DOCUMENT RESUME

ED 228 228 SP 022 158

AUTHOR Peterson, Ken; Kauchak, Don
TITLE Progress on Development of Lines of Evidence for the Evaluation of Public School Teaching.
INSTITUTION Utah Univ., Salt Lake City. Center for Educational Practice.
SPONS AGENCY Utah State Office of Education, Salt Lake City.
PUB DATE Apr 83
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Canada, April 1983).
PUB TYPE Information Analyses (070) -- Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Observation Techniques; Educational Research; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; *Evaluation Needs; Inservice Teacher Education; Peer Evaluation; Research Methodology; Student Evaluation; Teacher Effectiveness; *Teacher Evaluation
IDENTIFIERS *Utah

ABSTRACT

The method used for a developmental study of teacher evaluation techniques (part of the Utah Teacher Evaluation Project) was a review of research literature to find those teacher evaluation practices with the most promise. Each technique was further researched in order to bring it to a "recommended practice" condition with limitations, cautions, and format ground rules. Current practices were also analyzed, and empirical trials were initiated of the best practice conditions to gauge reliability and feasibility. In this preliminary report on the development of the teacher evaluation techniques, seven "lines of evidence" are identified as showing promise and are briefly discussed: (1) pupil report; (2) peer review of materials; (3) teacher tests; (4) administrator rating; (5) systematic observation; (6) pupil gain (special cases); and (7) other (special or idiosyncratic cases). (JM)

PROGRESS ON DEVELOPMENT OF LINES OF EVIDENCE FOR THE EVALUATION OF PUBLIC SCHOOL TEACHING

Ken Peterson, Ph.D.
Don Kauchak, Ed.D.

Utah Teacher Evaluation Project
Department of Educational Studies
and
Center for Educational Practice
University of Utah
Salt Lake City, Utah 84112

Presented at American Educational Research Association,
Montreal, April 1983

Views expressed in this paper do not necessarily reflect those of the Utah State Office of Education or University of Utah Center for Educational Practice which are funding R&D activities of the Utah Teacher Evaluation Project.

PROGRESS ON DEVELOPMENT OF LINES OF EVIDENCE FOR THE EVALUATION OF PUBLIC SCHOOL TEACHING

Ken Peterson and Don Kauchak, University of Utah

INTRODUCTION

The purpose of this paper is to give a preliminary report on the development of lines of evidence for teacher evaluation which is part of the Utah Teacher Evaluation Project (UTEP). Gathering evidence about teacher merit or value is a technical problem area of the larger question of improving teacher evaluation practice altogether. Significant improvement in teacher evaluation practice will require not only development of the technical problems of assessment, discussed here, but also considerable attention to teacher sociology and the politics of teacher evaluation. While these two latter problem areas are important parts of the UTEP, this paper will address only the technical problems of gathering evidence.

The UTEP is involved in teacher evaluation which is both formative and summative. This is an important distinction from many other current projects which are merely formative. Because of the needs for summative evaluation of many teacher evaluation audiences, e.g., lay public, legislatures, and higher education, it is somewhat misleading to label many formative projects as "evaluation;" inservice or professional feedback would be more appropriate terms.

The goals of the UTEP are to (a) make the value or impact of teacher efforts visible to a larger number of interested audiences, (b) provide a basis for the "authoritative reassurance" (Lortie, 1974) of teachers, (c) develop effective feedback procedures for the improvement of teacher education, and (d) increase the basic understanding of the sociological

and political dynamics of teacher evaluation. The current status of the UTEP is developmental. R&D are focused on sociological and political topics, as well as techniques of assessment. While work is currently underway in four school districts, a complete system of evaluation has not been installed in any. We recognize that a more simple approach to teacher evaluation than one which addresses sociology, politics, and multiple lines of evidence would be desireable; however at this time, no such system can be pointed out as successfully functioning in the country (Scriven, 1981).

BACKGROUND

A. Need Need for development of teacher evaluation is based on arguments which have been developed in the literature (Millman, 1981).

1. At present, many legitimate audiences for teacher evaluation are not addressed. Evaluation is not just for administrators and teachers improvement, it has import for the decisions and operations of others in the society.
2. New, multiple lines of evidence about teacher impact or value need to be developed. Single views of what makes a good teacher do not enable improved evaluation.
3. The profession is accountable to teachers for mechanisms which provide useable information about how well they are doing.
4. If improvements in summative evaluation are not made by the profession, then non-professionals will dominate with haphazard evaluation, or worse, with narrow and deficient systems of teacher review.
5. It is not entirely clear that the problem of the bad teacher is being adequately dealt with.

B. Evaluation Model Employed The model of evaluation for which the lines of evidence are to be gathered is what Glass (1981) has called a combined goals model, which in turn is a refinement of what Scriven (1973) first developed as goal-free evaluation. House (1980) further described goal-free evaluation as "utilitarian" in its assumptions, and having an "objectionist epistemology" which includes qualitative as well as quantitative objectivity. In this approach to evaluation, goals are defined in terms of the needs and understandings of audiences who use the evaluation

results. Thus, a combination of goals, and not just those of the teachers or districts themselves, must be looked at in terms of formative and summative decisions, and other determinations of value. In addition, the model is goal-free in the sense that value determinations, including comparisons, are free to a specific, standard set of criteria. This means that teachers may be evaluated, and compared, on the basis of different combinations of lines of evidence. While this may seem like a process of comparing apples and oranges, it is more a case of comparing the data which are most pertinent (whether apples, oranges, or pears) in order to determine the quality of the "fruitiness." A final component of goal-free evaluation is to examine the merit and adequacy of the goals themselves, as they are finally addressed in the evaluation.

Essentially, our procedures call for each teacher to make their best case for merit or value, using the most appropriate evidence. It is these cases which are compared for summative purposes. Each case, regardless of particular combination of lines of evidence, can be subjected to tests of credibility and adequacy. There are better cases than others. For example, a teacher who reports her own student ratings in one class will not have the credibility or adequacy of another who used an outside party to collect student ratings over a three year period, and who also included student achievement data and positive peer reviews of her materials used in class.

Identification of a deficient teacher is based upon the situation in which a person cannot provide credible evidence from any line, or too few lines, in order to be considered to have an adequate case. For example, if a teacher produced only evidence about student acceptance, this would be considered as positive, but perhaps inadequate in terms of quality of materials, student outcome, or administrator report. This summative judgment would have to be made in terms of teachers in a similar situation,

but in all likelihood would be found to be deficient: mere student acceptance is not strong enough ground, given the other possibilities for evidence of valuable performance or impact.

A teacher is free to inspect evidence before deciding to submit it. Of course this leads to the likelihood of teachers submitting only evidence which is positive. But this does not affect the outcome of the judgment. Lack of evidence is less important than the presence of a credible, positive case. In fact, this possibility of censorship enables teachers to exclude lines of evidence that are not appropriate in their setting, for example pupil gain scores where there are no good tests, or peer review where the teacher is unpopular with colleagues but otherwise effective with students.

This impact- or value-based evaluation is in distinction to criterion-based evaluation in which each individual line of evidence is compared with a fixed standard of performance, or with standing in the entire group in a normative sense. Competency-based evaluation, for example, compares each teacher with a uniform set of criteria of minimal performance capacities. Systematic observation and teacher knowledge tests, by themselves, likewise compare each teacher with a fixed standard or norms of performance. Our approach is to look to the value or impact, given the evidence, in order to make a summative judgment.

One problem that is avoided with a model which incorporates variable lines of evidence is that of the difficulty of using evidence which is inappropriate for even a minority of the teachers. For example, when teacher knowledge tests are used as the (or one among others) evaluation criterion, it can be shown that a small number of quite valuable teachers show poorly on the single measure, and that the irrelevance of the measure can be demonstrated in their case. The common result in practice is to protect these few teachers, and in doing so to obviate the entire evaluation

system. A fatal flaw of any criterion-based evaluation scheme will be the small minority of teachers who can make the credible case that the critical lines of evidence they are required to use are inappropriate, irrelevent, or misleading in their situation. As an alternative, we suggest that the teacher provide other evidence, not necessarily of the same kind, in order to make their case of merit.

STUDY PROCEDURES

The method for this developmental study of teacher evaluation techniques was to review the research literature in order to determine those with the most promise, then to design small scale applications in order to refine them. First, the range of possible techniques was limited to those which show most promise. For example, self-report may play an important role in professional functioning and development, but its use for summative purposes looks remote. On the other hand, peer involvement seems to be a desireable and possible line of evidence according to the research literature. Second, each line was further researched in order to bring it to a "recommended practice" condition with limitations, cautions, and format groundrules. For example, while peer involvement shows great promise, it is quite clear from the literature that actual classroom visits of peers are far too unreliable for summative and even many formative purposes. A third stage of inquiry for some of the promising lines of evidence has been to do an analysis of current practice. For example, administrator visits and ratings are a common technique in Utah, so an item analysis of the 30 rating forms which are in current use was completed. The fourth stage of study has been to initiate empirical trials of the best practices conditions, in which we determine estimates of reliability, make validity cases, rewrite the groundrules for use, document problems and successes, and make operational the potential of the different lines of evidence.

Future study and development of the UTEP will be to install these lines of evidence in an ongoing district evaluation program. At that time we will be able to address questions of the relationships among different lines of evidence, and teacher performance. The goal will not be a single global measure, since the lines are expected to operate with considerable independence. Quality teaching is made up from various teacher structures and combinations of performance.

FINDINGS

LINES OF EVIDENCE EXCLUDED AT THIS TIME

The literature review led to the exclusion of the lines of evidence for teacher evaluation listed in Table 1. Discussions of these lines of evidence have been well presented by McNeil and Popham (1974), Berliner (1977), Soar (1973), Scriven (1977), Borich (1977), Travers (1981), and Millman (1981).

Self-reports, self-ratings	Competency-based evaluation
Personal characteristics	Credentials
Performance tests	Parent reports
Classroom environment	Graduate followups

TABLE 1: Unpromising Lines of Evidence

LINES OF EVIDENCE IDENTIFIED AS PROMISING

Table 2 presents the lines of evidence of teacher impact which the preliminary research suggests should be developed for use by teachers.

Pupil report	Systematic observation
Peer review of materials	Pupil gain (special cases)
Teacher tests	Other (special cases)
Administrator rating	Other (idiosyncratic)

TABLE 2: Promising Lines of Evidence

Each of these lines will be briefly discussed in the following sections.

Pupil Report

There is much agreement in the literature that pupil reports are an important part of teacher assessment (Aleamoni, 1981). Students know their

own situation well, they have a vested interest in good teaching, and they are closely familiar with the work of teachers. They can reliably and economically provide data on some, but certainly not all, aspects of teaching performance and impact. The challenges with student reports are to limit the questions and to reliably collect the data. Types of items which the literature suggests are desireable and undesireable are listed in Table 3. This selection was based on relationships of items to learning

Desireable	Undesireable
I know what I am expected to do.	Teacher is fair.
Materials are available in class.	Teacher knows subject matter well.
Teacher lets us know how well we're doing.	Teacher makes me want to come to class.
Teacher does not shame, humiliate, intimidate.	Always treats us like individuals.

TABLE 3: Examples of Student Report Items

outcomes, and reliance on items for which students are good reporters. For example, the item "teacher is fair" is not desireable because students have trouble judging classroom events: fairness in their own case can be arguable (say, from the teacher's point of view), and students have difficulty reading the situation for others. On the other hand, items which reflect the opportunity to be engaged in learning, e.g., "we know what we are expected to do," and "the class is busy with learning" are consistent with research on the outcome benefits of task engagement.

The second problem of student reports, in addition to item selection, is that of reliable sampling (Gilmore, Kane, and Naccarato, 1978). This is an easier task in the high schools where a teacher has five to seven independent classes; multiple class averages in a single year are defensible. In the elementary school, however, one class per year presents a reliability problem. This situation calls for data collected over a period of years, with attention given to stability and trends. The UTEP is currently working to determine reliability limits of a variety of collection

procedures.

The third problem of student reports addressed by the UTEP is that of their use below grade six. Scriven (1981) recognized the demonstrated acceptable use of rating forms above grade six, but suggested that more work is needed below this age level. A number of studies (see Haak, Kleiber, and Peck, 1972) suggest that rating forms can be used with much younger students. We are currently working on reporting schemes for use with beginning readers in the primary grades, similar to those reported by Haak et al. (1972).

Peer Review of Materials

Teacher peers provide a perspective on teacher functioning which is unique and valuable. They are able to take into account many context variables such as student characteristics, actual local resources and problems, current expectations, and other factors which are important in estimating a teacher's adequacy.

However, there are many problems with involving teachers in peer review. The technical assessment difficulty appear solvable, but sociological barriers to use remain. For example, while many persons, professional and lay, feel that the best way to judge performance is to take a peek in the classroom, the literature provides many cautions about the casual visit (as compared with that of the trained observer in systematic observation). Specific studies, such as those by Centra (1975) and Cook and Richards (1972), and more general discussions by Evertson and Holley (1981) and Scriven (1981), point out the great difficulties inherent with peer classroom visits. It is apparent that matters of politics, friendships, styles, and role expectations account more for the variance in ratings than does the actual variety of teaching.

Other problems with peer evaluation hinge upon teacher sociology. At present, peer evaluation is limited by such professional phenomena as

isolation (Lortie, 1974) and professional modesty (Wolfe, 1973). A major R&D focus of the UTEP, not discussed at length in this paper, is the need for changes in the social role, expectations, rewards, and relationships for teachers at the district level. Without these changes, one should expect to see teacher evaluation practices continue much as they are today.

The literature on peer review suggests certain conditions and an absolute limit to the questions to which the peers may address themselves. First, materials need to be collected by the teacher. These may include lesson plans, assignments, written feedback to students, tests, examples of student work, readings, and instructional materials. Next, peers should be selected from similar settings, and contexts; perhaps three should be involved. Examples of the limited questions appear in Table 4. The compres-

Is there appropriate challenge and difficulty for these students?
Does feedback contain useful information for learning?
Are grades defensible?
Is the content up-to-date, relevant?
Are district guidelines addressed?

TABLE 4: Questions for Peer Review of Materials

sed judgments of peers should be limited to well functioning (majority), deficient, or exemplary. The latter two findings should be accompanied by specifications.

Administrator Visits and Ratings

Administrator visits and reports have long been the staple, if not the only, means of evaluating teachers. While there may be superficial appeal in the evaluative powers and sensitivity of an experienced administrator looking in at a classroom for even a brief time, the severely limited reliability of such visits for judgments and ratings has long been described (McNeil and Popham, 1974; Cook and Richards, 1972). However, it is apparent that administrator visits will continue because of their wide acceptance and lack of viable replacements, and the continued need for administrator

supervision in the classroom. Our developmental work has been to assess current practice, identify the worst strategies for elimination, and identify useful administrator contributions to overall evaluation.

The analysis of teacher rating forms relied upon teacher effectiveness research findings, (e.g., Borich, 1977; Rosenshine and Furst, 1974; Brophy and Good, 1978), the literature of interaction analysis for keys to reliable observation methods (e.g., Flanders, 1970; Amidon and Hough, 1967; Duckett, 1980; Hough and Duncan, 1970), the Wood & Pohland (1979) analysis of rating forms in New Mexico Schools, and our analysis of Utah rating forms which was patterned after Wood and Pohland. Examples of findings and conclusions are reported next.

Both our analysis of Utah administrator rating forms of teachers and that performed in New Mexico identified six categories of items.

One large category of items was that of Personal Characteristics. These are hard to defend in light of the research on this topic. It seems clear that they should not occupy the more than 25% of all items found both in Utah and New Mexico. Personal Characteristics should be greatly limited to those few cases where, in legal terms, there is material and substantial disruption of school operations, beyond the "mere discomfort" of school administrators, and the protected free expression of teachers. Likewise, the Administrator/Manager Role of the teacher could be limited to evidence that normal school functions and operations are carried out by the teacher. A third large category of items found in rating forms was Teaching Role. If there is no other evidence about this area, such as pupil reports, then administrator rating of this topic is justified. However, if other lines of evidence exist, particularly student report, peer review of materials, teacher knowledge of pedagogy, and systematic observation, then much more defensible data gathering and judgments can be made. The increased accuracy,

validity, and perspective of these additional sources should make the entire procedure more fair and useable for teachers. At the same time, the principal would be relieved of a complicated task of technical data gathering which is beyond the resources and even preferences of many administrators. The additional categories of Organizational Membership Role, Professional Role, and Social Role (together making up approximately 25% of current forms) are areas where a principal is in position to make more accurate and useful ratings.

Rating forms can be improved by limiting items to areas in which the principal has direct and reliable access to data, and to rely on other sources for additional information. This should improve overall data gathering, and may serve to make the principal's role more direct and less political. We suggest that current practice came into being by lack of other data, and by compromise on a least offensive and dangerous system. Evidence is that the current practice is not well liked by participants (Wolfe, 1973; Lortie, 1974; McNeil and Popham, 1974).

A final type of item which was rarely seen in Utah and New Mexico was that of an administrator's global rating; overall assessment of teacher. These global ratings have worked out well in other evaluation situations. If this type of rating received attention as only one part of a only one technique, it might provide value for the entire system. The evidence would have to be carefully tested with other lines of information. It has the potential of overweighting because of sociological or political reasons, beyond the technical contribution which it makes.

Teacher Tests

Teacher tests are generally of three knowledge types: subject matter, basic skills, and pedagogy (Harris, 1981). While some states have developed standardized tests of these types, they are represented on the national

level by the National Teachers' Exams (NTE) of ETS. There is little reason for individual states or districts to develop these tests, as has been done in a number of locations. Teacher tests are often overlooked in local evaluation systems because of their relatively low importance for teaching performance. It is often said that given minimal levels of teacher knowledge, that other factors such as classroom presence, human relations skills, and experience are far more important determinants of teaching impact. Another common statement is that high test scores are no guarantee of good teaching. Many individuals are ready to describe teachers who seemed to know their subject matter, but could not "get it across" in the classroom.

However, while teacher tests get downplayed in importance in teacher evaluation, especially within the profession, there are audiences which look to teacher tests as perhaps the most important indicator of teacher quality (Time, 1982; Lyons, 1979; Keisling, 1982). It is this line of evidence which has caused some of the strongest criticism among the public and legislative audiences. At the same time, teacher groups and other educators resist using teacher tests as part of evaluation, (Harris, 1981).

Teacher tests as evidence of teacher quality are perhaps the best example of how the principle of multiple audiences influences evaluation practice. The needs and perspectives of any one audience make their contribution to the data collection and analysis scheme, but do not exclude the interests of other audiences.

Systematic Observation

Systematic observation is a line of evidence in which a trained observer documents the manifest classroom performance of a teacher in some conceptual/theoretical framework of effective or otherwise meritorious teaching (Duckett, 1980). It is a technique often associated with Competency-Based Teacher Evaluation. Systematic observation differs from the less

formal (and reliable) administrator or peer visits in that it requires: (1) trained and monitored observers, (2) a reliable and representative number of visits, (3) demonstrably fair sampling of behavior, (4) limited observational categories, (5) systematic data recording and analysis procedures, and (6) a conceptually coherent framework for interpretation.

Systematic observation is perhaps the most expensive to obtain line of evidence, because it requires substantial time of a trained observer. This means that observation should be reserved for optimum periods of teacher development. Just at the time of certification might be one of these desireable times for systematic observation data, if it were not for the instability of beginning teacher performance. The next optimum period probably is at the end of two or three years of experience; it is at this time that performance stabilizes, and teacher tenure decisions are made. Another period for systematic observation data might be after 10 years of classroom experience. At this time there often are questions about long term inservice and professional development, and a need to check for the retention of the performance capacities. It is clear that there needs to be more experience with systematic observation programs before definitive recommendations about practice can be developed.

The main concern about using systematic observation as a line of evidence in the evaluation of teachers is the conceptual/theoretical framework and content of the observations themselves. While great progress has been made in process-product observational studies, there still is not universal agreement among educators as to a single framework (or some combination) for judgment of quality. Promising candidate systems include engaged time behaviors (Stallings, 1980), classroom managment variables (Kounin, 1974), equal opportunity strategies (Good and Brophy, 1974), and others (Good, 1980). Although each of these is defensible, none qualifies

as the solitary sine qua non of valuable teacher performance. While many claim that a common technology of instruction is just around the corner, it does not yet exist. What does exist at present is a more limited indicator of quality than many researchers in this field hope for.

It is important in the consideration of systematic observation as a line of evidence to be clear about what the strategy provides, and does not provide. What is assessed is capacity for a teaching strategy variable, and not evidence that this capacity is used appropriately and consistently. Research (Coker, Medley, and Soar, 1980) suggests that the mere capacity does not guarantee results in performance. The capacity for effective strategies also requires their use when the context makes them effective, and enough consistency in order to make a difference. For these latter two considerations other lines of evidence need to be considered, particularly student report and pupil outcome.

While there are obvious problems with systematic observation in terms of teacher evaluation, there is enough development in this area to warrant further exploration and trials. Cautions will have to be developed at the same pace as the promises.

Pupil Achievement (Student Gain)

Pupil achievement is a desireable line of evidence to some audiences and an ananthema to others (Millman, 1981). Lay audiences in particular can assume that student performance, regardless of the process or approach used, is the key indicator of teaching quality. This has lead to practices such as publication in newspapers of standardized test results by schools and grades (e.g., Los Angeles Times, 1982). However, the technical problems of using pupil achievement for the evaluation of teachers are legion, and well documented. They have led researchers such as Travers (1981) to conclude that "the difficulties of assessing teacher effectiveness in terms of

test scores of pupils seem to be almost insuperable at this time." Such an opinion is held by the majority of those involved in teacher evaluation development at this time. Technical difficulties include the fact that variance in gain in many conventional standardized tests is accounted for by well over 50% by pretest performance, pretests are rarely used in practice, classroom test reliabilities are deficient, and the fact that good tests do not exist for much of what is taught in schools.

However, Scriven (1981) has pointed out a limited situation in which pupil gain, as measured by standardized test scores could be a contributor for the evaluation of perhaps a small number of teachers in a district. These would be teachers associated with unusual student gain over a significant period of time. (Presumably this evidence would need to be based on gains adjusted for prior achievement - see Soar, 1973). Such teachers deserve recognition, and the school system should have some provision for demonstrating the value of student achievement. The fact that such a measure, or line of evidence, would be directly irrelevent for the majority of teachers in the district (perhaps 85-90%) does not tell the whole story about the indirect effect the evidence would have for the district and outside audiences. Often, as Scriven points out, the mere existence of evaluation can improve practice throughout that system.

Other Lines of Evidence

"Other" lines of evidence encompass two distinct types: (1) lines that apply to small, specific groups of teachers, and (2) lines that are entirely idiosyncratic in terms of individual teachers. An example of the first type is teachers who are able to demonstrate credible evidence that they are meeting an established need for students in their district in specific ways, for example minority or handicapped persons (Momsen, 1983). This provides an opportunity for teachers who give special services to be

recognized for their contribution. This evidence goes into the mix with other lines. Of course, the need needs to be clearly established at the district level or school before the teacher completes the second stage of making their case of individual merit and worth. An example might be the sudden presence of new Southeastern Asian students with concrete language and culture needs in the classroom, which are addressed by the teacher in question.

The second kind of "other" evidence is an open category in which the teacher provides credible evidence about some unique skill, results, approach, or process which gives outstanding impact in their position. An example is a debate coach whose teams consistently win state tournament honors, journalism teacher whose newspaper routinely is acknowledged as outstanding, or a science teacher whose students enroll in an inordinate number of science classes later, relative to district averages. Each of these idiosyncratic lines of evidence should be considered in light of other evidence gathered, as no single line presents a complete case.

Conclusion

This paper has presented a brief sketch of lines of evidence which show promise, and some lines which at this time do not. These decisions have been largely based on literature review. The UTEP is currently continuing the literature analysis of these lines with the aim of writing a preliminary description of recommended practice. The next step will be to empirically refine these recommendations with a series of studies in actual settings in order to study reliability requirements, validity in the greater context of teacher performance, and the sociology and political reactions of their use. The final stage of development which is anticipated is to install these techniques in a complete district wide system of teacher evaluation in which the technical, sociological, and political knowledge

is applied.

The anticipated product of the UTEP is a set of recommendations with which a school district can begin to analyze the current state of technical, sociological, and political realities of teacher evaluation, and can begin to move toward improved practice which has more satisfaction for audiences involved. This development will assume increased basic understanding of the dynamics and forces involved, as well as increased practical experience.

REFERENCES

Amidon, T. and J. Hough. Interaction-Analysis: Theory, Research, and Application. Reading, Mass.: Addison-Wesley, 1967.

Berliner, D. Impediments to measuring teacher effectiveness, in G. Borich (ed.) The Appraisal of Teaching. Reading, Mass.: Addison-Wesley, 1977.

Borich, G. The Appraisal of Teaching: Concepts and Processes. Reading, Mass:: Addison-Wesley, 1977.

Centra, J. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-337.

Cook, M. and H. Richards. Dimensions of principal and supervisor ratings of teacher behavior, Journal of Experimental Education, 1972, 41, 11-14.

Duckett, W. (ed.). Observation and the Evaluation of Teaching. Bloomington, Ind: Phi Delta Kappa, 1980.

Evertson, C. and F. Holley. Classroom observation, in J. Millman (ed.) Handbook of Teacher Evaluation. Beverly Hills, Sage Publications, 1981.

Flanders, N. Analyzing Teacher Behavior. Reading, Mass.: Addison-Wesley, 1970.

Gilmore, G., M. Kane, and R. Naccarato. The generalizability of student ratings of instruction: Estimation of the teacher and course components, Journal of Educational Measurement, 1978, 15, 1-13.

Glass, G. The Growth of Evaluation Methodology. Laboratory of Educational Research, University of Colorado. Mimeograph, 1981.

Good, T. and J. Brophy. Looking in Classrooms (2nd Ed.). New York: Harper and Row, 1978.

Haak, R., D. Kleiber, and R. Peck. Student Evaluation of Teacher Instrument II. Austin, Texas: R&D Center for Teacher Education, 1972.

Harris, W. Teacher command of subject matter, in J. Millman (ed.) Handbook of Teacher Evaluation. Beverly Hills: Sage Publications, 1981.

Hough, J. and Duncan, J. Teaching: Description and Analysis. Reading, Mass.: Addison-Wesley, 1970.

House, E. Evaluating with Validity. Beverly Hills: Sage Publications, 1980.

Keisling, P. The class war we can't afford to lose, American Education, 1982, 18 (7), 4-11.

Lortie, D. Schoolteacher: A Sociological Study. Chicago: University of Chicago Press, 1974.

McNeil, J. and W. Popham. The assessment of teacher competence, in R. Travers (ed.) Second Handbook of Research on Teaching. Chicago: Rand McNalley, 1973.

Los Angeles Times, School test scores: How students in L.A. compare statewide, December 26, 1982. Pp. 1-2, part II.

Lyons, G. Why teachers can't teach, Texas Monthly, 1979, 7 (9), 124 ff.

Millman, J. Student achievement as a measure of teacher competence, in J. Millman (ed.) Handbook of Teacher Evaluation. Beverly Hills: Sage Publications, 1981.

Scriven, M. Goal-free evaluation, in E. House (ed.) School Evaluation. The Politics and Process. Berkeley: McCutchan, 1973.

Scriven, M. The evaluation of teachers and teaching, in G. Borich (ed.) The Appraisal of Teaching. Reading, Mass.: Addison-Wesley, 1977.

Scriven, M. Summative teacher evaluation, in J. Millman (ed.) Handbook of Teacher Evaluation. Beverly Hills: Sage Publications, 1981.

Soar, R. Teacher assessment problems and possibilities, Journal of Teacher Education. 1973, 24, 205-212.

Stallings, J. Allocated academic learning time revisited, or beyond time on task, Educational Researcher, 1980, 9 (11), 11-16.

Time Magazine, Help, teacher can't teach. June 16, 1980. Pp 54-63.

Travers, R. Criteria of good teaching, in J. Millman (ed.) Handbook of Teacher Evaluation. Beverly Hills: Sage Publications, 1981.

Wolf, R. How teachers feel about evaluation, in E. House (ed.) School Evaluation: The Politics and Process. Berkeley: McCutchan, 1973.

Wood, C and P. Pohland. Teacher evaluation: The myth and realities, in W. Duckett (ed.) Planning for the Evaluation of Teaching. Bloomington, Ind.: Phi Delta Kappa, 1979.