, tests, and other materials for student assessment are integrated may provide a general understanding of how well the instructor has planned the course. In addition, inspection of course materials provides a framework for interpreting results from other evaluation sources.

Direct Assessment of Objectives

One obvious way to measure teacher effectiveness is through the use of student achievement *gain* scores. Popham (1968, 1971a, 1971b) has been a strong advocate

of using this approach in a systematic way and has outlined a system whereby the teacher is given instructional objectives for a course or module, a sample of the measurement procedure to be used, and time to plan the instruction. Following instruction, teacher skill is measured by increased learner performance.

Another direct assessment involves the use of delayed retention tests. Long-term retention of concepts and techniques are implicit goals of most courses, yet such retention is rarely measured.

Some evaluators separate the student's own estimates of performance from his actual performance. Students are asked to judge whether or not they feel that they possess the various concepts and techniques covered in the lessons. The assumption underlying use of this approach is that correct applications of techniques are meaningless unless the student also feels competent enough to actually use them.

Indirect Assessment of Objectives

It is also possible to measure attainment of objectives indirectly. If the course falls into some recognized departmental sequence, for example, students in more advanced courses could be measured on the skills supposedly obtained in the prerequisite courses. A course on the French Revolution might initially assess, directly or indirectly, students' knowledge of the pertinent portions of a prior course in modern European history.

Another indirect measure of a teacher's effectiveness might be the number or percentage of students choosing a major in the field, though interest in the area prior to taking the course would somehow have to be ascertained and held constant. Perhaps instead of counting majors, the number of courses later elected in the field could be used as a measure of instructor effectiveness. Some evidence exists that students of highly rated instructors elect more courses in the same field than students of instructors who are not so highly rated (McKeachie & Solomon, 1958).

Another measure would involve counting the number of times optional or supplementary material has been checked out of the library during the course. Even a measure as crude as average percentage of students attending lectures would give some indirect indication of teaching effectiveness, assuming that, if attendance is optional, students will only attend if the instructor is interesting or effective in facilitating student learning.

Assessment by Ratings

The last type of evaluation method discussed, rating of instruction, is also the most common. Rating can be done by direct observation or by indirect systems. Direct observation of teaching can be undertaken by students, colleagues, chairmen, or by the instructors themselves. Observation may be quantified by recording student or teacher behaviors with a checklist, by analyzing teacher-student interaction, or by observing the percentage of time the instructor is involved in various activities. Self-observations can be made through the use of audio and/or video recording equipment.

Research has not yet been able to ascertain the effects of the observer in the classroom (Masling & Stern, 1969). Furthermore, one review concluded that due to the errors inherent in the observer process and the ineffectiveness of supervisor or colleague ratings, observational methods should be discarded in favor of student ratings (Walker & Fischer, 1965).

Student ratings are most often obtained using some systematic scale but may also be obtained by interviewing a random sample of students from the class or, more simply (though probably less accurately), by "word of mouth." One virtually untapped measure is to interview samples of graduating students or ask them to list the five most effective instructors (and possibly the five least effective) from whom they had taken courses.

Alumni ratings are sometimes obtained but the turnover of faculty and teaching assignments are serious disadvantages of this method. Evaluation ratings of teaching effectiveness are often made by chairmen or deans, though often through second-hand information. Colleague ratings can most effectively be used in team-teaching situations when colleagues are able to observe each other over long periods. Self-appraisals can also be made. One particularly promising example of this method is the use of feedback from watching "expert teachers" in making a self-appraisal. Currently in successful use at the pre-collegiate level (Washington, 1970), this method involves groups of teachers in discussions and observations of excellence in teaching, followed by a continuing program of self-appraisal.

Types of Rating Scales

Most rating scales, as discussed in an excellent review by Remmers (1963), are a combination of numeric rating scales (which assign numbers to various reactions to statements) and graphic rating scales (which provide ordered categories along a continuum). For example, the University of Hawaii's *Faculty-Course Evaluation Scale* (J-5 Form) contains statements such as, "(Instructor is) able to explain difficult concepts; pulls abstractions down to earth." The student circles a letter correspond-

ing to one of the following reactions: strongly agree, agree, uncertain, disagree, strongly disagree.

Cumulated-point rating scales, which differ only slightly from the numeric and graphic forms, contain statements such as, "Do you like the teacher?". The responses, "no," "uncertain," and "yes" are weighted, respectively, 0, 1, and 2. Thus, answers to items are weighted according to their desirability and scores are summed across all items.

The use of checklists of teacher or student behaviors was mentioned under observational methods. Sometimes a set of concepts is listed and the ordered alternatives under each concept are weighted logically according to the favorability of the response. An example of such a multiple-choice rating form might be:

Explanation of concepts:

(−2) a. Totally confusing; initial understandings lost.

(−1) b. Fuzzy; doesn't interrelate concepts.

(+1) c. Good; meaning of terms clear in majority of cases.

(+2) d. Crystal-clear; even most difficult terms makes sense in various contexts.

(Weights are given in parentheses.)

Forced-choice rating scales are ipsative in nature; that is, they obtain judgments about the relative strength of certain traits or abilities in relationship to the rated person's other traits and abilities. An item from such a scale might ask the rater to select one of the following four alternatives as being *most* like the instructor:

a. Always on time.
b. Explanations are clear.
c. Good rapport with students.
d. Dresses well.

This type of rating scale would probably be used much more frequently if excessive pilot testing were not necessary to establish *a priori* the equality of various alternatives. Remmers (1963) presents evidence that norms from such rating scales are comparable to those obtained from graphic rating scales and that items which contain only desirable alternatives yield the most valid results.

There are several other rating systems which are slightly different from the scales discussed above. The semantic differential elicits information on the rated concept along three dimensions: evaluation, potency, and activity. Although several measures are usually taken of each factor for the rated concept, the hypothetical example below uses only one assessment of each:

For each of the following scales please put an "x" in the space you feel most accurately represents the teaching style of this instructor:

fair___:___:___:___:___:___:___ *unfair*
light___:___:___:___:___:___:___ *heavy*
slow___:___:___:___:___:___:___ *fast*

The semantic differential is one of the most widely used psychological instruments. Its major advantage is in effectively describing differences in courses and instructors that are not elucidated by other methods.

The Q-technique, or Q-sort, requires the rater to sort statements regarding the person or thing to be rated into an ordered set of piles (e.g., seven) according to the perceived applicability of the statement to the person or thing being rated. The number of statements to be placed in each pile is the same for each rater and is representative of the frequencies in a normal distribution. The Q-technique has not often been used to rate instructors, probably because the information gained is not worth the additional problems.

Another ipsative measure is the self-anchoring rating scale. Essentially, this technique requires that the rater make judgments about the person or thing being rated relative to that rater's ideal or standard.

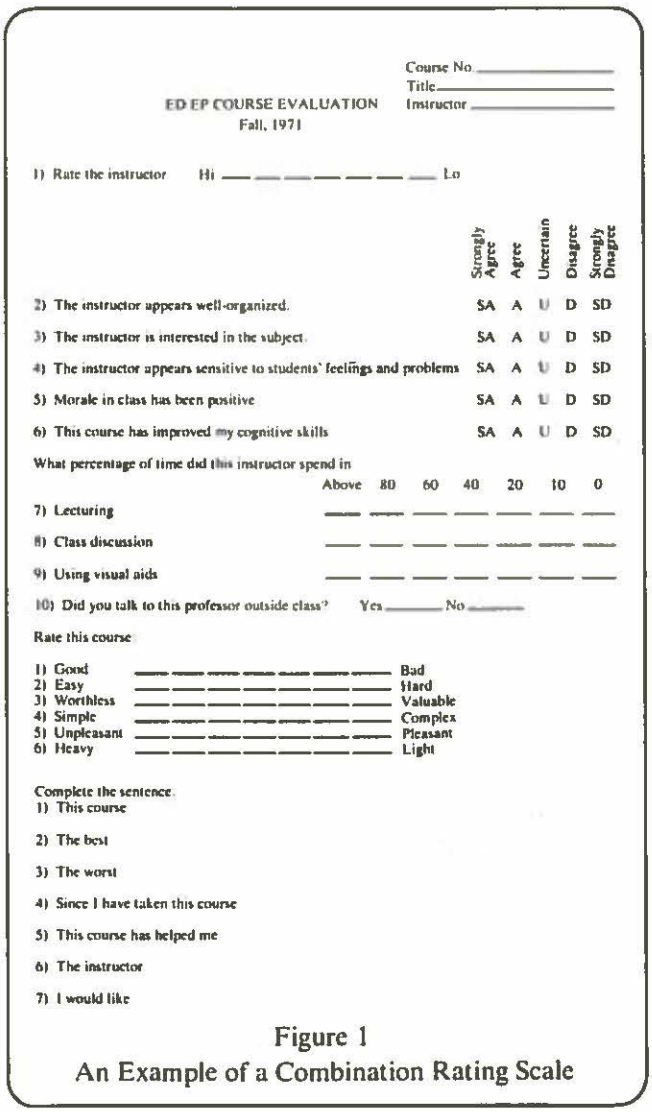
At least one projective technique, incomplete sentences, should also be considered with other types of scales due to its ability to elicit unique information or dimensions. For example, student raters might be asked to complete the following sentences:

This class_____.
The most striking thing about
the instructor_____.

An example of combining several of these methods in one instrument is presented in Figure 1. Such combination is valuable because different rating techniques provide different kinds of information.

Research on Student Ratings

The vast body of research dealing with the various aspects of student ratings has been discussed in a number of previous reviews (e.g., Eble, 1970; McKeachie, et al., 1971; Melnick, 1969; Remmers, 1963; Stecklein, 1960; Walker & Fischer, 1965). Though research concerning student ratings is far from unequivocal, the authors feel fairly confident in drawing three conclusions. First, it is possible for student rating scales to yield reliable measures. Second, it is possible for student rating scales to be free of subjective student biases. Last, validity of student rating scales is the most important



Course No. _____
Title _____
Instructor _____

ED EP COURSE EVALUATION
Fall, 1971

1) Rate the instructor Hi — — — — — — — Lo

	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
2) The instructor appears well-organized.	SA	A	U	D	SD
3) The instructor is interested in the subject.	SA	A	U	D	SD
4) The instructor appears sensitive to students' feelings and problems	SA	A	U	D	SD
5) Morale in class has been positive	SA	A	U	D	SD
6) This course has improved my cognitive skills	SA	A	U	D	SD

What percentage of time did this instructor spend in

	Above 80	60	40	20	10	0
7) Lecturing						
8) Class discussion						
9) Using visual aids						

10) Did you talk to this professor outside class? Yes _____ No _____

Rate this course

1) Good	— — — — — — —	Bad
2) Easy	— — — — — — —	Hard
3) Worthless	— — — — — — —	Valuable
4) Simple	— — — — — — —	Complex
5) Unpleasant	— — — — — — —	Pleasant
6) Heavy	— — — — — — —	Light

Complete the sentence.
1) This course
2) The best
3) The worst
4) Since I have taken this course
5) This course has helped me
6) The instructor
7) I would like

Figure 1
An Example of a Combination Rating Scale

consideration, but it is not always clear to what extent the typical rating scale is valid. An extensive review of student ratings of college teaching by Costin, Greenough, and Menges (1971), however, concluded that student ratings can provide valid information on the quality of courses and instruction.

The crucial problem is that validity is not an absolute concept. A specification of the validity of an index of teacher effectiveness entails a statement of validity for some purpose, thereby raising questions about the definition of the criterion. Unfortunately, relatively few attempts have been made to objectify criteria for "teacher effectiveness." The problem of measuring validity, however, is applicable to *all* methods of teacher evaluation.

Factors Underlying Judgments of Teaching Effectiveness

The number of factors derived from student rating questionnaires may vary somewhat from study to study; nevertheless, general factors with common features do seem to underlie the different questionnaires (Slobin & Nichols, 1969). At least three such factors usually appear in analyses of student ratings. One relates to the organization and clarity within the course; another relates to the personality of the instructor (e.g., enthusiasm, openness); and the third is an interaction factor that stems from the interrelationship of the instructor and students in the actual classroom situation. The fact that student ratings are not based on a single dimension is supported by research which indicates that student ratings are fairly independent of students' general attitude towards instruction (Feldhusen & Starks, 1970).

Purposes of Evaluation

Gustad (1966) found that systematic student ratings were about the fifth most used source of evaluation in 1961, but by 1966 they had dropped to about tenth place in importance. Whereas faculty are willing to accept the unsupported validity of "overall" evaluations by chairmen and deans, they are unwilling to acknowledge the empirical validity established for systematic student ratings. The absence of acceptable definitions of good teaching may be the cause of this paradoxical situation. Since most faculty members would have difficulty defending their teaching methods, it is little wonder that they fear measures that attempt to define and measure teaching effectiveness (McKeachie, 1969). It is far easier for them to assume that deans and chairmen all have a common understanding of "teaching effectiveness" and have standard, valid ways of assessing it. Though hardly acceptable, the situation is understandable. In the same way that "quantity of publications" replaces "quality of new knowledge" as a criterion of academic success, so do research and service outweigh teaching as criteria because they are easier to measure "objectively."

If our faculty review procedures were to stress development rather than judgment, as Eble (1970) suggests, faculty fear of systematic student ratings would be greatly allayed. Unfortunately, in order to interpret measures of teaching effectiveness to promote teacher

development, a person is needed who understands both the measuring system and the discipline involved. But this person is usually a senior member of the department and one who also participates in promotion and tenure decisions (Langden, 1966).

In order for instructors to develop and for measures of teaching effectiveness to improve, evaluation must take place in an atmosphere of cooperation rather than one of latent threat to the individual. In this sense, some student-published course guides are less than helpful. Criticisms should be constructive and most public and literal "grading" of faculty discouraged. A good example of an evaluation system that stresses course improvement rather than competitive judgment of faculty has been developed at the University of Minnesota (Parent, Vaughn, & Whatton, 1971). A class profile of student characteristics, experiences, and expectations is organized at the beginning of the course. A group of students from the class meet periodically with the instructor to enhance communication. More formal evaluation is undertaken midway through the term. Options are also provided for evaluation after completion of the course.

Summary

This paper was designed to give perspective to the evaluation of teacher effectiveness. Though some recent evaluation procedures have been discussed, the major portion of the paper has been devoted to promoting the use of currently available evaluation methods. It has been suggested that universities not only fail to obtain valuable evaluative information on teaching effectiveness but also neglect a prime purpose of evaluations, namely, the development of better courses.

There is evidence that the University of Hawaii is beginning to stress better measures of teaching effectiveness. A 1970 report from the Faculty Personnel Committee to the Faculty Senate called for more specific evaluations of teaching ability, with clear statements of the criteria used to judge an individual. Student rating summaries were one of the suggested objective criteria. It remains to be seen how well personnel committees at the various levels maintain these more specific criteria.

There is also evidence that the University is beginning to take a more realistic attitude towards the use of evaluation. A 1970 memo within the Department of Linguistics listed five possible uses of evaluation: professional advancement, optimal utilization of personnel, teaching improvement, improvement of student morale while stimulating student involvement in educational objectives, and advising students on choice of instructor.

The burden of constructing and validating an instrument may be eased by using one of the rating systems available through the Academic Evaluation Office. Departments may wish to develop more specialized instruments as the advantages of systematic evaluation become more obvious. Use of more than one method is probably preferable, as is the measurement of both absolute and relative strengths of an instructor. The most important concern, however, is to create a climate conducive to more precise measurement of effective teaching. Eble (1970) noted that the conditions surrounding the evaluation systems that seemed to work well were more important than the particular system used.

References

Astin, A. W., & Lee, C. B. T. "Current Practices in the Evaluation and Training of Teachers," In C. B. T. Lee (Ed.), *Improving College Teaching*, Washington D. C., American Council on Education, 1967, 296-311.

Costin, F., Greenough, W. T., & Menges, R. J. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research*, 1971, 5, 511-535.

Dressel, P. E. "The Current Status of Research on College and University Teaching," In W. J. McKeachie (Ed.), *The Appraisal of Teaching in Large Universities*, Ann Arbor, University of Michigan, 1959.

Eble, K. E. *The Recognition and Evaluation of Teaching*, Washington D. C., American Association of University Professors, 1970.

Feldhusen, J. F., & Starks, D. D. "Bias in College Students' Ratings of Instructors," *College Student Survey*, 1970, 4, 6-9.

Gustad, J. W. "Evaluation of Teaching Performance: Issues and Possibilities," In C. B. T. Lee (Ed.), *Improving College Teaching*, Washington D. C., American Council on Education, 1967, 265-281.

Langen, T. D. F. "Student Assessment of Teaching Effectiveness," *Improving College and University Teaching*, 1966, 14, 22-25.

Marsh J. B. Personal communication, 1971.

Masling, J., & Stern, G. "Effect of the Observer in the Classroom," *Journal of Educational Psychology*, 1969, 60, 351-354.

McKeachie, W. J. *Teaching Tips: A Guidebook for the Beginning College Teacher*, (6th ed.) Boston, Heath, 1969.

McKeachie, W. J., Lin, Y. G., & Mann, W. "Student Ratings of Teacher Effectiveness: Validity Studies," *American Educational Research Journal*, 1971, 8, 435-445.

McKeachie, W. J., & Solomon, D. "Student Ratings of Instructors: A Validity Study," *Journal of Educational Research*, 1958, 51, 379-382.

(References continued)

Melnick, M. (Ed.), *Course Evaluations*, Higher Education Research Abstracts No. 7, Center for the Study of Higher Education, Hofstra University, 1969, mimeo., 9 pps.

Parent, E., Vaughn, C. E., & Whatton, K. "A New Approach to Course Evaluation," *Journal of Higher Education*, 1971, 42, 133-138.

Pitts, J. R. "Can Students Evaluate Faculty?" *Academics*, 1970, 2, 3-4.

Popham, W. J. "The Performance Test: A New Approach to the Assessment of Teaching Proficiency," *Journal of Teacher Education*, 1968, 19, 216-222.

Popham, W. J. "Performance Tests of Teaching Proficiency: Rationale, Development, and Validation," *American Educational Research Journal*, 1971a, 8, 105-117.

Popham, W. J. "Teaching Skill Under Scrutiny," *Phi Delta Kappan*, 1971b, 10, 599-602.

Remmers, H. H. "Rating Methods in Research on Teaching," In N. L. Gage (Ed.), *Handbook of Research on Teaching*, Chicago: Rand McNally, 1963, 329-378.

Rosenshine, B. "To Explain: A Review of Research," *Educational Leadership*, 1968, 26, 303-305, 307, 309.

Slobin, D. Y., & Nichols, D. G. "Student Rating of Teaching, *Improving College and University Teaching*, 1969, 17, 244-248.

Solomon, D., Rosenberg, L., & Bezdek, W. "Teacher Behavior and Student Learning," *Journal of Educational Psychology*, 1964, 55, 23-30.

Stecklein, J. E. "Colleges and Universities: Appraisal of College Teachers," In C. W. Harris (Ed.), *Encyclopedia of Educational Research*, (3rd ed.) New York: Macmillian, 1960, 268-288.

University of Hawaii, Department of Linguistics. Mimeo., November, 1970.

University of Hawaii, Faculty Senate, Faculty Personnel Committee. Mimeo., June, 1970.

Walker, C. R., & Fischer, G. A. "Assessment of the Teaching-Learning Process at the College Level: A review of Research," *Journal of Research Services*, 1965, 5, 20-31.

Washington, E. "The Expert Teacher Action Study: A New Approach to Teacher Evaluation," *Journal of Teacher Education*, 1970, 21, 258-263.

Frederick T. Bail is Assistant Researcher in the Education Research and Development Center and Assistant Professor of Educational Psychology, College of Education, University of Hawaii. He holds an A.B. in Psychology from Bowdoin College and a Ph.D. in Educational Psychology from Cornell University.

Peter Dunn-Rankin is an Associate Researcher in the Education Research and Development Center and an Associate Professor of Educational Psychology. He holds a B.S. in Secondary Education and an M.S. in Language Arts from Florida State University, an M.A. in Mathematics from Louisiana State University, and an Ed.D. in Educational Research from Florida State University.

Both authors have had a continuing interest in the assessment of teacher effectiveness.