DOCUMENT RESUME

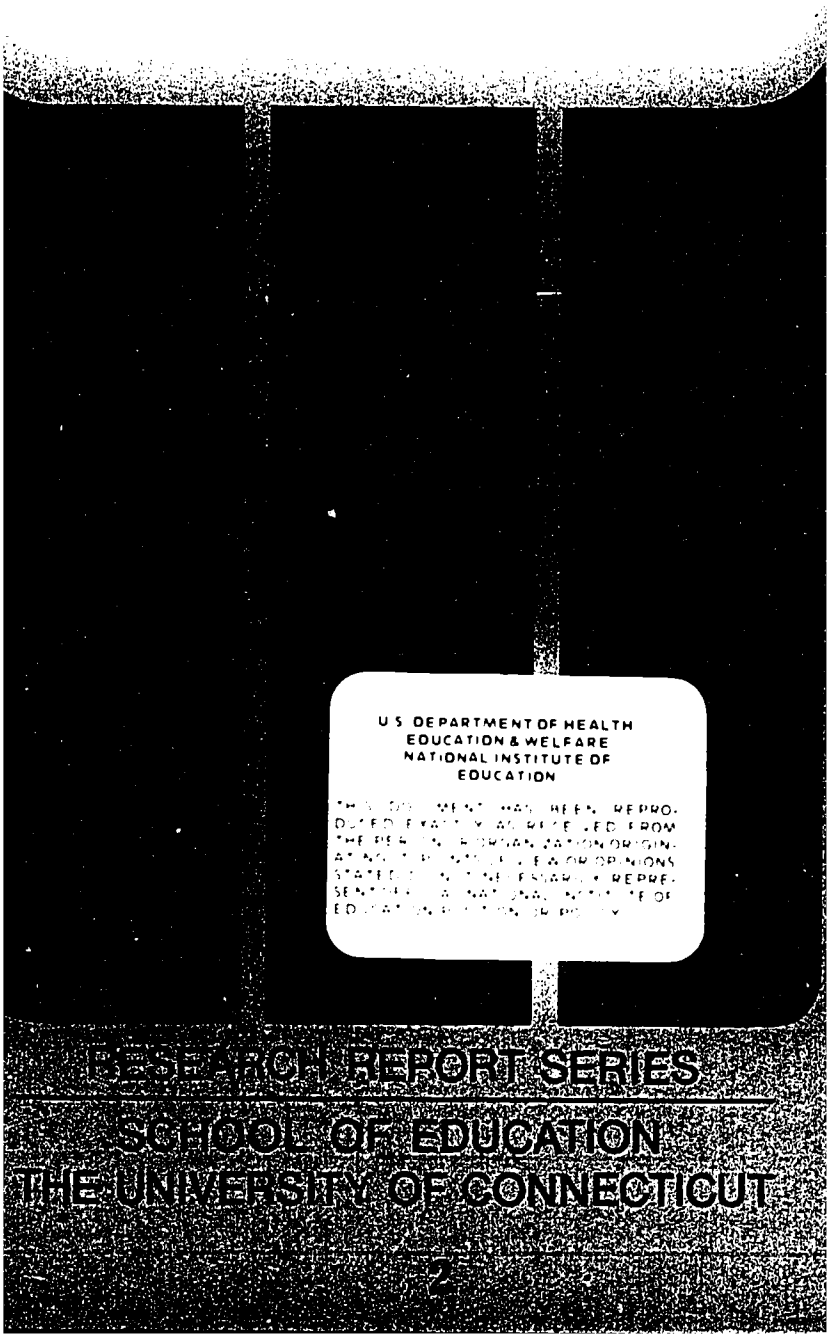
ED 129 902 TM 005 744

AUTHOR Owen, Steven A.
TITLE The Validity of Student Ratings: A Critique.
INSTITUTION Connecticut Univ., Storrs. Bureau of Educational Research and Service.
PUB DATE Apr 76
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 19-23, 1976); For related document, see TM 005 750

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; Class Size; Effective Teaching; Evaluation Criteria; Grades (Scholastic); Higher Education; Political Influences; Rating Scales; Secondary Education; *Student Evaluation of Teacher Performance; *Teacher Rating; Test Reliability; Test Validity; Units of Study (Subject Fields); *Validity
IDENTIFIERS Halo Effect

ABSTRACT

By considering such traditional features as validity and reliability, as well as utility, and political and ethical considerations, this paper attempts to establish that student ratings are not credible as sources of information about teacher effectiveness. Several of the most common problems in the related literature are outlined and include how ratings are related to grades awarded; the relationship between course content, or major areas, and student ratings; agreement between student raters and other raters; the relationship of student learning and teacher rating; the halo effect on teacher rating; the influence of course or class level and class size on instructor ratings; reliability of teacher rating; and the politics of evaluation. If sense is to be made of student rating instruments, the author suggests a threefold approach. First, he proposes a moratorium on student ratings as evaluative measures. Second, existing instruments need to be refined until they satisfy minimal criteria of objectivity, reliability, sensitivity, validity, and utility. Finally, he proposes that student ratings be studied in depth. Their use, for the time being, must be experimental and not evaluative. Suggested research directions are proposed. (RC)


U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
RESEARCH REPORT SERIES
SCHOOL OF EDUCATION
THE UNIVERSITY OF CONNECTICUT
2

The Validity of Student Ratings:

A Critique

Steven V. Owen

University of Connecticut

A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1976.

The Validity of Student Ratings:

A Critique

Steven V. Owen

University of Connecticut

Introduction.

In their latest edition of Learning and Human Abilities, Klausmeier and Goodwin (1975, p. 174) insist that, beyond product, process, and presage criteria, "There are no other generally accepted criteria, or procedures, for evaluating the effectiveness of classroom teachers." Since student evaluations are not classified as product, process, or presage criteria, Klausmeier and Goodwin imply—by omission—that student ratings are not credible as sources of information about teacher effectiveness. This paper will attempt to establish that Klausmeier and Goodwin are correct. To be considered are such traditional features as validity and reliability, as well as worth, and political and ethical considerations.

As Frey (1974) has noted, student ratings are currently enjoying a surge of popularity. The literature abounds with studies on student ratings, and often on the development of new rating scales. Yet we learn little from all of this literature, because new scales, and their often improper administration, rarely resolve the problems of old scales. For some observers, the increased use of student ratings as "measures" of teacher effectiveness has implied an increased acceptance of this form of assessment. A closer look, however, suggests that there is abundant

confusion and occasionally, skepticism about the meaningfulness, validity, and usefulness of student ratings. Some researchers have pointed out that the utility of student ratings depends on the purpose of the ratings. Doyle and Whitely (1974), for example, proposed that the interpretation of ratings depends upon whether they are intended for personnel decisions or for diagnostic purposes. Yet extremely little is known of the circumstances or conditions which permit useful decisions about student ratings (cf., McKeachie, 1973).

Some administrators, instructors, and researchers support strongly the use of student ratings for all purposes. Proponents have claimed, glibly, that those who dislike student ratings are simply reacting on the basis of a generalized fear of being evaluated. However, a rigorous appraisal of research and theory shows enough inconsistency, methodological shortcomings and naive acceptance of student ratings to cause genuine trepidation about their use as evaluative instruments. The purpose of this paper is to outline several of the most common problems in the related literature; these problems, I think, have contributed most to our lack of understanding about student ratings of teacher effectiveness.

Grades Awarded.

A plethora of studies have faced the question about how ratings are related to grades awarded, but the literature is replete with contradiction. A basic problem is that the relationship between grades and ratings may depend on which comes first. (It should be noted that few studies report whether grades or evaluations come first.) It seems reasonable that grades may have a greater influence on ratings if the ratings

are done after grades are awarded. If there are situational events that influence a student's feelings about a course or instructor, an unexpected grade may well change those feelings. Holmes (1972) found a powerful lowering of ratings after students were given a grade which was lower than they expected to get. Holmes' conclusion is that we should keep students "adequately informed of their proficiency, [so that] the possibility of disconfirmed expectancies will be decreased...."(p. 133). This suggestion refers to surprise grades at the end of the semester; but some students get surprises on exams throughout a course, which may well affect their ratings of the instructor. Bausell and Magoon (1972) supported this comment, showing that students whose grade expectancies decreased during a course also lowered their ratings of the course and instructior. Kennedy's (1975) study confirms the relationship between expected grades and student ratings, although the study was limited to a single course (across 15 sections).

Using a multivariate design, Lolli and Owen (1976) and Bausell and Magoon (1972) found significant differencies in ratings between three groups of students: Those whose expected course grade was lower than their GPA, those with congruent expected grade and GPA, and those whose expected grade was higher than their GPA. As hypothesized, the discrepancy between expected grade and GPA is a potent intervening variable in ratings. A superficial solution is to omit from ratings summaries all the "discrepant" students, and examine only the ratings of "non-discrepant" students. We have yet to see evidence, however, that this middle group of students provide more valid ratings than are otherwise obtained.

The potential for student "whimsy" is compounded when, as is sometimes

the case, ratings are collected after the course via mailed questionnaires. A return rate of 50 percent (Brown, 1974) does not guarantee a representative sample of all students in a course. It may be that those motivated to return the rating questionnaires are those with the strongest feelings about the course. While this outcome may ensure ratings at both ends of the like-dislike continuum, it may also underrepresent those with moderate likes and dislikes. The literature does not provide answers about the representativeness of voluntary questionnaire returns of student ratings.

The interpretation of a correlation between ratings and grades awarded is difficult. If students receiving higher grades tend to rate instructors more highly, it is possible that the instructor is, in fact, more effective for them. Another viewpoint is that the effectiveness of an instructor is a "truth," and student variation in ratings represents error variance in interpreting the "truth" (Deshpande et al., 1970). Nevertheless, the biasing influence of grades has not been resolved in the literature.

Subject Matter Content or Major Area of Study.

There are only a few studies on the relationship between course content, or major area, but there appears to be consistency in the findings. In particular, the nature of the content area (Veldman and Peck, 1969), or the student's major area (Slater, 1974; Paulus, 1973; Remmers, 1963; Centra, 1973a,b; Kennedy, 1972) influence student impressions of teachers. Veldman and Peck (1969) appear to dismiss this influence, commenting that the nature of course content explains only a small portion of ratings variance, and that prior self-selection of teachers into content areas

"undoubtedly" relates to differential ratings. Nevertheless, such minor variables continue to erode the trustworthiness of ratings.

Agreement Between Student Raters and Other Raters.

Several studies have examined the level of evaluative agreement between students and other sources of ratings. Tolor (1973), for example, compared high school students' choice of "most" and "least" effective teachers with administrators, other faculty, and parent choices. He found moderate agreement among groups about the effective teachers, but students labeled as "ineffective" a quite different set of teachers than did the other groups. Centra (1973b) compared teachers' self-evaluations with student ratings and found little agreement (median correlation = .21). In addition, the college instructors in his sample rated themselves better than students did. The widely-read review of Costin et al., (1971) gives other studies showing low to moderate relationships between student ratings and others' ratings. Interestingly, Costin et al. view this finding as "support [for] the contention that student ratings have a contribution of their own to make in the evaluation of teaching" (p. 517). A more cynical perspective is that students are not only different in their ratings, but also no better!

Student Learning.

Many have insisted that student achievement is the most decisive external criterion for ratings. But there is no agreement in the literature about the relationship between these two factors. While many studies have reported low, positive relationships, others (see Turner and Thompson, 1974) have found negative or no relationships. Costin et al., (1971) and Kulik

and McKeachie (1975) provide a representative sample of these studies. Recently, some researchers have decided to publicly debate the issue (Rodin and Rodin, 1972; Rodin, 1973; Frey, 1973, 1974). Although both factions would disagree, ~~with~~ the myriad differences in their research methodologies permits a decision of "no decision."

One of the great difficulties in the use of student achievement as a criterion is multiple meanings. Rosenshine's otherwise excellent work, Teaching Behaviors and Student Achievement (1971) has been roundly criticized (Gall, 1973) because of vagueness in defining "achievement." The Rodins-Frey debate is striking because of its lack of attention to validity and reliability of their achievement measures. Any research using student achievement should explicate clearly the meaning, context, validity, and reliability of such measures.

Halo Effect.

Some researchers have discounted a generalized halo effect[1] running through a set of teacher rating items, usually on the basis of factor analyses which reveal several separate dimensions underlying a complete scale. However, factorial validity and stability do not necessarily preclude a halo effect; in fact, another interpretation is that the halo effect is merely multidimensional! Other theory and research (Widlak et al., 1973) gives evidence for the notion of the halo effect in ratings. Cronbach (1958) suggested that raters often carry internal sterotypes (he called it "implicit personality theory") about clusters of attributes

[1]Other researchers have developed euphemisms for the halo effect in student ratings. Aleamoni (1973), for instance, called it the "general course attitude."

that people are "supposed" to have. Implicit personality theory can thus be an explanatory mechanism for the halo effect. Support for implicit personality theory's influence on ratings comes from a variety of sources. Person-perception research, for instance, would predict that students would rate lower those instructors whose attitudes were perceived to be different from those of the students. Good and Good (1973) and Levensen and LeUnes (1974) found this to be the case.

Passini and Norman (1966) found that highly similar factor structures emerged for two groups of raters: one group knew the people they were rating; the other group did not. Their implication that a priori impressions reduce rater objectivity formed the basis for later research by Magoon and Price (1972). Magoon and Price found congruence in rating factors between students who rated their instructors before the course began, and students who rated instructors after the course. Oddly, they discounted the halo effect as an explanation for this finding. Rather, they said, the "item relationships [seemed to be based on raters'] previous experience with other instructors" (p. 9). Nevertheless, they conclude (p. 9) that ratings may suggest more about "preconceptions of students than about real differences between courses and instructors." Another way of stating this conclusion is that interrater reliability may tell us more about the consistency of classification schemes than it does about the actual effectiveness of instructor.

Whitely and Doyle (1975) have supported this assertion by examining the congruence of factor analytic dimensions across raters, courses, and instructors. Most striking in their study was the correspondence between underlying dimensions of actual ratings, and clusters of rating items

which students were asked to group into homogenous sets. Ghiselli and Ghiselli (1972) have perhaps most clearly summed up the influence of implicit personality theory:

> [T]he report the rater makes about a stimulus person is not a faithful reflection of the qualities that person possesses or manifests, but rather is a report of his impression of that person, a description of his mental reaction to him. This reaction, of course, is conditioned by his social and cultural background (p. 270).

Course or Class Level and Course Size.

Some research suggests that systematic differences in instructor ratings occur as a function of the class level of students. Tolor (1973) found that high school students' judgments about teachers were related to the students' class level (e.g., sophomore, junior, etc.) Aleamoni and Graham (1974) discovered similar outcomes at the college level. Class size has been shown to influence ratings in a negative fashion. The larger the course, the lower the ratings (McKeachie, 1975; Paulus, 1973, Klafehn, 1975; Scott, 1975).

Reliability.

Almost all recent studies on student ratings appear to dismiss reliability quickly and cavalierly, by one of two methods. First instead of calculating a reliability estimate for the measure used, one can refer to, say, Costin, Greenough and Menges' (1971) review to find ample support for the reliability of such measures. Second, researchers can calculate their own estimates for a measure at hand. In either case, reliability estimates tend to be flawed, because the same erroneous techniques continue to be used.

Medley and Mitzel (1963, p. 253) correctly pointed out that "the term reliability coefficient refer[s]to the correlation to be expected between scores based on observations made by different observers at different times" (italics added). Rarely do we see this type of coefficient used.[1] Rather, we hear of "stability" estimates (e.g., student rankings now vs. student rankings a year later) and internal consistency estimates. Costin et al., (1971) reported rather high stability estimates—.48 to .89. The magnitudes are about that high in Bausell and Magoon's (1972) correlations between first day and last day ratings. However, Bausell and Magoon acknowledged that the high stability may mean either accuracy of ratings or durability of student bias.

Medley and Mitzel (1963) cautioned that a halo effect will add common variance to "different" rating items, and a scale must necessarily—and spuriously—build internal consistency. Also, they remarked, to the extent a halo effect is persistent over time, stability estimates will be inflated. The halo is likely to be detected by high correlations among items attempting to measure conceptually different teacher behaviors. As such correlations are fairly easy to find, one wonders about the level of exaggeration in "reliability" estimates of student ratings.

Another disconcerting influence of rating scale reliabilities occurs when rating scales are examined for their relationship to such variables as student and teacher characteristics. Often, in studies of how student and teacher personalities affect ratings, several measurements are regressed against some rating criterion. Grush and Costin (1975) and

---

[1] Kulik and McKeachie (1975, pp. 222-223) give a few examples of such estimates; the range shown is .34 to .67.

Treffinger and Feldhusen (1970) provide examples of this type of analysis. Treffinger and Feldhusen found modest multiple correlations (about .40) between a battery of student characteristics, and end of course ratings. They concluded that student variables "only" account for 21 percent of the criterion variance. Since the reliability of the criterion sets an upper boundary on its predictability (Cureton, 1965, p. 344)[1], Treffinger and Feldhusen's comment about 21 percent of the criterion variance is meaningless until we know how much criterion variance is reliable and thus predictable. Should Treffinger and Feldhusen's criterion have a reliability estimate of only .50, then they have actually accounted for 42 percent of the predictable criterion variance. Obviously, reliability estimates can change our ideas about the magnitude of relationships between ratings and other variables.

The Politics of Evaluation.

There is no question that student evaluations carry political overtones. Teacher organizations and unions are perhaps the most vocal opponents of student ratings[2]. If "excellence" in teaching is decided by students, and maintained by a reward system, it is feared that excellence may deterioriate to subservience: those who control merit rewards are in a position to "call the shots" about how teachers should behave (Bolton, 1972). Also, the emotional dimension of evaluative ratings

---

[1] "Variance accounted for" is the following proportion:

$$\frac{R^2}{\text{reliability of criterion}}$$

[2] They also have opposed most other types of teacher evaluation: see Selden's (1969) American Federation of Teachers position paper on evaluation.

often produces tension, hostility, and strain in interpersonal relationships (Gruenfeld and Weissenberg, 1966; Kerlinger, 1971).

The American Federation of Teachers (Selden, 1969) has claimed that, beyond an initial probationary period, evaluation of teachers is not a legitimate means of improving education. Bolton (1972) finds this attitude somewhat akin to disregarding the performance of a baseball player after he has played for a couple of years. Perhaps teachers are not ideologically ready to accept a critical evaluation of their classroom performance. We have been free from rigorous evaluation for a long time. Postman and Weingartner's proposal (1969, p. 139) that students should "classify teachers according to their ability" was once laughable; today the laughter has a nervous ring to it. Even if teachers are assured that student ratings are "merely" measures of satisfaction, their fears are not allayed. Teachers are afraid of students doing the evaluating, but as I have tried to establish in this paper, their fears are not entirely groundless. A cursory review of the literature is enough to make most of us stand in awe at the confusion surrounding student ratings.

An Immodest Proposal.

If we are to make more sense of student rating instruments, and the scores derived from them, I believe that we should begin a threefold approach. The three steps would seem to logically follow the sequence presented below.

First, I would propose a moratorium on student ratings as evaluative measures.[1] Admittedly, they may be one of the best available sources of

[1]Given the evidence about teacher preparedness in assessing student performance, maybe there should be a moratorium on all types of school evaluation...see Roeder (1973).

information about teacher competence (compared to such devices as administrator scuttlebutt, peer judgments, and self-rating). But the lack of clear meaning or validity in student ratings invites misuse and continued disagreement about their worth. Discontinuing student ratings seems to run counter to the ever mounting press for accountability; but it at least provides an opportunity to clear some of the evaluative smog that has been blurring our vision and stinging our sensibilities.

Second, we need to relearn some old lessons on important properties of rating scales. This remark implies that existing instruments need to be refined until they satisfy several minimal criteria outlined by Remmmers (1962, p. 330):

1. Objectivity. Use of the instrument should yield verifiable, reproducable data not a function of the peculiar characteristics of the rater.

2. Reliability. It should yield the same values, within the limits of allowable error, under the same set of conditions.... This criterion boils down to the accuracy of observations by the rater[s]....

3. Sensitivity. It should yield as fine distinctions as are typically made in communicating about the object of investigation.

4. Validity. Its content, in this case the categories in the rating scale, should be relevant to a defined area of investigation and to some relevant behavioral science construct; if possible, the data should be covariant with some other, experimentally independent, index....

5. Utility. It should efficiently yield information relevant to contemporary theoretical and practical issues.

Finally, student ratings should be studied hard and long. This supports the idea that their use, for the time being, must be experimental and not evaluative. There is enough exploratory research to give us some good ideas about programmatic research. Here, then, are a few suggestions

for research directions:

1. What is the influence of rater anonymity? We keep pretending that students will change their ratings if they have to reveal their identity. The caution of anonymity in ratings should be build on direct evidence not intuition.[1]

2. What are the differential effects on ratings when they are done before vs. after grades are awarded? Do these effects interact with student, teacher, or subject matter characteristics?

3. What is the criterion-related validity of ratings, using a variety of criteria, such as residualized student achievement, student affect toward course content, teacher behavioral change, student social behavior, and direct observations of teacher behavior by trained observers? Bolton (1972), Turner (1973) and Smith (1974) have proposed the use of "jury" models for weighting the various sources of evaluative information. Similarly, the conglomerate of criteria proposed here can be used as a multivariate view of the outcomes of teacher behavior. A variety of multivariate techniques are available to handle source, method (Halstead, 1970), and outcome variables; multiple regression, discriminant analysis, factor analysis, canonical analysis, and multiple analysis of variance and covariance methods have been used to rarely.

4. How are different rating formats related to other external criteria? There is some evidence that format changes produce rating changes (Follman et al., 1974), as well as evidence that they do not (Froman, 1976). Are there certain circumstances, (i.e., types of students, or types of rating items) which interact with format to produce higher or lower ratings?

5. Under what circumstances does provision of evaluative feedback help teachers improve? Is there a difference, for example, between the informational "worth" of high inference or low inference feedback? Many studies have addressed the issue of feedback; Trent and Cohen (1975 p. 1046) provide a good review. But it is not yet clear under what circumstances feedback improves instruction. Again, multivariate techniques allow the simultaneous consideration of many possible interactions between teacher, student, and school characteristics; content or subject area; type of feedback; and regularity of feedback.

6. What is the degree of interaction between teacher evaluation

---

[1] One unpublished research report (Anon., n.d.) supports the common view that identified student ratings will be higher than anonymous responses.

procedures and student evaluation procedures? For instance, does a teacher's use of norm-referenced vs. criterion-referenced grading system influence student ratings?

7. Can students be trained to be (more) objective raters? That is, can they be trained to make judgments about teacher behavior which are apart from, but not necessarily inconsistent with, "consumer satisfaction?" What types of students are most objective in rating what types of teachers? Is it possible for students to employ a common, consistent frame of reference? (That they currently do not is suggested by the research of Sanders and Lynch (1973)).

8. How can we build rating instruments that distinguish the middle ranges of teaching ability as well as the very good and the very poor?

9. Is it possible to build a rating system which balances the teacher attributes that students value with teacher characteristics that produce good learning?

10. Given the evidence that teachers show only moderate stability in producing student learning (Brophy, 1973), can we expect student ratings to show only moderate stability? (The stability issue implies an important but ordinarily forgotten rule of good research: replication.) What are the implications for ratings if broad intrateacher fluctuations are found? What are the implications if factorial invariance (or stability) of rating scales is not found (Villano and Rosenstock, 1973)?

11. Are the future answers to any of the previous questions moderated by the purpose of teacher evaluation, or can the findings be generalized across purposes?

Bibliography

Aleamoni, L.M. Evaluation by students to identify general instructional problems. Symposium presentation at AERA, New Orleans, February 1973.

Aleamoni, L.M. & M.H. Graham. The relationship between CEQ ratings and instructor's rank, class size, and course level. J. Educ. Meas., 1974, 11(3), 189-202.

Anon. Course evaluation questionnaire: anonymous vs. identified student responses. Office of Instructional Resources Research Report #202. Urbana, IL: Univ. of Illinois, n.d.

Bausell, R.B. & J. Magoon. The persistence of first impressions in course and instructor evaluations. Paper presented at AERA, Chicago, April 1972.

Bausell, R.B. & J. Magoon. Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educ. & Psychol. Meas., 1972, 32, 1013-1023.

Boltcn, D.L. Teacher evaluation (PREP Report #21). Washington, D.C.: HEW, Office of Education, 1972.

Brophy, J.E. Stability of teacher effectiveness. Amer. Educ. Res. J., 1973, 10(3), 245-252.

Brown, D.L. Faculty ratings and student grades: a university-wide multiple regression analysis. J. Educ. Psychol., in press.

Centra, J.A. The student as Godfather? The impact of student ratings on academia. Educ. Researcher, 1973a, 2(10), 4-8.

Centra, J.A. Self-ratings of college teachers: a comparison with student ratings. J. Educ. Meas., 1973b, 10(4), 287-296.

Costin, F., W.T. Greenough, & R.J. Menges. Student ratings of college teaching: reliability, validity, and usefulness. Rev. Educ. Res., 1971, 41(5), 511-535.

Cronbach, L.J. Proposals leading to analytic treatment of social perception scores. From R. Tagiuri & L. Petrullo (eds.), Person, perception, and interpersonal behavior. Stanford: Stanford Univ. Press, 1958.

Cureton, E.E. Reliability and validity: basic assumptions and experimental designs. Educ. & Psychol. Meas., 1965, 25(2), 327-346.

Deshpande, A.S., S.C. Webb, & E. Marks. Student perception of engineering instructor behaviors and their relationship to the evaluation of instructors and courses. Amer. Educ. Res. J., 1970, 7, 289-305.

Doyle, K.O. & S.E. Whitely. Student ratings as criteria for effective teaching. Amer.Educ. Res. J., 1974, 11(3), 259-274.

Frey, P.W. Student instructional ratings and faculty performance. Science, 1973, 178,

Frey, P.W. The ongoing debate: student evaluation of teaching. Change, 1974, 6(1), 47f.

Follman, J., M. Lucoff, L. Small, & F. Power. Kinds of keys of student ratings of faculty teaching effectiveness. A paper presented at AERA, Chicago, April 1974.

Froman, R.D. The influence of format change on the halo effect of student ratings. Symposium presentation at NCME, San Francisco, April 1976.

Gall, M.D. The problem of "student achievement" in research on teacher effects. (Report A73-2, Teacher Educ. Divn. Publication Series). San Francisco: Far West Lab for Educ. R & D, 1973.

Ghiselli, E.E. & W.B. Ghiselli. Ratings—kundgabe or beschreibung. J. Psychol., 1972, 80, 263-271.

Good, K.C. & L.R. Good. Attitude similarity and attraction to an instructor. Psychol. Rep., 1973, 33, 335-337.

Gruenfeld, L.W. & P. Weissenberg. Supervisory characteristics and attitudes toward performance appraisals. Personnel Psychol., 1966, 19, 143-151.

Grush, J.E. & F. Costin. The student as consumer of the teaching process. Amer. Educ. Res. J., 1975, 12(1), 55-66.

Halstead, J.S. A model for research on ratings of courses and instructors. Paper presented at APA, Miami, August 1970.

Holmes, D.S. The relationship between expected grades and students' evaluations of their instructors. Educ. & Psychol. Meas., 1971, 31, 951-957.

Holmes, D.S. Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. J. Educ. Psychol., 1972, 63(2), 130-133.

Kennedy, W.R. Grades expected and grades received—their relationship to students' evaluations of faculty performance. J. Educ. Psychol., 1975, 67(1), 109-115.

Kerlinger, F.N. Student evaluation of university professors. Sch. & Soc., 1971, 99, 353-356.

Klafehn, K.A. A multivariate analysis of teaching effectiveness. Quant. Meth., 1975, 2(1), 29-35.

Kulik, J.A. & W.J. McKeachie. The evaluation of teachers in higher education. From F.N. Kerlinger (ed.), Review of research in education III. Chicago: Rand McNally, 1975, 210-240.

Levensen, H. & A. LeUnes. Students' evaluation of an instructor: effects of similarity of attitudes. Psychol Rep., 1974, 34, 1074.

Magoon, A.J. & J.R. Price. Rating dimensions of course and instructor characteristics: the eye of the beholder. Paper presented at AERA, Chicago, April 1972.

McKeachie, W.J. Student evaluations keyed to function. Paper presented at AERA, Chicago, April 1972.

Medley, D.M. & H.E. Mitzel. Measuring classroom behavior by systematic observation. From N.L. Gage (ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 247-328.

Passini, F.T. & W.T. Norman. A universal conception of personality structure? J. Per. & Soc. Psychol., 1966, 4(1), 44-49.

Paulus, D.H. Normative information on the Univ. of Connecticut rating scale for instruction. Bureau of Educ. Research & Service, Storrs, CT: Univ. of Connecticut, 1973.

Postman, N. & C. Weingartner. Teaching as a subversive activity. N.Y.: Delacorte Press, 1970.

Lolli, A. & S.V. Owen. Student ratings: what is the frame of reference? Paper presented at NCME, San Francisco, April 1976.

Potter, D., P. Nalin, & A. Lewandowski. The relation of student achievement and student ratings of teachers. Paper presented at AERA, New Orleans, April 1973.

Remmers, H.H. Rating methods in research on teaching. From N.L. Gage (ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 329-378.

Rodin, M. & B. Rodin. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Rodin, M. Can students evaluate good teaching? Change, 1973, 5(6), 66f.

Roeder, H.H. Teacher education curricula—your final grade is F. J. Educ. Meas., 1973, 10(2), 141-143.

Rosenshine, B. Teacher behaviours and student achievement. London: Nat'l. Foundation Educ. Res. in England & Wales, 1971.

Sanders, J.R. & M. Lynch. Student evaluation of instruction: the analysis of discrepancies between perceived and ideal conditions. Bloomington, Indiana: Educ. Research & Eval. Lab, Indiana University, 1973.

Scott, C.S. Correlates of student ratings of professorial performance: instructor defined extenuating circumstances, class size, and faculty member's professional experience and willingness to publish results. Paper presented at AERA, Washington, April 1975.

Selden, D. Evaluate teachers? AFT Quest Paper No. 4. Washington, D.C.: American Federation of Teachers, 1969.

Slater, J.K. and S.V. Owen. Departmental differences in student perception of the "ideal" teacher. A paper presented at NERA, Ellenville, N.Y., October 1974.

Smith, F. A jury model for the evaluation of teacher competency. Paper presented at NERA, Ellenville, N.Y., October 1974.

Stewart, C.T. & L.F. Malpass. Estimates of achievement and ratings of instructors. J. Educ. Res., 1966, 59, 347-350.

Tolor, A. Evaluation of perceived teacher effectiveness. J. Educ. Psychol., 1973, 64(1), 98-104.

Treffinger, D.J. & J.F. Feldhusen. Predicting student ratings of instruction. Paper presented at APA, Miami, August 1970.

Trent, J.W. & A.M. Cohen. Research on teaching in higher education. From R.M.W. Travers (ed.), Second handbook of research on teaching. Chicago: Rand Mcnally, 1975, 997-1071.

Turner, R.L. Evaluating the validity of assessed performances: methodological problems. Paper presented at AERA, New Orleans, February 1973.

Turner, R.L. & R.P. Thompson. Relationship between student ratings of instructors and residual learning. Paper presented at AERA, Chicago, April 1974.

Veldman, D.J. & R.F. Peck. Influences on pupil evaluations of student teachers. J. Educ. Psychol., 1969, 60(2), 103-108.

Villano, M.W. & E.H. Rosenstock. A decade with the course attitude questionnaire: a factorial study. Paper presented at NERA, Ellenville, N.Y., October 1973.

Widlak, F.W., E.D. McDaniel, and J.F. Feldhusen. Factor analysis of an instructor rating scale. Paper presented at AERA, New Orleans, February 1973.