DOCUMENT RESUME

ED 221 147 HE 015 512

AUTHOR Benton, Sidney E.
TITLE Rating College Teaching: Criterion Validity Studies of Student Evaluation-of-Instruction Instruments. AAHE-ERIC/Higher Education Research Report No. 1, 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 Academic Achievement; *College Instruction; Criterion Referenced Tests; Educational Research; *Evaluation Criteria; Evaluation Methods; Higher Education; *Research Methodology; Student Evaluation; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; *Validity

ABSTRACT

Studies on the criterion validity of student evaluation-of-instruction instruments are analyzed, and recommendations are offered for future research into student evaluation of instruction. The main problem, and probably the reason for the lack of validity studies, is that it is difficult to agree on what the criteria of effective teaching should be. One method of dealing with the problems of research in student evaluation-of-instruction instruments is to select a measurable definition of teaching effectiveness. Since the ultimate criterion of teaching effectiveness is student learning, there is general agreement that an appropriate and defensible criterion is the amount that students learn as measured by achievement examinations. Attention is directed to: studies incorporating achievement scores and random assignment; studies incorporating achievement scores adjusted for ability; studies incorporating achievement scores not adjusted for ability; and a meta-analysis of student ratings and student achievement. Studies using criterion measures in conjunction with achievement and studies using criterion measures other than achievement are also reviewed. Tables are presented to summarize the studies that examined the relationship of student ratings of instruction and criterion measures. Although parallel data were not reported in all the studies, the table shows the largest significant correlation reported in each study. These largest correlations are squared to indicate the proportion of variance shared by the criterion and the student ratings. The majority of the investigations reported significant positive correlations between student ratings of instruction and criterion measures of effective teaching; however, the correlation between the ratings and criteria were usually modest. A bibliography is appended. (SW)

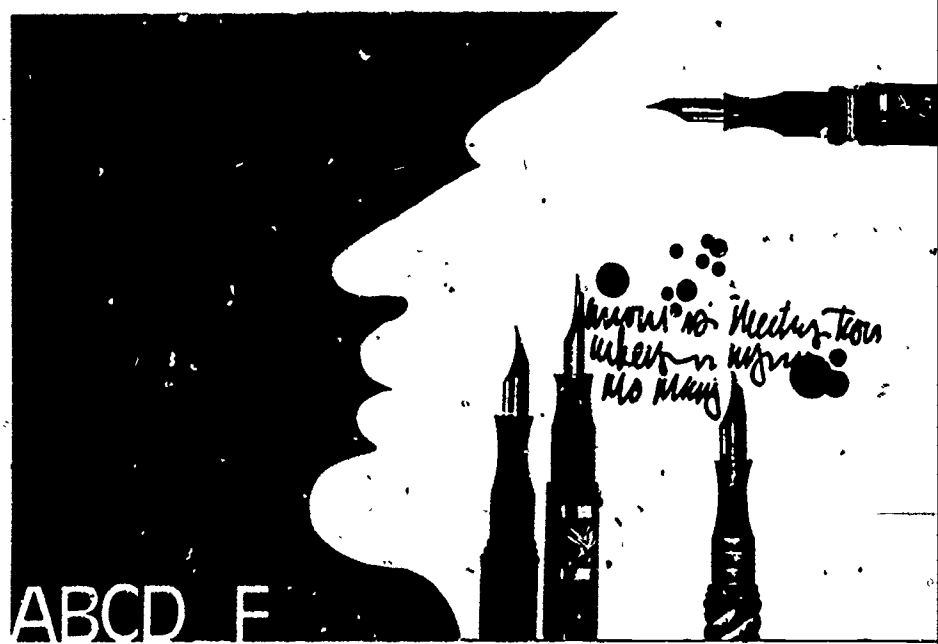
Report *1982*

1

Rating College Teaching:

Sidney E. Benton

Criterion Validity Studies of Student Evaluation-of-Instruction Instruments



ERIC®
AAHE

Rating College Teaching: Criterion Validity Studies of Student Evaluation-of-Instruction Instruments

Sidney E. Benton

AAHE-ERIC/Higher Education Research Report No. 1, 1982

Prepared by

ERIC®

Clearinghouse on Higher Education
The George Washington University

Published by

AAHE

American Association for Higher Education

Cite as:
Benton, Sidney E. *Rating College Teaching: Criterion Validity Studies of Student Evaluation-of-Instruction Instruments*. AAHE-ERIC/Higher Education Research Report No. 1, 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

Foreword

A great number of factors are pushing colleges and universities to examine the ways they evaluate their faculty. Enrollment changes and an immobile professoriate mean that institutions must tighten up their tenure and promotion policies. Limited financial resources constrain their ability to award merit raises and other perquisites. Changing patterns of student enrollment force them to consider terminating selected programs or faculty members. And the increase in litigation of personnel issues demands that they have definite policies and procedures for making such decisions. Finally, many institutions view evaluation of teaching as a way of helping their faculty develop skills, rather than only as a rating mechanism.

One of the primary methods colleges have used to evaluate faculty has been the questionnaire in which students rate the instruction they have received in their classes. In recent years, a large number of studies have looked at the criterion validity of student evaluation-of-instruction instruments. In this Research Report, Sidney E. Benton, professor of education at North Georgia College, analyzes these studies and presents a number of their problems and weaknesses, as well as their strengths. In doing so, he also makes recommendations for future research into student evaluation of instruction and how it can more properly serve the purposes it is designed to accomplish.

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

Acknowledgments

The author is indebted to Jonathan D. Fife and the staff of the ERIC Clearinghouse on Higher Education for their assistance and guidance during the writing of this monograph. The author is grateful to Owen Scott, University of Georgia, who introduced the author to the area of student evaluations of college instruction and encouraged him to develop a body of work in the area. The author wishes to thank Christopher Sharp, North Georgia College, who has continually urged the author to practice good teaching as well as to research and write about it. A special thanks is due to Bob Jerrolds, University of Georgia, who provided invaluable assistance in editing and reviewing the writing. Finally, Andrea Byrd and Debbie Eaton, secretaries in the Department of Education at North Georgia College, gave generously of their time in typing the various drafts of the monograph.

Sidney E. Benton
North Georgia College

Overview

Let such teach others who themselves excel,
And censure freely who have written well.

Alexander Pope
An Essay on Criticism I

A number of procedures are used to measure instruction in higher education. These include evaluation by colleagues, appraisals by the dean or department head, evaluations by means of audio or video tapes, appraisals of the instructor's course material by a faculty committee, and student evaluations of instructors.

Of these procedures the use of student evaluations has gained the most support. Writers have pointed out that these evaluations are made by those who have actually experienced the teaching. Student evaluation-of-instruction instruments are widely used and written about.

These student ratings have been used primarily to improve instruction and to make decisions about faculty tenure, promotion, and merit pay. The basic assumption behind this use is that such ratings provide evidence of quality teaching.

Many faculty members, however, criticize the use of student rating forms, especially in matters of tenure, promotion, and pay. Faculty resistance to the use of these forms stems from the fact that many rating forms have been prepared by groups or individuals who merely sat down and developed items that in their judgment had face validity with respect to measuring effective teaching behaviors. Repeatedly college instructors point out that insufficient attention has been given to criterion validity checks. Criterion validity is perhaps best defined as "the extent to which test performance is related to some other valued measure of performance" (Gronlund 1981, p. 72). In this case, "test performance" is the students' ratings of their instructor on a student evaluation-of-instruction instrument. The "valued measure of performance" is some other measure of the instructors' teaching effectiveness. These other measures typically have been students' scores on a course examination, student gain scores, students' scores on national examinations, students' interest in advanced courses, ratings of video tape clips, and ratings by trained observers. If the effectiveness of an instructor is to be evaluated in any part by student evaluation-of-instruction instruments, it should be important to examine this relationship between the results of the ratings on such instruments and good teaching performance as indicated by other measures deemed to be valid.

Since the ultimate criterion of teaching effectiveness is student learning, there is general agreement that an appropriate and defensible criterion is the amount that students learn as measured by achievement examinations. The majority of criterion-validity studies reviewed involved the use of such examinations for establishing criterion validity. One of the problems in such studies involves finding courses that have a large number of sections with a common examination. Such requirements are necessary to avoid statistical and research design problems. Statisticians generally

agree that relationships based on a small number of course sections are apt to be unstable.

When courses with a large number of sections are located, it is not easy for the researcher to ensure that the students in all the sections have the same aptitude at the beginning of the courses. If the sections do not consist of students with this equal aptitude at the beginning of the course, the researcher must assume that any differences extant at the end of the course might have resulted from the initial differences rather than from the differential effects of the teaching. Random assignment to sections provides the best assurance that groups are equal at the onset of a study. However, in many situations the researcher cannot make such arbitrary assignments. When randomization is not possible, many researchers have statistically adjusted for ability. Some have ignored that such differences may exist. Other researchers have had students select sections without knowledge of who the instructor will be, thus reducing a possible systematic bias in the selection process and giving some assurance that the groups are equal at the beginning of the study.

Other investigations have involved measures other than course examinations in establishing criterion validity. These studies are reviewed and discussed in a second section of this monograph.

It should be noted that a number of very weak studies are discussed in this monograph with the weaknesses delineated. There are two reasons for including these weak studies. In the first place, they shed some light on the subject at hand, even though the data base is weak. To ignore these studies would be to eliminate some important information. It has been pointed out:

> *A common method of integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies . . . and then advance the one or two "acceptable" studies as the truth of the matter. This approach takes design and analysis too seriously. . . . To integrate research results by eliminating the "poorly done" studies is to discard a vast amount of important data. (Glass 1976, p. 4)*

Secondly, these weak studies are often cited in books and articles on student evaluation of instruction without drawing attention to their weaknesses. Frequently in reviews of the literature of individual articles the statistical results of related pieces of research are summarized in a sentence or two, but the limitations of the research are not mentioned. Thus, these "findings" become incorporated into the mainstream of education thought and practice without their legitimacy being questioned. Specifically, some weak studies are being used inappropriately by colleges and universities to justify or reject the use of specific instruments or student evaluation-of-instruction instruments in general.

A review of the literature suggests four major observations. The first of these observations on criterion validity studies is that the majority of the investigations reported significant, but modest, positive correlations

between student ratings and criterion measures held to be indications of effective teaching. The synthesis of the findings of the studies indicates that student evaluations are tapping into an important dimension of teaching. Therefore, there is a legitimate basis in using them to evaluate the performance of college teachers.

A second major observation is that, overall, the findings are highly inconsistent, with a range of significant correlations reported between −.75 and .96. However, since the correlations are not highly positive, it must be recognized that there is a great deal more to instruction than is accurately reflected in student evaluations. Such evaluations should be an important part of an overall assessment of an instructor's teaching performance, but it would appear that an administrator or committee that makes decisions about a professor's teaching based on student evaluations alone is on shaky ground indeed.

There are at least eight possible reasons for the inconsistent results reported in the various studies. These include: small sample sizes, a diversity of the types of courses using the evaluation forms, the number of types of evaluation forms used, the failure to distinguish who was being evaluated (teaching assistants or full-time professors), a lack of standardized procedures for the administration of forms, the use of criterion measures with unknown psychometric properties, a lack of adequate control for initial ability of students in various course sections, and differences in the times during the course the evaluation forms were administered. It seems reasonable to expect that in practice student evaluations would parallel the research. Those who use student evaluations must realize that such evaluations will vary a great deal, according to whether the class is large or small, whether the course is of one type or another (basic or advanced, theoretical or practical, elective or required, etc.), whether the types of evaluation forms fit the types of instructional procedures used, and whether the instructor is a teaching assistant or full time professor. Variations can also be expected when the procedures for the administration of the instrument are not standardized, when the psychometric qualities of the instrument itself are lacking, when the students are of differing abilities and attitudes, and when the instrument is administered at different points during the course.

A third major observation is that there is an identifiable trend in the frequency with which certain student rating variables emerge as significant predictors of effective teaching. Although an overall evaluation of instruction item or an overall score is often listed as an indicator of effective teaching, neither is generally useful to instructors as an aid for improving their teaching. The two specific categories or factors that emerge most often in studies as significant predictors of effective teaching relate to the skill of the instructor and organization and planning. Instructors and evaluation committees should, therefore, pay particular attention to their ratings on items that reflect these two factors.

The fourth major observation from the review of the literature is that a definite need still exists for more studies of student evaluation of in-

struction instruments. With the increasing demand for objective data in the evaluation of professors, there is every reason to expect that these instruments will continue to be used. If so, it is in the best interests of higher education in this country that we learn more about these instruments so that they can be used more fairly and justly.

The Uses of Student Evaluation Instruments

A number of procedures are used to evaluate college instruction. These include ratings by colleagues, appraisals by deans or department heads, evaluations by means of audio or video tapes, appraisals of the course material by faculty committees, and evaluations by students. The central purpose of this monograph is to examine studies that relate to student evaluations and to make suggestions for the uses of these evaluations as well as suggestions for future studies of student evaluations. First, however, the other procedures used to evaluate college instruction mentioned above will be discussed briefly.

The evaluation of an instructor's performance by colleagues and administrators has been criticized. In such evaluations:

> *those raters seldom have observed the individual in the classroom. Therefore they base their ratings of his teaching on his performance in rather different situations and/or on statements made by some of his students. These students may or may not be a representative sample of the teacher's classes. Further, the sample may or may not be comparable from one teacher or another. (Voeks 1962, p. 212)*

There is a danger that in these evaluations the rater will "screen the teacher's performance too much through his own selective perceptions of what constitutes good teaching" (Miller 1974, p. 31). This caution is applicable to classroom visitations by superiors and colleagues and to the use of audio and video tapes.

Evaluation of the instructor's course material also has its failings. It is easy for an instructor to get together an impressive syllabus, an array of objectives, and a list of readings for an appraised committee. This set of materials may bear little relationship to what goes on in the actual teaching situation. Although many who produce such materials are also good teachers, there is no guarantee that these materials accurately represent a good teacher.

In recent years the use of student evaluations of instructors has gained much support. "Of several procedures, the student instructional rating approach is apparently being marketed most vigorously" (Frey 1973a, p 3). "With increased demand for more careful assessment of teaching, administrators are incorporating student ratings of instructional effectiveness into their personnel decisions" (Sheehan 1975, p. 687). Another position is that "student ratings constitute one of the most credible indicators of professorial performance available" (Scott 1975, p. 445).

It also has been pointed out that student evaluations of instructor effectiveness are made by those who have actually experienced the teaching. "Students are the only persons who see the teacher day after day in the classroom. They are not experts on how to teach, but they can furnish valuable evidence concerning the way their teachers teach" (Hayes 1963, p.168). Both the importance and the usefulness of the opinions of students concerning their instructors have been emphasized by a number of sources.

What seems to be most lacking in current practices is carefully accumulated information about a teacher's actual performance. Student opinion is of particular importance here because it represents an important addition to the data customarily used to judge faculty competence. It is the one source of direct and extensive observations of the way teachers carry out their daily and long-range tasks. (Eble 1971, p. 14).

Student evaluations of instructors have three major uses: to help institutions make decisions about faculty tenure and promotion, to help students select courses or instructors, and to provide information that instructors can use in changing their courses or teaching methods (Centra 1980; Blount, Gupta, and Stallings 1976). The need for evaluation of teaching definitely exists if for no other reason than to improve teaching performance.

It has been suggested that many faculty members *do* use the ratings for purposes of course improvement and self-improvement (Romine 1973). "Many faculty regard student evaluation of their courses as an indication of their teaching success, and may actually allow the results to shape their subsequent pedagogical behavior" (Bausell and Magoon 1972, p. 1013).

Many faculty members criticize the use of student rating forms. However, after conducting one of the most comprehensive reviews of the empirical studies pertinent to these criticisms, Costin, Greenough, and Menges (1971) concluded that these ratings can provide reliable and valid information on the quality of courses and instruction. However, they point out that "faculty resistance to the use of student rating forms may stem partially from the fact that many rating forms have been prepared by groups or individuals not qualified to construct such instruments" (p. 511). This claim seems founded. According to Miller, "Too many procedures for evaluation consider only the first step, the development of evaluative criteria" (1974, p. 15).

In the past, many of these forms were constructed by people who merely sat down and developed items that in their judgment had face validity with respect to measuring effective teaching behaviors. In many instances, insufficient attention was given to the rationale for devising items, to revision of the items, and to reliability and criterion validity checks. Many student rating forms are considerably lacking in attention to predetermined criteria as a basis of their construction (Costin, Greenough, and Menges 1971; Miller 1974).

Today is the "Age of Litigation" for institutions of higher education. Those who make decisions about faculty salaries, tenure, and promotions have to be able to produce evidence to support their decisions. In the search for data that can be so employed, they have frequently mandated the use of student evaluation-of-instruction instruments. Those who use, or require the use of, such instruments often know little about the development or the psychometric properties of the instruments they choose. Since such critical decisions are affected by the use of these instruments, it is important to learn as much about them as possible.

Selection of a Student Evaluation-of-Instruction Instrument

There are a great number of reports in the literature on instruments for student evaluation of instruction and their use. Pettman's (1972) annotated bibliography on student evaluation articles published betwen 1965 and 1970 listed 107 articles. Biddle's (1980) annotated bibliography, incorporating the ERIC files dating between 1976 and 1978, listed approximately 280 items.

An extensive search of the literature to identify the best and most widely usable student evaluation-of-instruction instruments was made by Benton (1974). His search was guided by three predetermined criteria: (1) the instruments had to be applicable to the various academic areas, not specific to one area (such as psychology), (2) the instruments had to be designed to evaluate college and university teaching, and (3) the instruments had to have been designed to provide information that could be used to improve instruction. Only 39 instruments, of the hundreds reported, were located that met even these very basic criteria. Since 1974, a number of other instruments have appeared in the literature, but only a limited number of them meet these criteria.

Considering all the checklists and other forms available for students' evaluations of instruction and the amount of literature available on the topic, it is understandable that instructors, faculty committees, and administrators find it difficult to select one instrument in which they can have confidence.

Problems of Criterion Validity

One criterion suggested in selecting an instrument for student evaluation of college instruction was that "validity, beyond simple content validity, has been substantiated" (Benton 1979, p. 15). The type of validity appropriate in this case is called criterion validity, sometimes referred to as empirical or statistical validity. It is defined as the degree to which scores on the instruments for student evaluation of instruction are in agreement with some given criterion measure of effective teaching. Although it is easy to say that student evaluation instruments should have criterion validity clearly established, it is not easy to find studies that report such information. Validity is one of the typical faculty concerns in the use of such instruments (Aleamoni 1974).

To establish criterion validity of instruments of student evaluation of instruction, the following three steps generally are involved:

- The instrument is administered to a group of individuals.
- A criterion measure of effective teaching is obtained.
- The two measures are correlated.

The resulting correlation, or validity coefficient, is an indication of the criterion validity of the student evaluation instrument. The range of the coefficients can be from .00 (indicating no relationship between the two measures) to 1.00 (indicating a perfect relationship between the two mea-

sures). The closer the correlation is to 1.00 the higher the criterion validity. No student evaluation-of-instruction instrument is expected to have a perfect criterion validity coefficient; therefore, predicting teaching effectiveness based on these instruments will always be somewhat imperfect. However, the larger the validity coefficient, the less the error in predicting effectiveness and the more effectively the two measures reflect each other.

The chief problem in establishing criterion validity is the difficulty in obtaining a satisfactory criterion measure (Thorndike and Hagen 1969). *So the main problem, and probably the reason for the lack of validity studies, is that it is difficult to agree on what the criteria of effective teaching should be.* "Validating student ratings at the university level is difficult since there are no clearly defined criteria of instructional quality" (Marsh, Fleiner, and Thomas 1975, p. 833). "Validating a measure of a construct like teaching effectiveness requires the use of many alternative criteria" (Marsh 1977, p. 442).

"Most studies of validity have used correlations with peer ratings or supervisor ratings as the criterion" (Sullivan and Skanes 1974, p. 584). However, what is needed is a focus on criterion validity studies that relate to the direct outcomes of effective instruction. Therefore, as stated at the beginning of this chapter, the central purpose of this monograph is to examine studies that relate to these outcomes and to give suggestions regarding the present uses of student evaluations and suggestions regarding future studies of student evaluations.

Student Achievement and Evaluation of Instruction

One method of dealing with the problems of research in student evaluation-of-instruction instruments is to select a measurable definition of teaching effectiveness. Since the ultimate criterion of teaching effectiveness is student learning, there is general agreement that an appropriate and defensible criterion is the amount that students learn as measured by achievement examinations.

One of the usual approaches to studies that examine the relationship of student evaluations of instruction and student achievement is for the researcher to select a course that has several sections taught by different instructors but has a common examination. "In this case there is an agreed upon, measurable, and common educational outcome which can be used as a criterion of teacher effectiveness" (Smultz 1978, p. 15). For each section of the course, the mean examination score is then correlated with the mean of the students' ratings of instruction. A significant positive correlation is held to be empirical evidence of the criterion validity of the evaluation instrument.

There are several problems inherent in such an approach to establishing criterion validity. One problem is that courses with a large number of sections and a common examination are difficult to find even in large universities. Also, there are many studies that involve a large number of student responses but compare only a small number of instructors. These comparisons are likely to be unstable. Even if the conditions of goodly numbers of sections are met:

> *the statistical tests are generally not very powerful. With 10 different sections, a validity coefficient would have to be .55 to reach even the .05 level of significance. Extremely high validity coefficients cannot be expected since performance depends upon many variables besides instructional quality and evaluations depend upon many factors besides learning measured by a final examination. (Marsh, Fleiner, and Thomas 1975, p. 834)*

Another problem is that even when courses with large number of sections are located, it is not easy for the researcher to ensure that the students in the various sections have the same aptitude at the beginning of the course.

> *If it cannot be demonstrated that predisposing factors such as student ability and motivation have been equated across the different sections of the multisection course, then it may be these variables that produce the correlation between student ratings and exam performance. (Marsh and Overall 1980, p. 469)*

In order to compensate for these possible initial differences in aptitude, various researchers have randomly assigned students to course sections, statistically adjusted for initial ability, or had students select course sections without knowledge of who the instructor was to be. Some researchers simply have ignored the existence of differences in course sections.

Studies Incorporating Achievement Scores and Random Assignment

The best method of controlling for possible initial differences among various class sections would be to randomly assign students to the sections. Centra (1980) suggests that "randomization of students is one of the steps needed to draw a cause and effect relationship between rated teacher effectiveness and student learning" (p. 37). However, researchers in college settings rarely are able to do this. Only two studies were located in which students were randomly assigned to the sections.

Sullivan and Skanes (1974) used 130 sections of ten courses at Memorial University of Newfoundland, Canada. Students were randomly assigned to sections in each of the courses. Sullivan and Skanes reported low to moderate correlations for mean instructor ratings and mean final examination scores for the ten courses. Of the ten correlations, eight were above .32, and the average correlation was .39. However, only two of the ten correlations and the average correlation were significant. The researchers pointed out that one possible reason for the small correlations was that the range for the two variables was restricted. The overall rating for the instructors was based on a five-point scale, and there was little variability in the examination scores.

One of the major strengths in the Sullivan and Skanes study, other than the random assignment of students, was the development and scoring of the final examination for the ten courses. Examination committees constructed the examinations and set guidelines for grading each answer. The examinations were scored by boards with a "small group of faculty members marking one answer on all papers" (p. 585). The student evaluations were done anonymously. The correlations involved only a global rating of instructor effectiveness rather than a number of dimensions.

The study involved two different biology courses, and one course each from physics, psychology, and science. Two of the courses had six sections, two had eight sections, two had nine sections, and two had 14 sections. The remaining two courses had 16 and 40 sections.

In the course that had 40 sections, the correlation between instructor ratings and examinations was .41. When the correlation was calculated for two subgroups, 27 full-time instructors and 13 part time teaching assistants (TAs), the correlation was .53 for the full time instructors and .01 for the part time TAs. When the amount of experience was considered for the 27 full-time instructors, the correlation between ratings and achievement for experienced faculty (one or more years of full time teaching) was .69, but for the inexperienced (those in their first year of full-time teaching) the correlation was .13. Sullivan and Skanes thus suggest that their results may provide some answers to some of the contradictory results of previous studies. They further conclude, "valid ratings are much more common and are easily obtained in the case of experienced and full-time instructors than in the case of inexperienced or part time instructors (p. 587). Again because of the size of these subgroups, the data must be regarded as tentative. It would seem appropriate to do further research in the areas of ratings of full time versus part time instructors and experienced versus

inexperienced instructors and the relationship of the achievement of the students.

A second study in which true random assignment of students to sections was employed was reported by Centra (1977). The study also included sections of courses in which randomization was not used. In the Centra study there were 72 sections of seven courses. In two of the seven courses, a biology course and a chemistry course, students had been randomly assigned to sections. As in the Sullivan and Skanes (1974) study, the subjects were from Memorial University in Newfoundland. Instead of a single global item, Centra used nine variables from the Student Instructional Report (SIR). These variables were. "Overall Teaching Effectiveness," "Value of Course to Student," "Teacher-Student Relationship," "Course Objective and Organization," "Reading Assignments," "Course Difficulty and Workload," "Examinations, Lectures, and Student Effort." For the two courses in which the students had been randomly assigned to the sections, the highest correlations with mean final examination performance were for the area of Value of Course to Student. The correlations were reported to be .73 and .92 for the two courses. Other significant correlations reported were .81 (Examinations) for the biology course, and .76 (Lectures) and .79 (Student Effort) for the chemistry course.

In the Centra study almost all the instructors were experienced teachers, none were graduate teaching assistants. The final examinations were developed and scored as in the Sullivan and Skanes study. However, again the results of the study must be interpreted with caution. Of the two courses that had students randomly assigned to the sections, there were only seven sections of each course. Regarding all the 72 sections Centra concluded:

> *The pattern of correlations across the courses indicated that the global ratings of teacher effectiveness and of the value of the course to students were most highly related to mean exam performance (12 out of 24 product-moment and partial correlations were .58 or above). Ratings of course objectives and organization and the quality of lectures were also fairly well correlated with achievement. Ratings of other aspects of instruction, such as teacher-student relationship or the difficulty/workload of the course, were not highly related to achievement scores. (p. 17)*

Studies Incorporating Achievement Scores Adjusted for Ability

Since students usually know who the instructor will be when they select their course section, it is possible that different sections could differ markedly in student abilities and attitudes. For example, the best students might choose the teachers with a reputation for good teaching and high standards for students. The poorer students might select the teachers who are less demanding and who give higher grades. A comparison of two such sections, thus, would be contaminated by the way the students came to be in those particular sections in the first place.

When sections are unequal in abilities and attitudes at the beginning

of a study, some statistical adjustment is necessary to account for these initial differences. The process is usually one of computing residual scores; that is, scores statistically adjusted for initial ability or attitude. Researchers generally use some nationally normed aptitude test (e.g., Scholastic Aptitude Test), a pretest in the course area, or grade point averages to compute residual examination scores. Although the procedure is defensible, simply adjusting scores statistically is not as satisfactory as random assignment of the students to sections. However, in the studies reviewed in this section, the researchers made some adjustment for ability in a portion of their study.

One of the most controversial investigations cited in the literature is a study by Rodin and Rodin (1972). The study is cited first in this section because it has had so much visibility; the results have provoked much discussion and some of the studies cited later were conducted as a reaction to the Rodin and Rodin findings. Rodin and Rodin reported a strong *negative* correlation between achievement and instructor ratings. Rodin and Rodin used teaching assistants in an undergraduate calculus course. The students met three days a week for a lecture with a professor, and on the remaining two days met with individual teaching assistants in 12 small sections. The teacher rating form used in the study was not specified; moreover, only the responses to one question on the form were used in the analysis. The question was, "What grade would you assign to his total teaching performance?" Numbers were assigned to these ratings (A = 4 to F = 0). A measure of the students' initial ability in calculus was obtained from the previous quarter. Mean grades in the course for the 12 sections and mean section ratings were used in the calculation of a partial correlation. This partial correlation between the objective measure (the grade determined by the number of problems passed) and the subjective measure of teaching ability (the one question on the student evaluation), with initial ability held constant, was −.75. "The instructors with the three lowest subjective scores received the three highest objective scores. The instructor with the highest subjective rating was lowest on the objective measure" (p. 1165). The researchers concluded, "Students rate most highly instructors from whom they learn least" (p. 1164).

Many researchers (Bryson 1974, Frey 1973a, 1973b, 1978; Gessner 1973; Marsh, Fleiner, and Thomas 1975, Rippey 1975) have cited methodological problems in the Rodin and Rodin study. One problem is that it is "inconsistent with common sense as well as with accumulated results of previous research on this topic" (Frey 1973a, p. 4). Another weakness is that the research had assessed the effectiveness of graduate teaching assistants (TAs) who has only complemented the activities of the professor (Frey 1973a). It should be noted specifically that (in contrast to a great many other studies where the TAs were, indeed, the instructors for the courses) these TAs, though designated as instructors, really were only assistants who had a minor role in instruction.

Frey (1973b) investigated the conclusions of Rodin and Rodin in his study examining two different calculus courses that had a regular faculty

member and teaching assistants. The students met with the faculty member three times a week for lectures and with a teaching assistant once a week for a quiz. Each course had a common syllabus and a common final examination. Approximately 75 percent, or 354, of the students completed an instructional rating form used at Northwestern University. The form was mailed to the students who had completed the examination and whose Scholastic Aptitude Test (SAT) scores were on file at the university. The average final examination score for each instructor (adjusted for initial difference in sections using composite SAT scores) was the criterion for validation of the student rating. One special strength of this study concerns the reliability of the grading system. All examination papers were scored in a common session with an instructor grading the same item for all sections.

Frey factor analyzed instructor ratings using individual responses from these and other classes and found six factors, indicating that the evaluation form was measuring six different areas. A Pearson product-moment correlation was calculated between the adjusted final examination score and each of the six factors. In the introductory calculus course (eight instructors), three factors—"teacher's presentation," "organization-planning," and "student accomplishment"—showed high positive correlations with the regressed final examination scores. .91, .87, and .84, respectively. In the multidimensional calculus course (five instructors) the correlation between "student accomplishment" and the examination was .90. When the correlations for the two calculus courses were averaged, the "student accomplishment" factors and the "teacher's presentation" factor were the highest predictors of achievement (.87 and .75). "Teacher accessibility" and "work load" correlated the lowest (.31 and .44).

A weakness in Frey's study was the small number of sections involved in the analyses. Frey admitted "correlation coefficients based on such a small number of observations are notoriously unstable" (p. 84). The use of factors, obtained by analyzing individual responses, in prediction of class mean achievement scores is also open to question.

Like Frey (1973b), Doyle and Whitely (1974) used examination scores in conjunction with student ratings of college instructors. The premeasure of ability of 174 beginning French students taught by 12 graduate students at the University of Minnesota was the Minnesota Scholastic Aptitude Test. The Student Opinion Survey (SOS), with an addition of seven general items, was used in rating the instructors. Two types of data were included in the study, between-sections data and across-sections data. Between-sections data compared class trends and involved correlations of section means. Across-sections data were from all sections, pooled, and involved correlations of raw item responses. When the seven general items were analyzed across sections (174 students), six of the items had significant correlations with residual examination scores. The correlations ranged from .18 to .25. However, when the same seven items were analyzed between sections (12 instructors), only two of the items had significant correlations (.51 and .49) with residual examination scores. These two items

related to "general teaching ability" and "overall teacher effectiveness." Since the SOS had been factor analyzed using individual responses, the correlations of the factors with residual achievement were done only across sections. Two of the factors, "motivation of interest" and "expository skills," correlated significantly (.36 and .31) with residual achievement.

The Doyle and Whitely study did not provide multiple correlations using factors of SOS to predict residual achievement. Further, the stability of the correlations in the across-section analyses is open to question because of the small number of classes in the study. Also, the seven general items that were added to the SOS must be questioned. No information is given as to the origin of the items and the reasons for their selection.

Another study using student ratings to predict residual achievement was by Turner and Thompson (1974). Unlike the Frey (1973b) and the Doyle and Whitely (1974) studies, in which small numbers of classes were used, the Turner and Thompson investigation used one sample of 16 sections of beginning French students and another sample of 24 sections of beginning French students all taught by TAs. Residual achievement (final examination corrected for first examination) was computed. Members of the French Department selected 30 items from the student rating instrument reported by Deshpande, Webb, and Marks (1970). Five items specific to teaching beginning French were added to this list of 30 items. These items related to the instructor giving students opportunities to speak in French, having a good command of French, having a knowledge of the culture of French-speaking peoples, making pronunciation errors in French, and being enthusiastic about speaking French. Two subscales (labeled "Instructor Cognitive and Affective Merit Versus Student Cognitive and Affective Stress" and "Motivation and Workload") and a total subscale score were then used as the student rating variables. When the two subscales and the total subscale scores were used to predict residual achievement, negative correlations of −.51, −.51, and −.52 were obtained for the first sample and −.41, −.31, and −.41 for the second sample.

Of all the studies herein reviewed, this is only the second case in which a significant negative relationship between ratings and achievement was reported. The authors suggested that the "stress/overload" produced by the instructor was the important factor in obtaining greater residual achievement and that the positive behaviors of the instructor appeared to lead to less residual gain. Since the vast majority of studies in this area show opposing results to those of Turner and Thompson, their study should be noted, but the findings should be viewed with caution. Turner and Thompson concluded:

> *the results of the study suggest that student ratings of college instructors should be treated with great caution by college administrators and promotion and tenure committees. Although such ratings may express student observations of and attitudes toward an instructor, they clearly cannot be routinely interpreted to be positive indicators of student residual achievement in the instructor's course. (p. 3)*

Without a substantial number of other studies with similar negative correlations, it is perhaps most useful to try to determine why this study had such different results from the general body of the literature rather than generalize from this study about the whole question of student ratings. The article itself gives no basis for speculation as to why these results were different. One reason might be because these instructors were all TAs rather than full-time professors. Also, in the Turner and Thompson study not enough information was given concerning the achievement examinations. Although the authors stated that the first examination covered grammar and the final examination covered grammar, dictation, composition, and reading comprehension, they did not state whether the test items were objective, essay, or a combination of the two. The type of items on the test is an important consideration because the scoring of essay items is generally not as reliable (consistent) among various instructors as the scoring of objective items, and no information was reported about this scoring.

In a fifth study, only a portion of the scores used in the analyses was adjusted for ability. Frey, Leonard, and Beatty (1975) collected ratings of instructors from 16 sections of introductory calculus at Northwestern University, ten sections of educational psychology at Purdue, and five sections of introductory calculus at North Dakota State. Each of the three institutions used the Endeavor Instructional Rating Form. At each institution instructors used a common syllabus, textbook, and final examination. A factor analysis of the responses from Northwestern and Purdue indicated similar factors. For three of these factors the correlation with final examination performance was "fairly strong" at the three institutions. The mean correlations for the factors and achievement at the three schools were .59 for "student accomplishment," .58 for "presentation clarity," and .51 for "organization-planning." It should be noted that the best predictor of final examination performance found in any comparison in the study was "organization-planning." (At Purdue this correlation was .85.)

For various reasons the correlational analysis was not based on all the original sections. Four sections were eliminated from the Northwestern data, and one section was eliminated from the Purdue data. The researchers do not specify how many of the instructors were teaching assistants. Mathematics SAT scores were used for the Northwestern analysis to adjust final examination scores for the sections. No adjustment was made for the other sets of data. An overall consideration indicates that the study provides moderate support for the use of student ratings.

The data of the Frey, Leonard, and Beatty (1975) study constituted a "qualitative improvement over that which was available in the Frey (1973b) study" (Scott 1975, p. 445). Apparently this judgment is based upon the increased number of course sections used in the study. In addition, the findings of the study provide:

> *additional support for the contention that at least some information from student ratings is positively related to student achievement, a trend which*

must be substantiated if widespread use of student ratings for merit, promotion, and/or instructional improvement is to be continued. (Scott 1975, p. 445)

Another study concerning the relationship between regressed examination scores and achievement was by Frey (1976). Frey compared the final examination performance of students in seven sections of introductory calculus at Northwestern University to student ratings of the instructors. Randomization was used in assigning subjects in each section to two time-of-rating groups. The researcher compared the mathematics SAT scores for the two groups. Some subjects were reassigned after this comparison to ensure that the two groups were equal in mathematics aptitude. When students signed up for the sections, they did not know which instructor was to teach each section. Students who later requested section changes were "actively discouraged."

Ratings of the instructors were conducted by a mail survey; half the students rated the instructors during the final week of classes and the other half during the first week of the subsequent term. Frey reported that the two different times (before and after the examination) did not significantly affect the ratings of the instructors, although the ratings made after the examination showed a slightly stronger correlation. Results of the study indicated a strong relationship between instructor ratings and final examination scores, based on regressed mathematics SAT scores. The highest correlation reported for the "before exam" group was .90 between "planning" and the final examination. "Student accomplishment," "personal attention," and "presentation skill" were the three best predictors of final examination performance for the "after exam" rating group. The correlations reported using these three aspects of instruction were .83, .85, and .78 respectively, providing reasonably strong validation of student ratings.

Instructors of the seven sections were full-time faculty members who used a common text and a common syllabus. In one group 68 percent returned the rating form, and in the other group 70 percent did so. Frey reported similar mathematics SAT scores and a similar final examination scores for the responders and nonresponders. Frey reported that the evaluation form used in the study, stressing student observation rather than student opinion, was the result of a long development process. The major criticism of the Frey study relates to the small number of course sections.

Whitely and Doyle (1979) also investigated the relationship of student ratings to achievement. The researchers compared the ratings of five professors and 11 teaching assistants of a beginning mathematics course at the University of Minnesota. When the data were calculated for between classes, "overall teaching effectiveness" was significantly correlated with the residualized final examination for the professors (.80), but it was not significantly correlated with achievement for the teaching assistants. The premeasure of ability was the Minnesota Scholastic Aptitude Test (MSAT).

As in previous studies, because of the small sample size, the data of

the Whitely and Doyle study must be interpreted cautiously. In the study the Student Opinion Survey was the evaluation instrument, and the MSAT was the ability measure used to residualize examination scores. Students supplied identification numbers but were assured that the results would be confidential. No information is given in the study about the construction of the final mathematics examination. To ensure reliability in grading, each teaching assistant graded one problem from all the students. The report seems to indicate that the teaching assistants also scored the papers from the professors' section, although this was not specified. Thus, once again, there is some support for the use of student evaluations with full-time professors, but not for their use with teaching assistants.

In one article, McKeachie, Lin, and Mann (1971) reported five studies that pertained to criterion measures and student ratings of instruction. In one study, scores were adjusted for intelligence, but the intelligence test was not identified in the report. All correlations in the reported studies were done using mean section scores on the student evaluation instruments and class mean achievement scores, no multiple correlations were reported.

In the first study, students in 33 (in the table they report 37) sections of general psychology evaluated 17 instructors with the Isaacson et al. (1964) evaluation instrument. Four factors of the instrument, "skill," "feedback," "interaction," and "rapport" correlated significantly (.28, .35, .30, and .42, respectively) with the Introductory Psychology Criteria Test, labeled a "thinking" test.

The study was then replicated with 34 sections of general psychology, and results were analyzed separately for men and women. For a second criterion, 25 items were taken from old examinations to make a "knowledge" test. For males, "interaction" correlated significantly with the "thinking" test (.33), and "overload" correlated significantly with the "knowledge" test (.39). For females "feedback" correlated significantly with both the criterion measures (.33 for the "thinking" test and .40 for the "knowledge" test).

In the second study students in 32 sections of general psychology evaluated 16 instructors. None of the factors was significantly correlated with the "thinking" or the "knowledge" test for either females or males.

In the third study, only six instructors were involved, and the number of sections was not reported. The criterion measures were a multiple-choice test of knowledge and an essay test. "Skill" correlated significantly with the essay test for females (.65). This correlation was the only significant correlation in the study.

The sample of the fourth study consisted of 16 sections of second-year French. Criterion measures of the study were a test of grammar, a test of reading, and a departmentally administered test of oral expression. None of the student rating factors correlated significantly with any of the three French criterion measures for either females or males.

In the final study, 18 advanced graduate students, who were the instructors, were evaluated by their students in sections of introductory

economics. The rating scales used in the study consisted of 12 items with high loadings from the Isaacson et al. (1964) scale plus items previously used in the economics course. The two criterion measures were a numerical grade based on course examinations stressing "thinking" and an economics attitude sophistication change score. For males, "structure" was significantly negatively correlated with the grade (−.41). For females "changes in beliefs" correlated significantly with the attitude sophistication change score (.44) and "skill" correlated significantly with the numerical grade and the attitude sophistication change score (.72 and .43).

The five studies by McKeachie, Lin, and Mann (1971) illustrate a point made earlier—namely, that when one uses different populations, different examinations, and variations in the evaluation instrument (with different factors), one can expect wide variations in the results. Criticism of these five studies as reported by McKeachie et al. mainly has been concerned with what was not reported. In three of the studies the variations of the student evaluation instrument were not described clearly. In some of the studies not enough information was given to determine the worth of the measures of achievement. The researchers report that intelligence was partialled out of the correlations of the first study, but no indication is made of this adjustment in the other four studies. If no adjustments were made concerning initial ability in the section, the results are open to further question. Also, one of the studies is based on a sample of only six sections. In two of the studies the authors specified that graduate students taught the classes; no mention is made of the status of the instructors in the other studies.

Two studies (Canaday, Mendelson, and Hardin 1978, Doyle and Crichton 1978) dealt with adjusted achievement scores and student evaluations, although the main focus of these studies was on other research concerns. The present discussion deals only with those dimensions of these studies that have to do with student achievement as the criterion related to student evaluation.

Canaday, Mendelson, and Hardin (1978) investigated the effect of timing on the validity of student evaluation in a one-section course in anatomy. They reported a significant relationship between the course achievement, as measured by multiple-choice examinations, and the course ratings of students in the College of Medicine, Medical University of South Carolina. The researchers reported a partial correlation of .42 between achievement and ratings, when GPAs were controlled. A 31-item student evaluation instrument was designed for the study, and examination reliabilities were reported to be .81 and .85. Because of attrition (some of the ratings were collected three weeks after the final examination) and other factors, the data of the study were based on only 93 of the original 158 students, but the study does lend moderate support to the use of student ratings.

Doyle and Crichton (1978) investigated the relationship of student, peer, and self evaluations to student achievement. They had usable data from 263 student ratings of 12 instructors in a course in introductory

communications. Most of the instructors were graduate students. No student, peer, or self-evaluations of instruction correlated significantly with residual examination scores. Final examination scores were adjusted by using verbal scores from the Preliminary Scholastic Aptitude Test. The student evaluation instrument consisted of four items from factors identified by Doyle and Whitely (1974) plus two overall evaluation items. Thus, once again, ratings using mostly teaching assistants as instructors were not related to achievement.

Finally, Benton and Scott (1976) did not calculate residual achievement scores, but used self-reported grade point averages (GPAs) as one of the independent variables in the calculation of a multiple correlation. Benton and Scott selected two instruments, the Student Instructional Report (SIR) by Centra (1972) and the Inventory of Student Perceptions of Instruction (ISPI) by Scott (1973), that best exemplified the rational and empirical approaches to developing student evaluation-of-instruction instruments. These two instruments were administered at the University of Georgia in 31 sections of freshman English that had a common final examination. A random half of each class was given the SIR and the other half, the ISPI. Students were asked to supply their identification numbers and were assured that the results would be confidential. Mean self-reported GPAs and two empirical sections of SIR (labeled "adjustment of individual needs" and "work load") were statistically significant predictors of class mean examination performance. The multiple correlation obtained using the self-reported GPAs and the sections of SIR as predictors was .62. There was no empirical section or rational section of ISPI or combination of sections with self-reported GPAs that contributed significantly to the mean final examination scores of the English classes. (The largest multiple-R obtained was .42). The authors suggest that results of the study lend some support to the use of instruments developed empirically over those developed rationally.

There are certain problems inherent in the design of the Benton and Scott study that may have influenced the lack of relationship between student ratings and final examination scores. One problem involved the lack of anonymity of the ratings. Students may have responded differently if they had not been required to supply their identification numbers. Another factor that may have influenced the lack of relationship was the use of the common essay examination. Even though the researchers gave each instructor a list of recommendations for scoring essay examinations, it may be that the scores given by each instructor did not truly reflect achievement in the course. Benton and Scott did compare the means of self reported GPAs with actual GPAs of a randomly selected portion of the sample. The means were not significantly different, and the correlation of actual GPAs and self-reported GPAs was .94, indicating that the use of self-reported GPAs in research of this nature is a defensible procedure.

All in all, when achievement scores adjusted for ability are correlated with student ratings, most studies have found a great deal of variability but enough of a relationship to warrant the use of student evaluation

instruments for full-time professors. However, there is little support for using these instruments to examine the instruction of teaching assistants.

Studies Incorporating Achievement Scores Not Adjusted for Ability

In all the previously mentioned studies, students were either randomly assigned to course sections or the researchers incorporated some measure of ability to adjust examination scores. Simply adjusting scores for ability is not an entirely satisfactory substitute for student randomization. Course sections can be significantly different in other variables, such as motivation. If the students in one course section are more highly motivated than students in another section, they may spend more time preparing for the examination regardless of the deficiencies in the instruction. It is further suggested that many:

> *researchers probably misuse ability pretests when residualizing achievement and may remove from section-to-section achievement variation the portion produced by differences in teaching ability in addition to the portion produced by differences in student ability. (Leventhal, Perry, and Abrami 1977, p. 363)*

In spite such criticism, the researchers discussed in the previous section did make some attempt to compensate for differences in ability of course sections. In the following studies (Orpen 1980; Bendig 1953; Cohen and Berger 1970; Bryson 1974; Costin 1978; Hsu and White 1978; Blass 1974; and Endo and Della Piana 1976) apparently no attempt was made to adjust achievement scores of course sections. Because of the possible initial differences in the sections, the reported results should be interpreted cautiously.

Even though no adjustment was made for possible differences in ability of students in ten sections of an introductory course in mathematics, Orpen (1980) did compare mean scores on the aptitude pretest, consisting of a short form of the Scholastic Aptitude Test, and the mean grades the students expected to obtain prior to the final examination. Results revealed no significant differences among the sections on these two measures. The students completed the Teaching Rating Form (derived from the form in McKeachie, Lin, and Mann 1971). Each of the ten sections (taught by different graduate students) used the same content, textbook, and assignments. The common examination was scored by the course director and three specially trained graduate students. Each section's average on the final examination was correlated with the section subscale means on the Teaching Rating Form. Six of the eight correlations, ranging from .52 to .74, were significant. A multiple correlation of .75 was calculated using these six subscales together to predict the examination scores. Even though the results of this study are somewhat equivocal, overall they support the use of student ratings. This result is different from other similar studies where teaching assistants were the instructors.

Even though sections were not significantly different on the aptitude

pretest or expected grades, there are other student characteristics that could have made the sections different. Also, in this study the enrollment for each section was between 10 and 12. This small enrollment is not typical of the other studies that have used multisections of courses, and may be the reason why this study using TAs got basically positive relationships whereas most other similar studies did not.

Bendig (1953) also investigated the relationship between course ratings and achievement in an introductory psychology course at the University of Pittsburgh. Three of the five instructors for this course were predoctoral graduate students. The three achievement tests in the study were all multiple-choice, were used in all the classes, and had been constructed on a departmental basis. Bendig found that correlations between instructor ratings and achievement varied greatly from section to section. Only one of the five ratings correlated significantly with achievement of the students (.37), and only one of the five section ratings correlated significantly with achievement (.46). The total correlation of .28 for the five sections of achievement and course rating was significant. However, the total correlation of achievement and instructor rating was not significant.

The sum of each student's standard scores on the three achievement tests was the criterion measure. Course ratings and instructor ratings were determined by summing students' ratings on the Purdue Rating Scale for Instruction. Students' ratings forms were signed by the students, but they were assured that the instructors would not see the individual forms and that their grades would not be affected by their ratings. The small sample of five instructors greatly limits the findings of the study. The equivocal findings may have resulted from the use of both full time professors and TAs.

Cohen and Berger (1970) reported significant correlations between mean final examination performance and three dimensions and the total scale of the Michigan State University Student Instructional Rating Report (SIRR). The three dimensions of SIRR that were significantly correlated with achievement were "student interest" (.39), "student-faculty interaction" (.37), and "course organization" (.31). The total scale correlated .48 with achievement. None of the dimensions or the total scale correlated significantly with mean class grade point averages at the onset of the study.

The sample of the study consisted of 25 sections of a basic natural science course at Michigan State University. The instructors had a course syllabus designed by the staff. Each instructor was asked to administer the SIRR "at his convenience" within a two-week period to one of his sections. The researchers do not state whether the instructors were professors or TAs. The final examination was a 100-item objective examination that had been validated, and 93 percent of the students who took the final examination completed the evaluation form.

Bryson (1974) also examined the relationship of student ratings and achievement of students. Subjects were 582 students in 20 sections of college algebra taught by 14 instructors who used a common syllabus and

textbook. The mean section scores on each of 14 items of a student rating instrument were correlated with students' performance on the Cooperative Intermediate Algebra Test. All ratings were anonymous, a strength many such studies do not have. Six of the correlations were significant and ranged from .44 to .68.

There are two apparent problems with the Bryson study. First, no attempt was made to adjust for the initial ability in the classes. Second, the evaluation instrument was not named. It was only stated that the items "were selected from a routinely administered faculty and course evaluation form" (p.12). No information is reported about the validity or reliability of the original instrument. If the original had acceptable validity and reliability, the use of a portion of the items without substantiation of such use may have reduced these values.

Costin (1978) reported significant correlations between mean ratings of instructors of an introductory psychology course at the University of Illinois and the mean final examination scores over a four-year time span. The four correlations ranged from .41 to .56. The number of graduate teaching assistants who were in charge of the classes ranged from 21 to 33. Ratings of the instructors were anonymous. The percentage of students rating the instructors ranged from a low of 76 percent to a high of 93 percent for the four years. The final examinations were constructed by the supervisor of the course, and the instructors did not see the examination until it was administered. Although the evaluation instrument remained the same for the four-year period, the final examination in the study was not the same over those years.

One of the criticisms of the Costin study concerns the instrument used for the evaluation of the instructors. Five items were selected from a 46-item instrument reported by Isaacson et al. (1964). Even if the original instrument reported by Isaacson et al. possessed adequate validity and reliability, the use of only five of the 46 items raises serious questions about the reliability and validity of the "new" instrument. No indication of any recheck of reliability was reported. If all 46 items did indeed represent content validity, then the reduction of the instrument to five items probably reduced the content validity considerably. In contrast to a number of other studies, this study does lend some support to the use of student ratings of TAs.

Hsu and White (1978) found significant correlations between achievement scores and students' ratings of instructors on two different evaluation forms. The overall correlations, relating scores with the factors of the instruments, were .74 and .65 for the two instruments. The sample consisted of 308 students enrolled in 12 undergraduate education courses from West Chester State College in Pennsylvania. The instructors of the courses were six full-time professors. The two evaluation instruments were the Inventory of Student Perceptions of Instruction (ISPI) by Scott (1973) and the Instructional Improvement Questionnaire (IIQ) by Pohlmann (1972). In the study, the same graduate assistant used standardized instructions to administer all the student evaluations.

The ISPI was adminstered halfway through the semester, and the IIQ was administered toward the end of the semester. It is possible that results would have been different if the instruments had been administered closer in time. Certainly an evaluation of an instructor can change from the middle of the semester to the end.

Another question arises in the Hsu and White study concerning the achievement measures. Hsu and White state that the first two scores were students' scores on the mid-term examinations and the third was the final examination score. Ordinarily an instructor does not give two mid-terms, so it is not clear how the three measures were obtained. No other information is given regarding the examinations. Also, it is not stated whether the same examinations were given for all the courses. If the courses were really different and common examinations were not used, then the analyses in the study should be questioned. Overall, however, the study provides support for the use of student ratings of professors.

Blass (1974) investigated the relationship between mid-term grades and course evaluation of students who were classified as "subjective" and "objective." The sample of the study consisted of 48 nursing students in an introductory psychology class at Brooklyn College, Brooklyn, New York. When mid-term examination scores were correlated with each of nine student evaluation-of-instruction items for all 48 students, six of nine correlations were significant. The range of significant correlations was from .34 to .60 for the total group. Also in the study, this positive relationship between grades and teacher evaluations was true for students with low scores on the Blass Objectivity-Subjectivity Scale (classified as "subjective"), but was not true for students with high scores (classified as "objective"). The largest correlation reported between examination scores and any single evaluation item for the "subjective" students was .73. The largest correlation reported between examination scores and any single evaluation item for "objective" students was .44. In the study students were asked to indicate their mark on the mid term examination they had taken two weeks previously.

Endo and Della-Piana (1976) found no significant correlations between student ratings and common final examinations for eight combined sections (n = 111) of trigonometry at the University of Utah. Apparently there were five instructors for the eight sections. No description of the rank or experience of the instructors was given. Correlations between student ratings and achievement were also calculated for each instructor, but there were no consistent trends across the instructors. The highest correlation between any item and achievement for any instructor was .76. Over one-half the initial enrollment was not included in the results because students either did not turn in course evaluation cards or withdrew from class. The researchers admitted that this fact is a "serious sample attenuation which somewhat limits generalizability of results" (p. 84). The evaluation form used in the study consisted of seven items to be rated on a seven-point scale. The authors stated that the validity and the reliability of the form are questionable; no reliability or validity data were reported.

In summary, the studies using achievement scores not adjusted for ability compared to student ratings showed more variation than studies using other types of comparisons. However, in general they tended to support the use of student ratings of professors. Student ratings of TAs were highly varied across the various studies, perhaps too varied to merit their use.

Studies Incorporating Achievement Scores and Sections Selected Without Identity of the Instructor

In addition to the Frey (1976) study mentioned earlier, three other studies have been reported in which students selected their sections without knowing who their instructors were to be. Although this procedure is defensible, it is not as rigorous in research design as random assignment would have been. Certainly the procedure is better than simply ignoring the fact that differences between the sections might exist at the onset of a study. In two of the studies, the researchers stated that a pretest indicated no statistical differences in the initial ability of the students in the sections. However, it is possible that the sections were different in other critical areas than those evaluated by the pretest.

The Marsh, Fleiner, and Thomas (1975) study involved 18 sections of an introductory course in computer programming at the University of California at Los Angeles. Students chose sections on the basis of the time the sections met without any knowledge of who would teach each section. A 46-item evaluation-of-instruction instrument developed at the University of California was used. The section averages of 12 of the 46 items were significantly correlated with the average of the student examination scores for the sections. A multiple correlation of .74, using four of the 12 significant items as predictors of average section achievement, was also significant. In addition, two factors of the instrument, "course organization" and "class presentations," as well as two summary items correlated significantly with achievement. These correlations were .55, .43, .44, and .42, respectively.

In the Marsh, Fleiner, and Thomas study only 72 percent of the students completed the evaluation forms. Also, students in the study were asked to include their registration numbers on the evaluation forms. Even though the students were assured their evaluations would be anonymous, it is possible that the results would have been more valid if students had not been asked to include their registration numbers. A random spot check indicated no variations in the scoring of the objective final examination. The sections of classes in the study were generally taught by graduate students who used a common course outline developed by the director of the course. A major strength of the study is that the instructors had been randomly assigned to the sections.

A second study in which students selected their sections without knowing who the instructor was to be was the Marsh and Overall (1980) study. The subjects, again, were students enrolled in 31 sections of a course in computer programming application at the University of California at Los

Angeles. Instructors were, again, mostly graduate teaching assistants who were supervised by a course director who had developed the final examination. There were no statistical differences on pretest measures of ability and interest in the 31 sections. Results of the study were based on the 73 percent of the enrolled students who completed the required examinations and forms. As in the previous study, students were asked to supply their registration numbers. The evaluation form consisted of 33 items intended to measure seven factors of teaching.

Partial correlations were calculated between ratings given by students at mid-term and at the end of the term and criteria of effective teaching. Ratings given at the end of the term correlated higher with the criteria than did the ratings given at mid-term. Regarding the end of term evaluations, the final examination correlated highest with an "instructional improvement" item (.42), "overall instructor" item (.38), and the factor labeled "instructor enthusiasm/concern" (.40). When the results of this study were compared with the Marsh, Fleiner, and Thomas (1975) study, Marsh and Overall (1980) stated:

> *Both studies reported that achievement was significantly related to overall instructor and instructional improvement summary ratings but was not significantly correlated with overall course ratings. The two studies did not, however, agree on which specific components of the student ratings were most highly correlated with final examination performance. In particular, the Organization factor that was most highly correlated with final examination performance in the earlier study was not significantly correlated with any of the criteria in this study. (p. 474)*

The inconsistent results of the two studies is especially interesting since the samples, courses, examinations, and the procedures were the same or similar.

In a third study, one conducted by Braskamp, Caulley, and Costin (1979), instructors during two subsequent semesters were assigned to sections after students had registered. There is no indication that the researchers checked for, nor controlled for, any possible initial differences in the sections. Instructors in the study were teaching assistants of a psychology course at a "large midwestern university." None of the three global items or the five scales of a student rating form significantly correlated with student performance on a final examination for the fall semester group. For the spring semester group, only one of the scales, labeled "teacher control," was significantly correlated with achievement (.58).

In the study, 80 and 79 percent of students completed the evaluation form for the two semesters. The researchers reported Kuder-Richardson (KR 21) reliabilities of .83 and .86 for the multiple-choice final examination. The researchers reported that 23 instructors taught 47 sections one semester, 19 instructors taught 38 sections the other semester, and 17 of these instructors taught the course both semesters. Means in the study were calcualated by averaging the students' scores in all sections taught

by each instructor. However, in one table the means reported were based on 19 and 17 instructors for the two semesters. The researchers did not explain how these 19 and 17 were chosen, but apparently data for four instructors one semester and two instructors the other semester were not included in the analyses.

Generally speaking, the studies in which initial differences between sections is somewhat controlled for by having students select their sections without knowing who the instructor is to be have not shown a very consistent relationship between student ratings of their instructors and student achievement. It should be noted, however, that these instructors were TAs rather than full-time, experienced professors.

A Meta-Analysis of Student Ratings and Student Achievement

One of the most recent as well as most important studies concerning student ratings and student achievement was a meta-analysis by Cohen (1981). Meta-analysis has been defined as an "analysis of analyses" or "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass 1976, p. 3). Cohen integrated and reanalyzed primary data analyses from some 41 independent validity studies that had incorporated 68 multisections of course ratings in the prediction of student achievement.

The average correlation reported in the studies between student achievement and an overall course rating (available in 22 of the 68 multisection courses) was .47, and the average correlation between student achievement and an overall instructor rating (available in 67 of the 68 courses) was .43. Cohen reported that if no relationship existed between student achievement and overall course ratings or between student achievement and an overall instructor rating, then an equal number of positive and negative correlations would be expected, with the majority of the correlations around zero. However, the majority of the courses reported positive relationships. "Instructors whose students achieved the most were also the ones who tended to receive the highest instructor ratings" (Cohen 1981, p. 296).

Cohen also reported the average correlations between achievement and seven specific teaching dimensions. None of the 41 studies had reported all these seven correlations. The average correlations between achievement and the teaching dimensions were: skill (.50), structure (.47), feedback (.31), rapport (.31), evaluation (.23), interaction (.22), and course difficulty (-.02). The average correlation for student progress, students' self ratings of their learning, and achievement (reported in 11 of the studies) was .47. Cohen concluded:

> *While large effect sizes are found for the Skill and Structure dimensions, other dimensions such as Rapport, Interaction, Feedback, and Evaluation show more modest effects. The Course Difficulty dimension shows no relationship with student achievement. Finally, students' self-ratings of their learning correlate quite highly with student achievement. (p. 298)*

Cohen says that his meta-analysis provides strong support for student evaluation-of-instruction instruments as a measure of teaching effectiveness when the effectiveness is defined as achievement in the course. The data also seem to indicate that in using a student evaluation-of-instruction instrument the greatest emphasis of teaching effectiveness should be placed on an overall course rating item, an overall instructor rating item, or on factors that measure skill, structure, or student progress. Emphasis should not be placed on rating factors that relate to course difficulty.

Relationship of Student Evaluation and Other Criterion Measures

All the previously mentioned researchers have used scores on course examinations in establishing criterion validity. Other researchers have employed other criterion measures in conjunction with achievement. These have included students' gains in the course (Morsh, Burgess, and Smith 1956), students' scores on a national examination (Gessner 1973), scores on a problem-solving exercise (Wiviott and Pollard 1974), and students' interest in advanced courses and attitude toward the course subject (McKeachie, Lin, and Mendelson 1978).

Other researchers have used criterion measures that did not include student achievement. Among these are students' interest in advanced courses (McKeachie and Solomon 1958), judges' ratings of video tape clips of instructors (Stallings and Spencer cited in Aleamoni and Spencer 1973), and the use of ratings of trained observers (McKeachie and Lin 1978).

Studies Using Criterion Measures in Conjunction with Achievement

One criterion of teaching effectiveness could be gains that students make in a course. Morsh, Burgess, and Smith (1956) correlated student gains on a test of knowledge with instructor ratings. They also used gains on a performance examination. The gains made on the written examination, the gains made on the performance test, and the combined gains correlated significantly with the overall ratings of the instructors (.32, .39, and .40, respectively). When only student ratings of the instructors' teaching ability were correlated with the three gains criteria, the correlations were slightly higher (.41, .41, and .46, respectively).

In the study, complete data were available on 106 of 121 instructors of a hydraulics phase of an aircraft mechanics course at Sheppard Air Force Base. Classes consisted of about 14 students each, and this phase of the course lasted only eight days. One possible confounding variable in most of the reported studies is that the instructors who were being rated administered the criterion tests. The way the students felt about this examiner conceivably could have affected their performance on the criterion test. In contrast, a strength of the Morsh, Burgess, and Smith (1956) study is that the criterion tests were administered by personnel other than the instructors of the classes.

Other variables that were compared with gain scores were peer rankings and supervisor ratings and rankings, verbal facility ratings, instructors' knowledge of hydraulics, instructors' general intelligence. Morsh, Burgess, and Smith (1956, p. 86) concluded that "student ratings of their instructors were the only instructor measures which seemed to predict the student gains criterion." Although the study involved instructors at an Air Force base, the researchers suggest that the results "would find application to other teaching situations" (p. 87).

"A confounding factor that has not been sufficiently recognized is that in many instances the persons who developed the measure of achievement and the persons who were rated by the students were presumably the same individuals" (Costin 1978, p. 86). Gessner (1973) recognized this problem and used not only departmental examinations but also a nation-

ally normed examination that the instructors in the study had no part in developing.

Students in the Gessner study were 119 second-year medical students in a basic general science course. Ten faculty members taught the 23 subject areas in the course. Students attending the last lecture of the course evaluated the subject areas with regard to "content and organization" and "presentation." A three-point scale was used: good, fair, or poor, and ratings were assigned values of +1, 0, and −1. A weighted mean rating for each subject area was calculated. A departmental committee prepared departmental examinations from questions submitted by the individual faculty members. On this examination, performance for an area was determined as the mean class performance for that area. Five weeks after the course was completed, 116 students also took Part I of the National Medical Board Examination. Questions from this examination were classified into the subject areas by two members of the faculty. The difference between the percentage of the class and the nationwide sample who answered each question correctly was calculated for each item. These units of differences were averaged for each subject area and were used as the measure of class performance in the subject areas of the national examination. It is not clear why Gessner chose to use these units of difference. The use of the class and the national group in the study has been criticized, and one writer states that the design of the Gessner study lacked internal validity (Leventhal 1975).

The significant correlation between class performance in the subject areas on the national examination and ratings on "content and organization" was .77, and for the subject areas of the national examination and "presentation" the correlation was .69. However, when partial correlation coefficients were calculated for these variables, with "relative emphasis" (the amount of time devoted to a topic) held constant, the correlations dropped slightly to .74 and .62. The correlations between class performance in the subject areas on the departmental examination and the two rating dimensions were only .11 and .17, respectively. Gessner concluded that:

> *It appears quite clear that student ratings of instruction and class performance on national examinations are positively related. the higher the student ratings of the instruction they receive, the higher the class score relative to a nationwide norm. On the other hand, no significant correlation is found between student ratings and class performance on institutional examinations. This suggests that both student ratings and class performance on national normative examinations are valid measures of teaching effectiveness. (p. 569)*

A readily apparent problem in this study is the loss of 41 (of the 119) students who did not attend the last lecture and, therefore, did not rate the subject areas on the two dimensions. The results may well have been different if the ratings of these 41 students had been included.

Too, one could question the two dimensions that were chosen for the ratings. The author did not describe the rationale used for choosing those particular dimensions. Although the students rated 23 subject areas, only 20 of these were used in the computation of correlations. The other three areas are not accounted for. It is possible that these dimensions of the course were not included in the National Medical Board Examination. Certainly, since the two achievement measures were calculated in different ways, there is some question in making comparisons of the correlations involving them. Overall the study adds credibility to the use of student ratings.

In addition to using an achievement test as a criterion, Wiviott and Pollard (1974) also used a problem-solving exercise to measure "ability to analyze, synthesize, and evaluate course content" (p. 37). Results of the research suggested that student ratings of instruction were not related to scores on the achievement test and only contributed "slightly" to a regression model for problem solving.

Again, as in previous studies, one of the criticisms of the Wiviott and Pollard study is the small number of course sections. The sample consisted of six introductory educational psychology sections at the University of Wisconsin-Milwaukee. Whereas the researchers reported the sample was composed of 138 undergraduates in the sections who had completed the criterion measures and the course evaluation forms, one does not know the percentage of students enrolled in the course who were eliminated from the study.

Subjects in the study were assured that their scores on the criterion measures would not affect their grade in the course and that their scores would not be given to their instructors. Since anonymity was not provided for, it is possible the scores were not true measures of the students' ratings of their instructors.

It is not clear from the report whether the instructors of the sections were teaching assistants. The researchers stated, however, that teaching assistants had administered the student evaluation instrument. The researchers administered the tasks. A model answer was used to score the problem-solving exercise, and "inter-rate" reliability was reported to be .78.

In addition to scores on an achievement examination, McKeachie, Lin, and Mendelson (1978) also used interest in advanced psychology courses and attitude toward psychology as criterion measures. Although most researchers have measured achievement by a final examination given at the end of the grading period, McKeachie, Lin, and Mendelson also looked at delayed measures. They state:

> *Probably the oldest objection to student ratings is the comment, "I did not really appreciate some of my best teachers until sometime after the course had ended." Another common quote is, "Most of what a student learns and puts on a final examination is forgotten by the next week." (p. 352)*

McKeachie, Lin, Mendelson, (1978) therefore, compared students' interest in advanced psychology courses, scores on the Introductory Psychology Criteria Test, and scores on the Attitude Toward Psychology Scale to student ratings of instruction at the end of an introductory psychology course and again 14 months after completion of the course. The ratings of instruction were not highly related to the criterion measures at either of the two time periods. The measure of attitude toward psychology was the only follow-up criterion that had a rank order correlation with student ratings that the authors labeled as "substantial" (.66).

The sample consisted of only six instructors at the University of Michigan who were all advanced graduate students. The student rating instrument consisted of items derived from the form reported by Isaacson et al. (1964). In the study, the researchers were able to locate 124 of the original 152 students enrolled in the six courses. Students were sent letters with the questionnaire and were offered three dollars to complete the questionnaire. The students located, 92 (74%) responded (61% of the original sample). The researchers reported that the respondents did not differ significantly from the nonrespondents on the measure that had been completed at the end of the course. On the questionnaire, only ten items of 48 that were used at the end of the course were included from the Introductory Psychology Criteria Test. Other than the positive relationship of attitude toward psychology and student ratings, the study does not lend much support to student ratings, it should be noted that the instructors were TAs rather than full-time professors.

Studies Using Criterion Measures Other Than Achievement

Final examinations are the most commonly used criterion measures of teacher effectiveness. Many instructors argue that some of their most important objectives cannot be measured by final examination scores. The ability to arouse interest in the subject matter should be one of the criteria of effective teaching. It has been stated:

> *While awakened interest is not an educational outcome, we might expect that when a teacher has aroused interest in his field, his students will be likely to elect another course in that field. Thus in comparing the effectiveness of instructors in a multisection course, we might compare the percentages of their students who elected advanced courses. (McKeachie and Solomon 1958, p. 379)*

McKeachie and Solomon then proceeded to validate instructor ratings against the percentage of students who took advanced courses. Data were collected over a period of about three years from students in about eight advanced psychology courses. Students were asked to report the instructor and semester they had taken the beginning psychology course. At the end of some semesters students in the beginning psychology course responded to two items that had to do with an overall rating of the instructor's effectiveness and an overall rating of the course. Instructors were ranked

on the ratings on the two questions and on the percentage of students taking advanced courses. In two of the five semesters, the ratings of the instructors were significantly correlated with the percentage of students electing to take advanced courses (.63 and .41).

Stallings and Spencer (cited in Aleamoni and Spencer 1973), employed a different type of criterion measure. They compared student ratings of nine instructors of accountancy at the University of Illinois with ten judges' ratings of the instructors. The judges, who were measurement specialists and teaching assistants from the speech department, rated video tape clips of the instructors on a three-point scale. Students rated the instructors on the Illinois Course Evaluation Questionnaire (CEQ). Total scores on the CEQ for each instructor were averaged and ranked, and the average instructor ratings by the judges were also ranked. A significant Spearman rank order correlation of .70 was obtained between ranks on the CEQ and the average rating ranks.

Finally, McKeachie and Lin (1978) compared student ratings with ratings given by trained observers. McKeachie and Lin used graduate student observers trained in a categorization system to evaluate 20 teachers in three introductory psychology courses at the University of Michigan. The researchers report that such data are difficult to obtain because of the costs in training observers. Since a factor of many student evaluation of instruction instruments is "rapport," student ratings on that factor of the Student Perceptions of Teaching and Learning were correlated with observed teacher acts relating to "warmth" and "agreement." The "rapport" factor consisted of three items that related to the instructor being permissive, friendly, and inviting criticism. One correlation between the students' ratings of the instructor being friendly and the observed behavior of "agreement" was significant (.61). McKeachie and Lin state that the "study lends some empirical support to the presumption that student ratings of teaching are based on teacher behavior" (p. 47).

Teachers in the study ranged in experience from zero to 27 years. The graduate students observed each class approximately six times during the term. Since the three introductory psychology courses were not described in the study, it is perhaps possible that the nature of the different courses could cause the same teacher to have quite different ratings on "rapport," "warmth," or "agreement" for the different courses.

Taken together, comparisons of teacher ratings to criterion measures other than achievement lend some criterion validity to the use of the ratings. However, as in the use of achievement as a criterion, these comparisons with other measures indicate there is a great deal more to evaluating instruction than can be accounted for in this type of comparison.

Conclusions and Implications

An overall examination of criterion validity studies of student evaluations of college instruction suggests four major observations. The first of these observations is that the majority of the investigations cited in this monograph reported significant positive correlations between student ratings of instruction and criterion measures held to be measures of effective teaching. Therefore, there appears to be sufficient criterion validity data to support the use of student evaluations of college instruction. This synthesis of the findings of many studies indicates that the student evaluations of instruction are tapping into an important dimension of teaching. Thus, student evaluations can be a defensible part of an instructor's evaluation and can contribute to the improvement of teaching.

The second major observation that can be drawn from these studies is that the relationship between student evaluations of instruction and criterion measures is by no means perfect. Although the majority of the investigations cited in this monograph reported a significant positive correlation between ratings and criterion measures of effective teaching, that correlation was almost always a modest one. Several researchers reported no significant correlations. Apparently a great deal more goes into effective teaching than can be easily evaluated with student evaluation-of-instruction instruments, therefore, these evaluations should not be the sole vehicle for judging the effectiveness of instruction.

Student evaluation instruments have been increasingly widely used in making decisions about tenure, promotion, and merit pay of college instructors. Obviously, considering the modest correlations and the methodological problems in the literature cited in this monograph, these ratings should not be the sole criterion. Aleamoni (1976) states that it would be invalid to use student ratings as the only basis for decisions about an instructor's effectiveness. He further adds:

> *it is important that instructional evaluation systems designed for administrative personnel decisions include evaluations of colleagues, course content, course materials, course objectives, instructor self-ratings, quality of student learning, and so forth, in addition to student ratings. (p. 609)*

Even when student ratings are not used for personnel decisions and are used only for the improvement of instruction, the instructor should realize that the research indicates that student evaluations do not tell the whole story. Additional sources of feedback would appear to be needed.

A third general observation is that there is a discernable trend in the frequency with which certain student rating variables appear as leading indicators of effective teaching. First, items relating to an overall rating of the instructor or overall scores on the instrument were often listed as significant predictors of teaching effectiveness. Cohen (1981) also pointed out in his meta-analytic study that an overall course or an overall instructor item correlated highly with student achievement. Such an overall item or an overall evaluation score is of some use in decisions about tenure, promotion, and merit pay, but it would not be very useful in the improve-

ment of instruction, one of the primary reasons for student evaluation-of-instruction instruments. Teachers need much more information relating to specific strengths and weaknesses if they are to make any adjustments in their teaching.

Benton (1979) examined 19 studies that reported the names of factors of student evaluation of instruction instruments. He found that the 113 named factors of the various studies could be classified into eight categories. These categories, in order of the frequency of appearance, were: skill of instructor, student-teacher interaction, course organization and content, feedback to students, course difficulty and workload, motivation, importance of the course, and attitude of instructor. The overall examination of the studies reviewed in the present monograph indicates that the first three categories listed by Benton were often listed as *significant* predictors of some measure of teaching effectiveness.

The category most often mentioned in the studies as a significant predictor of teaching effectiveness related to the skill of the instructor. This category appeared more than two times as often as the second best predictor. Factors labeled "skill," "lectures," "presentations," "presentation clarity," "presentation skill," "expository skills," and "class presentations" were included in this skill-of-the-instructor category.

The second category most often listed as correlating significantly with achievement was organization and content. Factors labeled "organization-planning," "planning," and "course organization" were included in this category.

The third category most often found to significantly correlate with instructor effectiveness was interaction and included factors labeled "interaction" and "student-faculty interaction." Although factors relating to the other five categories reported by Benton were sometimes reported as significant predictors of effective teaching, the infrequency of their appearance gives them much less credibility than the three mentioned above.

It is interesting to note that in the meta-analysis reported by Cohen (1981) skill and structure correlated "highly" with achievement, whereas interaction correlated "moderately" with achievement. The present study and the Cohen meta analysis seem to indicate that factors relating to skill of the instructor and to organization and planning or structure are factors that correlate highest with teacher effectiveness. Therefore, in selection of an instrument designed to measure teacher effectiveness, it is recommended that an instrument should definitely possess these two factors, and that they should carry more weight in faculty tenure, promotion, and pay decisions than other factors of the instruments. Further, instructors seeking to improve their teaching should attend most carefully to these factors.

The fourth major observation drawn from the present analysis is that a definite need still exists for more criterion validity studies of student evaluation of instruction instruments. Future researchers in this area need to give more attention to the methodological problems previously cited. Specifically, of the recommendations listed by Benton (1974), the following

appear to have application for future studies investigating the relationship of ratings and criterion measures:

1. Subjects should be randomly assigned to sections, and then instructors also should be randomly assigned.
2. A large number of sections should be used.
3. Subject-matter content should be essentially the same across the sections.
4. Examinations with better psychometric qualities should be used (such as rationale for devising and revising items and validity and reliability information).
5. In addition to achievement, other appropriate criterion measures that cover the spectrum of instructor objectives should be used (e.g., attitude measures).

Also more standardized procedures for administration and scoring of instruments should be evidenced in future reports.

It would appear to be advisable to replicate studies that have already been reported. If student rating forms with adequate reliability and validity information are used, one would be justified in determining whether the relationships reported in the reviewed studies could be further generalized. In addition, further research that is most urgently needed is: (1) the comparison of ratings of TAs and full-time professors, (2) the effects of the time during the semester or quarter the forms are administered on the ratings of the instructors, (3) criterion validity studies that involve advanced classes, (4) criterion validity studies that involve graduate classes, and (5) the comparison of rating forms developed empirically and those developed rationally.

The table on pages 36–40 is a summary of the studies reported in this monograph that examined the relationship of student ratings of instruction and criterion measures. Although parallel data were not reported in all the studies, the table shows the largest significant correlation reported in each study.

These largest correlations are squared to indicate the proportion of variance (common variance) shared by the two variables the criterion and the student ratings. Common variance has to do with the variation in one variable that can be attributed to its tendency to vary with the other. For example, if an obtained correlation of .50 is squared, the resulting value is .25. This indicates that we know 25 percent of what we need to know to make a perfect prediction of one variable (the criterion) from the other (the result of the student rating).

The range of the significant correlations (−.75 to .96) indicates that the findings are highly inconsistent. Further, when there was a significant relationship between the ratings and the criterion, the amount of variance accounted for was usually not large. Examination of the table suggests several possible reasons for the inconsistency of the findings. One of the most obvious possibilities is that many of the studies were based on small

Studies Examining Relationship of Student Ratings of Instruction and Criterion Measures

Study	*Sample*	*Student Evaluation Instrument*	*Largest Significant Correlation Reported*	*Largest Significant Correlation Squared*
Bendig (1953)	5 sections of intro. psychology	Purdue Scale for Instruction	.46	.21
Benton & Scott (1976)	31 sections of freshman English	Student Instructional Report, Inventory of Student Perceptions of Instruction	.62	.39
Blass (1974)	1 intro. psychology course	Course Rating Sheet	.73	.53
Braskamp, Caulley, & Costin (1979)	19 and 17 instructors of psychology	3 global items, items from Costin (1971), items from form described by Isaacson et al. (1964)	.58	.34
Bryson (1974)	20 sections of college algebra, 14 instructors	12 items from a "routinely administered faculty and course evaluation form"	.68	.46
Canaday, Mendelson, & Hardin (1978)	one-section anatomy course	total scores on a 31-item questionnaire (further described)	.42	.18

Centra (1977)	72 sections of 7 different courses taught by 74 teachers (no TAs)	Student Instructional Report	.96	.92
Cohen & Berger (1970)	25 sections of a basic natural science course	Michigan State University Student Instructional Rating Report	.48	.23
Costin (1978)	96 TAs of intro. psychology over a 4-year period	5 items from a form described by Isaacson et al. (1964)	.56	.31
Doyle & Crichton (1978)	12 instructors of communications	4 items from factors identified by Doyle & Whitely (1974) + 2 summary items	——	——
Doyle & Whitely (1974)	12 TAs teaching beginning French	Student Opinion Survey	.51	.26
Endo & Della-Piana (1976)	8 sections of trigonometry, 5 instructors	Associated Students of the University of Utah Course Evaluation	.76	.58
Frey (1973b)	8 instructors of intro. calculus & 5 instructors of multidimensional calculus	A Northwestern University instructional rating form	.91	.83

Frey (1976)	7 sections of intro. calculus	not specified	.90	.81
Frey, Leonard, & Beatty (1975)	12 and 5 sections of intro. calculus, 9 sections of ed. psychology	Endeavor Instructional Rating Form	.85	.72
Gessner (1973)	10 faculty members teaching 23 subject areas of a basic science course	ratings of each of the subject areas regarding content, organization, and presentation	.77	.59
Hsu & White (1978)	12 classes of undergraduate education courses, instructors	Inventory of Student Perceptions of Instruction, Instructional Improvement Questionnaire	.74	.55
Marsh, Fleiner, & Thomas (1975)	18 sections of intro. computer programming	46-item instrument developed at the Univ. of California, Los Angeles	.74	.55
Marsh & Overall (1980)	31 sections of intro. to computer programming applications	33 items from 7 factors and 3 summary items	.42	.18
McKeachie & Lin (1978)	20 instructors of 3 intro. psychology courses	"rapport" factor of Student Perceptions of Teaching and Learning	.61	.37

McKeachie, Lin, & Mann (1971)	5 studies: (1) a. 33 sections of general psychology b. 34 sections of general psychology (2) 32 sections of general psychology (3) 6 instructors, no. of sections not specified (4) 16 sections of 2nd yr. French (5) Intro. economics taught by 18 TAs	variations of the Isaacson et al. (1964) scale	(1) a. .42 b. .40 (2) —— (3) .65 (4) —— (5) .72	(1) a. .18 b. .16 (2) —— (3) .42 (4) —— (5) .52
McKeachie, Lin, & Mendelson (1978)	6 sections of intro. psychology, 6 TAs	items derived from form described by Isaacson et al. (1964)	.66	.44
McKeachie & Solomon (1958)	approximately 8 advanced undergrad. psychology courses over a 3-year period	2 global items	.63	.40
Morsh, Burgess, & Smith (1956)	106 instructors of a hydraulics phase of an aircraft mechanics course	global item + 4 ratings on qualities of the instructors	.46	.21
Orpen (1980)	10 sections of mathematics, 10 TAs	Teaching Rating Form, a version of the form of McKeachie et al. (1971)	.75	.56

Rodin & Rodin (1972)	"teaching assistants" of 12 sections of undergraduate calculus	one global item	−.75	.56
Stallings & Spencer (cited in Aleamoni & Spencer 1973)	9 instructors of beginning accounting	Illinois Course Evaluation Questionnaire	.70	.49
Sullivan & Skanes (1974)	130 sections of 10 different courses	researcher-designed form	.53	.28
Turner & Thompson (1974)	16 and 24 sections of beginning French taught by TAs	30 items from scale of Deshpande, Webb, & Marks (1970) + 5 additional items	−.52	.27
Whitely & Doyle (1979)	5 professors, 11 TAs of "beginning mathematics"	Student Opinion Survey	.80	.64
Wiviott & Pollard (1974)	6 sections of intro. ed. psychology	25-item rating scale and the grade A–F assigned to the course	——	——

sample sizes. Very few of the studies had large enough sample sizes to merit confidence in the stability on the results. Approximately one-third of the studies were based on samples of less than ten sections.

A second possibility results from the diversity of the types of courses that used the evaluation forms. Among the subject areas reported in the studies were psychology, English, mathematics, science, communications, French, government, and computer programming. Most of the courses were beginning courses; a few were advanced. It is entirely possible that one subject area may require a different type of teaching than another, and that the type of teaching in advanced courses is very different from the teaching in beginning courses. Perhaps in the advanced courses the number of students in various sections would be smaller than the number of students in beginning courses, and this difference could affect the ratings. None of the studies concerned graduate classes, yet many colleges and universities are presently using evaluation forms in graduate classes.

Another possibility for the diversity involves the number of types of evaluation forms used in the studies. Indeed, it was rare to find two studies that used the same form. Many of the researchers were so vague in describing the instruments they used that it would be impossible to replicate their studies. Other researchers used forms that lacked rationale for devising items, lacked provisions for revising items, and that had no reliability and validity information. Some researchers used a portion of items from other instruments but offered no reliability or validity for those items. Marsh and Overall (1980) suggest that even if different evaluation instruments that had similar factor labels were used, there is no guarantee that the factors are indeed the same. Rating forms have been developed in several different ways. Benton (1979) reports there are two approaches to developing items to be included in the final form of instruments. a rational approach and an empirical approach. The review of the literature does not indicate which of the two types of instruments would be the better predictor of criterion measures. In many cases one does not know, when reviewing the research, whether an instrument was developed by one of the two approaches or whether the items on the instrument simply have face validity.

Another possiblity is that many of the studies did not distinguish between who was being evaluated, TAs or full-time professors. Although apparently the majority of these studies used TAs, the findings have been overgeneralized to represent college and university teachers in general. It is not easy to set up such studies involving full-time professors. One suspects it is exceedingly difficult to get a large number of professors to use a common examination, textbook, and syllabus. Unfortunately many full-time professors' salaries, tenure status, and promotions are being determined, at least in part, by these instruments that have too little empirical research with full-time professors.

There is some evidence that the evaluation of TAs and full-time professors is significantly different with such instruments, and that there is greater criterion validity support for the use of student ratings for full-

time, experienced professors than for their use with TAs. Until there is evidence to support the practice, it is recommended that full-time professors not be evaluated using instruments on which the only reliability and validity data have to do with TAs and vice versa.

There are other possibilities not reflected in the table that could have contributed to the variation in the reported findings. Some of the studies present little evidence that student evaluation instruments were administered under standardized conditions. It is, perhaps, common knowledge that a lack of anonymity affects ratings and that if an instructor remains in the classroom, the ratings will be different than if the instructor leaves. Only a few researchers reported whether the latter was a part of the procedures. Any number of other variations in the administration of the rating instruments could have contributed to differences in results.

Another possible source of the diversity concerns the criterion measures used in the studies. In some studies psychometric properties of the instruments are not known. In other instances no information was reported about the scoring of these criterion measures.

Another possible reason for the inconsistent findings is that many of the studies have not provided adequate control for initial differences in the sections of the courses. Some researchers adjusted for initial student ability but were not consistent in the measures used to adjust for ability. Other researchers used no control for initial ability. It has been previously mentioned that the sections could be different in other areas, such as motivation, which could affect evaluation of instruction. Some researchers used samples in which the students selected courses without knowing who the instructor was to be, a procedure that is less adequate than randomization. In only two studies were students randomly assigned to sections.

Another factor that could have caused results to differ was the time the evaluations were administered. There is no ideal time to do so. There is evidence that evaluations administered as early as mid-term will have different results from those administered at the end of the course. When administered before the final examination the students have not experienced an important part of the course that should be a part of the instructor evaluation. Also, when the ratings are administered at the time of the final examination, test anxiety might contaminate the instruction evaluation.

Frey (1973b) indicates that when student ratings are made after the grades are known the course evaluation might simply reflect the students' acceptance of their instructors' evaluations of them. Frey also mentions a "retaliation hypothesis," i.e., the students may tend to mark lower an instructor who has given them a low grade. Rodin, Frey, and Gessner (1975) mention the "reward hypothesis," i.e., the students may tend to mark higher an instructor who has given them a high grade.

Although it is rarely mentioned in any of the research and could not be accounted for in the summary table, one research design problem further clouds the issue. Most research projects of this nature depend upon instructors who will volunteer to be evaluated. It may be that fewer of the poor instructors volunteer, thus, there is not as much variability in

the teacher rating scores in the samples as might be representative of the total population of college instructors. In other words, there might not be a truly representative range of teaching abilities in the various subjects of many of these studies. Perhaps if the range of instructors were increased, a greater relationship between student ratings of instruction and measures of teacher effectiveness would exist.

However, in defense of these studies, one should note that most of the actual use of student evaluation instruments depends on the faculty members volunteering to use them. Most colleges and universities simply say that some kind of defensible evidence must be produced to support a teacher's candidacy for promotion, tenure, or merit pay. Many professors turn to student evaluations for such evidence. Obviously those who know they are going to get poor student evaluations are not going to use them if they can possibly avoid it. Thus, even though the volunteer aspect of the student evaluation-of-instruction studies may be a limitation because it does not accurately represent the total population of college teachers, it probably is a strength because the volunteer aspect may accurately represent the actual present use of such instruments.

As long as a great many unresolved questions remain about academic freedom and evaluation of teaching, it is likely that a certain amount of volunteerism will continue with the use of student ratings of instruction.

Student evaluations of instruction have long been used by individual instructors to help them improve their teaching. In recent years colleges and universities have had to become acutely aware of the possibility of litigation with personnel decisions. This concern with litigation has forced institutions of higher learning to look for evidence to substantiate personnel decisions. Those seeking the improvement of teaching and those seeking a more objective data base for decision making have turned more and more to student evaluations of instruction.

As with all cases of evidence, concern must eventually turn to the quality of that evidence. Many criticisms have been leveled at student evaluations of instruction. Some of this criticism has come from measurement and evaluation experts. Much of it comes, one suspects, from professors and teaching assistants who do not get very good student evaluations. In considering these criticisms of student ratings one must turn to fundamental questions of their legitimacy. No rating procedure should be used to modify teaching methods or in university governance unless that procedure has established validity.

Of first consideration in matters of student ratings is the question of criterion validity, i.e., how well do student ratings hold up when compared to accepted indicators of good and poor teaching.

It seems quite clear that student ratings of instruction provide good evidence of the quality of teaching. However, they provide evidence only, they should not be considered to be more than evidence. They should never be considered alone as positive proof. It is quite clear that there is something more to teaching than can ever be totally accounted for by those who are taught.

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*. Most of these publications are available through the ERIC Document Reproduction Service (EDRS). For publications cited in this bibliography that are available from EDRS, ordering number and price are included. Readers who wish to order a publication should write to the ERIC Document Reproduction Service, 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).

Aleamoni, Lawrence M. "Typical Faculty Concerns about Student Evaluation of Instruction." Paper presented at the Symposium on Methods of Improving University Teaching, March 1974, at Haifa, Israel. ED 113 995. MF-$1.11; PC-$3.49.

———. "On the Invalidity of Student Ratings for Administrative Personnel Decisions." *Journal of Higher Education* 47 (September/October 1976):607–10.

Aleamoni, Lawrence M., and Spencer, Richard E. "The Illinois Course Evaluation Questionnaire: A Description of Its Development and a Report of Some of Its Results." *Educational and Psychological Measurement* 33 (Autumn 1973):669–84.

Bausell, R. Barker, and Magoon, Jon. "Expected Grade in a Course, Grade Point Average, and Student Ratings of the Course and the Instructor." *Educational and Psychological Measurement* 32 (1972):1013–23.

Bendig, A. W. "The Relation of Level of Course Achievement to Students' Instructor and Course Ratings in Introductory Psychology." *Educational and Psychological Measurement* 13 (1953):437–48.

Benton, Sidney E. "A Comparison of Two Types of Student Response Inventories for Appraising Instruction." Ed.D. dissertation, University of Georgia, 1974.

———. "Instruments of Students' Assessment of College Instruction." *Phi Delta Kappa CEDR Quarterly* 12 (Winter 1979):13–15, 22.

Benton, Sidney E., and Scott, Owen. "A Comparison of the Criterion Validity of Two Types of Student Response Inventories for Appraising Instruction." Paper presented at the annual meeting of the National Council on Measurement in Education, April 1976, at San Francisco. ED 128 397. MF-$1.11; PC-$5.14.

Biddle, John C. "Annotated Bibliography of ERIC Reports Concerning Student Evaluation of Faculty Performance, 1974 to 1978." Bakersfield, California State College-Bakersfield: 1980. ED 184 483. MF-$1.11; PC-$6.79.

Blass, Thomas. "Measurement of Objectivity–Subjectivity. Effects of Tolerance for Imbalance and Grades on Evaluations of Teachers." *Psychological Reports* 34 (1974):1199–1213.

Blount, H. Parker, Gupta, Venu G., and Stallings, William M. "The Effects of Different Instructions on Student Ratings of University Courses and Teachers." Paper presented at the annual meeting of the National Council on Measurement in Education, April 1976, at San Francisco.

Braskamp, Larry A.; Caulley, Darrel; and Costin, Frank. "Student Ratings and Instructor Self-Ratings and Their Relationship to Student Achievement." *American Educational Research Journal* 16 (Summer 1979):295–306.

Bryson, Rebecca. "Teacher Evaluations and Student Learning. A Reexamination." *Journal of Educational Research* 68 (September 1974):12–14.

Canaday, Stephen D., Mendelson, Marilyn A.; and Hardin, James H. "The Effect of Timing on the Validity of Student Ratings." *Journal of Medical Education* 53 (December 1978):958–64.

Centra, John A. *The Student Instructional Report. Its Development and Uses. SIR*

Report Number 1. Princeton, N.J.; Educational Testing Service, 1972.
———. "Student Ratings of Instruction and Their Relationship to Student Learning." *American Educational Research Journal* 14 (Winter 1977):17–24.
———. *Determining Faculty Effectiveness*. San Francisco: Jossey-Bass, 1980.
Cohen, Peter A. "Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies." *Review of Educational Research* 51 (Fall 1981):281–309.
Cohen, Stanley, H., and Berger, Wallace G. "Dimensions of Students' Ratings of College Instructors Underlying Subsequent Achievement on Course Examinations." *Proceedings of the 78th Annual Convention of the American Psychological Association* (1970):605–606.
Costin, Frank. "Empirical Test of the 'Teacher-centered' Versus 'Students-centered' Dichotomy." *Journal of Educational Psychology* 62 (1971):410–12.
———. "Do Student Ratings of College Teachers Predict Student Achievement?" *Teaching of Psychology* 5 (April 1978):86–88.
Costin, Frank; Greenough, William T.; and Menges, Robert J. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." *Review of Educational Research* 41 (December 1971):511–35.
Deshpande, Anant S.; Webb, Sam C.; and Marks, Edmond. "Student Perceptions of Engineering Instructor Behaviors and Their Relationship to the Evaluation of Instructors and Courses." *American Educational Research Journal* 7 (May 1970):289–305.
Doyle, Kenneth O., Jr., and Crichton, Leslie I. "Student, Peer, and Self Evaluations of College Instructors." *Journal of Educational Psychology* 70 (October 1978). 815–26.
Doyle, Kenneth O., Jr., and Whitely, Susan E. "Student Ratings as Criteria for Effective Teaching." *American Educational Research Journal* 11 (Summer 1974):259–74.
Eble, Kenneth E. *The Recognition and Evaluation of Teaching*. Washington, D.C.. American Association of University Professors, 1971. ED 046 350. MF-$1.11; PC-$10.48.
Endo, George T., and Della-Piana, Gabriel. "A Validation Study of Course Evaluation Ratings." *Improving College and University Teaching* 24 (Spring 1976):84–86.
Frey, Peter W. "Student Instructional Ratings and Faculty Performance." Paper presented at the annual meeting of the American Educational Research Association, February 1973a at New Orleans.
———. "Ratings of Teaching: Validity of Several Rating Factors." *Science* 182 (October 1973b):83–85.
———. "Validity of Student Instructional Ratings: Does Timing Matter?" *Journal of Higher Education* 47 (May/June 1976):327–36.
———. "A Two-Dimensional Analysis of Student Ratings of Instruction." *Research in Higher Education* 9 (1978):69–91.
Frey, Peter W.; Leonard, Dale W.; and Beatty, William W. "Student Ratings of Instruction: Validation Research." *American Educational Research Journal* 12 (Fall 1975):435–44.
Gessner, Peter K. "Evaluation of Instruction." *Science* 180 (May 1973):566–70.
Glass, Gene V. "Primary, Secondary, and Meta-analysis of Research." *Educational Researcher*, 5 (November 1976):3–8.
Gronlund, Norman E. *Measurement and Evaluation in Teaching*. 4th ed. New York: MacMillan Publishing Co., Inc., 1981.
Hayes, Robert B. "A Way to Measure Classroom Teaching Effectiveness." *Journal*

of Teacher Education 14 (June 1963):168–76.

Hsu, Yi-Ming, and White, William F. "Interactions Between Teaching Performance and Student Achievement." Paper presented at the annual meeting of the American Educational Research Association, March 1978, at Toronto. ED 151 332. MF-$1.11; PC-$3.49.

Isaacson, Robert L., McKeachie, Wilbert J.; Milholland, John E.; Lin, Yi G.; Hofeller, Margaret; Baerwaldt, James W.; and Zinn, Karl L. "Dimensions of Student Evaluations of Teaching." *Journal of Educational Psychology* 55 (1964):344–51.

Leventhal, Les. "Teacher Rating Forms: Critique and Reformulation of Previous Validation Designs." *Canadian Psychological Review* 16 (October 1975):269–76.

Leventhal, Les; Perry, Raymond P., and Abrami, Philip C. "Effect of Lecturer Quality and Student Perception of Lecturer's Experience on Teacher Ratings and Student Achievement." *Journal of Educational Psychology* 69 (August 1977):360–74.

Marsh, Herbert W. "The Validity of Students' Evaluations: Classroom Evaluations of Instructors Independently Nominated as Best and Worst Teacher by Graduating Seniors." *American Educational Research Journal* 14 (Fall 1977):441–47.

Marsh, Herbert W., Fleiner, Howard; and Thomas, Christopher S. "Validity and Usefulness of Student Evaluations of Instructional Quality." *Journal of Educational Psychology* 67 (June 1975):833–39.

Marsh, Herbert W., and Overall, J. U. "Validity of Students' Evaluations of Teaching Effectiveness: Cognitive and Affective Criteria." *Journal of Educational Psychology* 72 (August 1980):468–75.

McKeachie, W. J., and Lin, Y. G. "A Note on Validity of Student Ratings of Teaching." *Educational Research Quarterly* 4 (Fall 1978):45–47.

McKeachie, W. J.; Linn, Yi-Guang; and Mendelson, Cynthia Neigler. "A Small Study Assessing Teacher Effectiveness: Does Learning Last?" *Contemporary Educational Psychology* 3 (October 1978):352–57.

McKeachie, W. J., and Solomon, Daniel. "Student Ratings of Instructors. A Validity Study." *Journal of Educational Research* 51 (January 1958):379–82.

Miller, Richard I. *Evaluating Faculty Performance*. San Francisco. Jossey-Bass, 1974.

Morsh, Joseph E., Burgess, George G.; and Smith, Paul N. "Student Achievement as a Measure of Instructor Effectiveness." *Journal of Educational Psychology* 47 (February 1956):79–88.

Orpen, Christopher. "Student Evaluation of Lecturers as an Indicator of Instructional Quality: A Validity Study." *Journal of Educational Research* 74 (September/October 1980):5–7.

Pettman, Phillip J. "Student Evaluation of Faculty. 1965–1970. An Annotated Bibliography." Minneapolis: Minnesota University Measurement Services Center. 1972. ED 054 735. MF-$1.11; PC-$3.49.

Pohlmann, J. T. *Factor Analyses of Parts I and II of the IIQ. Technical Report 6. 1–72*. Carbondale, Ill.: Testing Center, Southern Illinois University, 1972.

Rippey, Robert M. "Student Evaluations of Professors. Are They of Value?" *Journal of Medical Education* 50 (October 1975):951–58.

Rodin, Miriam; Frey, Peter W., and Gessner, Peter K. "Student Evaluation." *Science* 187 (February 1975):555–59.

Rodin, Miriam, and Rodin, Burton. "Student Evaluations of Teachers." *Science* 177 (1972):1164–66.

Romine, Stephen. "A Decade of Experience with Student Ratings of College Instruction." *Phi Delta Kappan* 54 (February 1973):415–16.

Scott, Craig S. "Some Remarks on 'Student Ratings: Validation.'" *American Educational Research Journal* 12 (Fall 1975):444–47.

Scott, Owen. "The Measurement and Use of Student Perceptions to Improve College Instruction." Paper presented at the annual meeting of the National Council on Measurement in Education, February 1973, at New Orleans.

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

Shultz, Charles B. "Some Limits to the Validity and Usefulness of Student Ratings of Teachers: An Argument for Caution." *Educational Research Quarterly* 3 (Summer 1978):12–27.

Sullivan, Arthur M., and Skanes, Graham R. "Validity of Student Evaluation of Teaching and the Characteristics of Successful Instructors." *Journal of Educational Psychology* 66 (April 1974):584–90.

Thorndike, Robert L., and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education*. 3rd ed. New York: John Wiley and Sons, Inc., 1969.

Turner, Richard L., and Thompson, Robert P. "Relationships Between College Student Ratings of Instructors and Residual Learning." Paper presented at the annual meeting of the American Educational Research Association, April 1974, at Chicago. ED 090 826. MF-$1.11; PC-$3.49.

Voeks, Virginia W. "Publications and Teaching Effectiveness." *Journal of Higher Education* 33 (April 1962):212–18.

Whitely, Susan E., and Doyle, Kenneth O., Jr. "Validity and Generalizability of Student Ratings from Between-Classes and Within-Class Data." *Journal of Educational Psychology* 71 (February 1979):117–24.

Wiviott, Suzanne P., and Pollard, Diane S. "Background, Section, and Student Evaluation Variables as Predictors of Achievement in a College Course." *Journal of Educational Research* 68 (September 1974):36–42.

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 Dupon 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

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