

Teacher Evaluation — A Review of Research

Benjy Levin

Research provides little support for most current teacher evaluation practices. Evaluation by students may be more valid and reliable than observations by supervisors.

Teacher evaluation can serve one or both of two purposes. The first is to guide decisions about hiring, retention, and promotion. The second is to help improve teaching. In each case the purpose suggests certain important questions. The first purpose, personnel decisions, requires comparison among teachers, so techniques used for evaluation must be fair. In research jargon this is a question of reliability—the extent to which an evaluation produces consistent results at different times and for different people—and validity—that an evaluation measures what it is supposed to measure.

The situation is different for the improvement of teaching. Here the crucial question is the extent to which evaluation influences the subsequent behavior and attitudes of teachers and students.

Scope of the Report

The literature on teacher evaluation is enormous. There are hundreds of articles and documents dealing with theories of teacher evaluation, models of teacher evaluation, forms for teacher evaluation, and examples of teacher evaluation. This is not to mention the even larger amount of material on evaluation in general. The guiding principle of this report, therefore, is to focus on research that provides evidence about the reliability, validity, and effects of various techniques of teacher evaluation.

When this criterion is applied, the mountains of documents previously referred to melt away. As one reviewer commented, "Most of the literature on evaluation in schools is not based on empirical research. It is parochial at best." (Natriello, 1977). In fact, for all the pages of writing, there is not very much evidence about the reliability, validity, or effects of most of the techniques used in teacher evaluation. What there is, however, makes interesting reading.

Modes of Teacher Evaluation

Six general approaches to teacher evaluation were identified in the literature. These provide a useful way to categorize the research results, although much of the research is applicable to more than one mode. The six are:

1. The use of students' ratings of teaching through questionnaires and other survey instruments;
2. Evaluation based on observation by supervisors, such as principals;
3. Evaluation using an observation instrument or system, such as the Flanders Interaction Analysis System;
4. Self-evaluation by teachers;
5. Evaluation based on gains shown by students on various tests;
6. Evaluation through specially designed "teaching tests."

Teacher Evaluation Through Ratings by Students

Evaluation of teachers by students became popular in universities and colleges a decade ago, and has remained in wide use. It is less common in public schools than in postsecondary institutions, but is used more extensively than most other modes of evaluation.

Most of the research on students' evaluations—and there has been a considerable amount of it—has been conducted using college students, although students from grade six and up have been involved in studies. The problems of varying sample sizes, research designs, methods of data analysis, and measurement instruments are significant.

The first concern that arises about students' evaluations of teachers is reliability. Although this varies from one rating form to another, the general consensus is that there are several highly reliable student rating tools now available (Mintzes, 1977; Costin, 1971; Lehman, 1974). Furthermore, on most instruments ratings of teachers and courses are consistent among students and over time. That is, different groups of students tend to give similar ratings to the same teachers, and the same students will rate teachers similarly at two different times.

A second question has to do with bias in students' ratings. Is it true, for example, that easy teachers receive better evaluations? How do grades in a course affect evaluation results? The evidence on these points is not complete, but it does provide some strong indications. As far as is known, student characteristics such as age and sex are not related to evaluation results (Smith and Brown, 1976), and easier courses do not receive higher ratings (Painter and Granzin, 1973). Students are serious about their ratings; two studies showed that ratings did not change significantly after students were told that the results would be used for making decisions about whether or not to promote the teachers being rated (Centra, 1976).

Some factors do influence ratings, however. Class

"The most difficult problem with students' evaluations of teachers has to do with their validity. Do they actually measure important aspects of teaching?"

size has an impact, although just what that impact is remains unclear. There is some evidence that a teacher's reputation influences the evaluations he or she receives (Abrami and others, 1976). Ratings are related, as might be expected, to student interest in the subject and general attitude toward school. Grades are also a factor (Costin, 1971); students who get higher marks tend to give better evaluations. However, which factor causes the other is uncertain.

The most difficult problem with students' evaluations of teachers has to do with their validity. Do they actually measure important aspects of teaching? This is a difficult question to answer because a rating is an expression of personal opinion. One way of assessing validity is to compare students' ratings of teachers with those of other groups, such as supervisors or other teachers. Such comparisons show that students' ratings are substantially different from those of the others. That is, students will identify different teachers as being good (Swartz, 1975). This has been interpreted by some to mean that students are poor judges of good teaching (Morrow, 1977), and by others to indicate that students have a unique view of teaching, not shared by others, and therefore that their input in evaluation is especially important. Gage and others (1971) found that student ratings of teachers and the students' scores on a test of lesson comprehension correlated well, indicating that students are good judges of their own learning. Also, the many different measurement instruments used rank the same elements of teaching as important. Clarity of presentation, enthusiasm, and empathy with students appear repeatedly as desirable characteristics of good teachers. If students do have a common view of what good teaching is, their ratings are more likely to be valid.

Finally, does feedback from student ratings have an effect on subsequent teaching? Evidence on this point is scanty and inconclusive, but it suggests that

"Overall, it appears that many of the fears about students' evaluations of teachers are not well founded, and that such evaluations can provide reliable, useful data for evaluation purposes."

formal ratings are not particularly effective in changing teachers' behavior (Medley, 1977); Natriello, 1977). As mentioned, this conclusion is in accord with other research in a wide variety of areas concerning behavior change.

Overall, it appears that many of the fears about students' evaluations of teachers are not well founded, and that such evaluations can provide reliable, useful data for evaluation purposes.

Ratings by Supervisors

Supervisor ratings are undoubtedly the most common form of teacher evaluation currently in use. The principal usually does this evaluation (NEA, 1964). The modal practice consists of two or three half-hour visits to the classroom followed by a meeting between the teacher and the evaluator. Robinson (1978) noted that most evaluators had had no training in observational techniques, and did little or no preparation before observing a teacher.

How valid are supervisors' ratings? Across time they tend to be reliable, of course, since the same evaluator rates the same teacher. However there is some disagreement among principals and other raters in judging teaching. Tuckman (1977) reported that the criteria used in evaluating teachers varied between elementary and secondary principals, with the former tending to value warmth, creativity, and organization, while the latter look for systematic, task-oriented, structured teaching. Wilcox (1976) found that principals, current students, and graduates differed in their ratings of secondary teachers. In a study by Swartz (1975), 72 vocational teachers were rated by principals, supervisors, other teachers, students, and through self-ratings. Significant differences were found among the groups, although the ratings of principals and supervisors were quite similar. McNeil and Popham (1973), reviewing research on the assessment of teacher competence up to 1973, criticized

supervisor ratings as being unsatisfactory on several grounds, including the confusion of a teacher's personality and staff relations with his or her teaching ability. Chan (1973) found that evaluations were dependent largely on the principal's philosophy of education.

No studies were found that examined the effect of supervisors' ratings on subsequent teaching behavior. Robinson (1978) reported that half or less of the Connecticut teachers he studied found their evaluations useful to them. Wolf (1973) suggested that most teachers did not see evaluation as being in their interests. Reavis (1978) concluded that teachers generally did not like traditional methods of supervision and evaluation.

In summary, serious questions exist about the validity and fairness of supervisor ratings of teachers as presently conducted.

Evaluation Using an Observation Instrument or System

Systematic observation of teaching involves the use of an instrument such as the Flanders Interaction Analysis System, which guides the observer in terms of the behavior to be looked at and the ways in which it is to be understood. The use of such instruments for evaluation is quite rare at present.

Generally classroom observation instruments have been carefully developed and validated. That is, they have been compared with other rating systems, and compared against themselves over time. However many such instruments exist, and not all have received the necessary care in development. The user needs to look for reliability data before selecting an instrument for evaluation purposes.

Researchers have used observation instruments extensively in the study of classroom behavior and in attempts to determine elements of teaching that affect students' learning. As has been noted earlier, this search has not shown any great rewards, causing Popham and McNeil to comment:

... observations are most beneficial for recording and analyzing the teaching act—not judging it. . . . Effective teaching cannot be proven by the presence or absence of any instructional variable . . . (1973, p. 233).

Evidence indicates that observation instruments can be very useful for giving teachers feedback on aspects of their teaching because they are reliable and focus on discrete aspects of the teaching process. However, the effect of such feedback on teachers' subsequent behavior is less certain. The mere provision of feedback is no guarantee that teaching will improve or change.

Self-evaluation involves the improvement of in-

struction through having teachers reflect on their own teaching and modify it accordingly. Issues of reliability and validity are of little moment here, because the use of self-evaluation is limited to the individual teacher. The key question becomes the extent to which self-evaluation produces changes in teaching practices.

Neely (1972) studied the attitudes of Oregon teachers about self-evaluation. These ranged from neutral to slightly favorable. In other words, Neely did not find great enthusiasm for self-evaluation among the teachers he surveyed. This supports the contention of Popham and McNeil (1973) that:

... there are only a few studies indicating that some teachers are self-directing in their learning and expend effort in judging their behavior on the basis of the consequences of their teaching as revealed by the actions of pupils. (p. 231)

On the other hand, Natriello (1977) cites evidence from U.S. armed forces studies that self-evaluation does result in behavior change. Johnston (Balzer, 1973) compared the effects of traditional and self-evaluation on 84 student teachers and found that those involved in self-evaluation had higher scores on indices of indirectness in teaching. He concluded that self-evaluation could produce changes in subsequent behavior, but that these changes were not always at the level of significance. Some of the uncertainty about self-evaluation might be removed, however, if the approaches described by Medley (1977) or Roper and others (1976) involving careful development of the self-evaluation process were used. Both systems involved teachers in identifying desirable classroom practices and developing ways of measuring their occurrence.

Finally, the effects of self-evaluation cannot be separated from teachers' overall attitudes towards evaluation. In a study by Wolf (1973), 58 percent of the teachers indicated that they were not encouraged to evaluate their classroom behavior, indicating that self-evaluation was unlikely to occur or to be productive.

Evaluation Based on Students' Gains on Tests

In an attempt to link teacher evaluation with students' learning, some school districts have tied evaluation to gains by students on tests or other previously agreed-on measures of learning, attitudes, or behavior. Again, however, this technique is used infrequently. There has been practically no research on the use of this method of evaluation, although experience with other forms of performance contracting in the United States has been disappointing. The disadvantages and dangers of such a system seem clear—teaching to the test, the loss of long-range objectives

"Issues of reliability and validity are of little moment here, because the use of self-evaluation is limited to the individual teacher. The key question becomes the extent to which self-evaluation produces changes in teaching practices."

in favor of short-term gains on test scores, and so on. Additionally, there is dispute about the extent to which teachers' abilities to produce gains in students' learning is stable. That is, a teacher may be more successful one year or with one group of students than in other settings. Thus evaluation based on a single year or class would be invalid. Rosenshine (Glass, 1974) reviewed several studies of stability with the conclusion that stability was low. Brophy (1976), while more optimistic about the stability of teachers' effects, still felt that these were not high enough to justify their use for accountability purposes. Similar conclusions were reached by Soar and Soar (1975) and Shavelson and Dempsey (1975). The former referred to the use of gain scores as having difficulties that were "extremely serious, if not disabling." Thus it appears that the use of student gains to evaluate teaching is not a particularly desirable approach.

Evaluation Through "Teaching Tests"

The idea of developing a "test" for teachers has been promoted most vigorously by Popham. Popham suggests that a number of teachers be given one or more identical objectives and samples of the measures that will be used to assess them. Teachers would then plan and teach a lesson to a controlled group of students, under controlled conditions. Following the lesson, the measurement tools are administered, and the students' scores on the posttest become an index of the teachers' effectiveness. The technique is aimed at the evaluation of teachers for personnel purposes rather than at the improvement of instruction. The advantages cited for it are objectivity and the measurement solely of student outcomes rather than of teaching methods.

In a critique of this approach, Glass (1974) argues that to be fair this method must be able to demon-

strate high reliability across topics and groups of students. That is, teacher ability to produce changes in test scores must be stable. However as was noted in the previous section, there is doubt about whether this is, in fact, the case. Consequently, the technique of evaluation through teaching tests has, as yet, uncertain reliability.

Current Practices

There is little systematic information about current teacher evaluation practices. Formal programs



Benjy Levin is Past Executive Director, The Manitoba Educational Research Council, Inc., Winnipeg, Canada.

of teacher evaluation, involving clearly defined criteria and procedures, are rather infrequent, with studies indicating that about 50 percent of schools have such a program (NEA, 1964). Of the six modes listed above, evidence suggests that supervisors' ratings are by far the most commonly used technique (Robinson, 1978). However the data supporting this contention are limited and somewhat out of date. A 1974 study of 500 school districts in the United States revealed that about one quarter of them had some sort of student evaluation of teachers or courses (Oldham, 1974). Use of all the other modes listed appears to be quite rare.

Some Cautions

Interpretation of the results of educational research needs to be made in the light of a number of problems. Most of these center around the fragmented nature of the research effort—single researchers, often graduate students, working without proper support or coordination with other research efforts. As a result, studies frequently have very small samples, use unique methodologies that are not comparable with those of other studies in the same area, and involve a vast array of measurement tools and instruments. All these factors make it difficult to generalize about the significance of the research. Thus, all the points made in this review about the meaning of the research done to date could well be disputed on the basis of methodological weaknesses in the studies. This is an important, if depressing, point. The entire field of teacher evaluation has suffered from a surplus of opinion and a shortage of evidence. Thus,

few of the findings of this paper should be regarded as more than tentative.

The second point to be made in summary is that research provides little support for current practices in teacher evaluation. One of the few things that can be safely said is that the prevalent system of evaluation for purposes of hiring or promotion through observation by supervisors is biased and subjective. The use of techniques that have greater promise for providing objective data, such as observation instruments or student ratings, is as yet uncommon.

In view of the research reported here, schools should reexamine their practices in teacher evaluation carefully. Certainly more extensive use of student evaluations and less reliance on ratings by principals and other supervisors would be warranted changes. Schools might also do well to involve teachers in the development of supervision and evaluation policies, since this is likely to increase teachers' commitment to and use of the results. Reliance on a single evaluation technique is unwise. Using several approaches to ensure that each teacher is judged as fairly as possible is one way to counteract the biases of the various methods that have been discussed. *EL*

References

- Abrami, Philip, Les Leventhal, and Raymond Perry. "Do Teacher Evaluation Forms Reveal as Much About Students as About Teachers?" *Journal of Educational Psychology* 68(4): 441-45; 1976.
- Alberta Teachers' Association. *Report on Opinions of Principals on the First Year Experience of Teachers Prepared in Alberta Universities*. Edmonton: ATA, 1973.
- Balzer, Levon, editor. *A Review of Research on Teacher Behavior*. Columbus, Ohio: ERIC/SMEAC, 1973.
- Blue, Terry. *The Effect of Written and Oral Student Evaluative Feedback and Selected Teacher and Student Demographic and Descriptive Variables on the Attitudes and Ratings of Teachers and Students*, ERIC # ED 135 855. 1977.
- Brophy, Jere. "Reflections on Research in Elementary Schools." *Journal of Teacher Education* 27(1): 31-34; 1976.
- Centra, John. "The Influence of Different Directions on Student Ratings of Instruction." *Journal of Educational Measurement* 13(4): 277-82; 1976.
- Chan, Peter. *A Study of the Relationship Between Principals' Philosophies and Teachers' Philosophies in Educational Practices and Teacher Evaluation*. Unpublished doctoral dissertation; University of Michigan, 1973.
- Costin, Frank, and others. "Student Ratings of College Teaching." *Review of Educational Research* 41(5): 511-35; 1971.
- Dunkin, Michael, and Bruce Biddle. *The Study of Teaching*. New York: Holt, Rinehart & Winston, 1974.
- Emmer, Edmund. *Instructor Perception, Content of Scale, and Feedback Effectiveness*. ERIC # ED 103 399. 1974.
- Gage, Nathaniel, and others. *Research Into Classroom Processes: Recent Developments and Next Steps*. New York: Teachers' College Press, 1971.
- Glass, Gene. "Teacher Effectiveness." In: Herbert Walberg, editor. *Evaluating Educational Performance*. Berkeley, California: McCutchan, 1974.

Lehman, Irvin. "Evaluating Teaching." In: William Gephart, editor. *The Evaluation of Teaching*. National Society of Professors of Educational Research, 1974. ERIC # ED 148 894.

McDonald, Frederick. "The State of the Art in Performance Assessment of Teaching Competence." In: Theodore Andrews, editor. *Assessment*. ERIC # ED 102 200. 1974.

McNeil, John, and James Popham. "The Assessment of Teacher Competence." In: Robert Travers, editor. *Second Handbook of Research on Teaching*. Chicago: Rand McNally, 1973.

Medley, Donald. *An Approach to the Definition and Measurement of Teacher Competency*. ERIC # ED 144 952. 1977.

Mintzes, Joel. *The Student Opinion Survey of Teaching*. ERIC # ED 146 195. 1977. Available from Department of Biology, University of Windsor.

Morrow, James. "Some Statistics Regarding the Reliability and Validity of Student Ratings of Teachers." *Research Quarterly* 48(2): 372-75; 1977.

National Education Association. *The Evaluation of Classroom Teachers*. Washington, D.C.: NEA Research Report, 1964-R14.

Natriello, Gary. *A Summary of Recent Literature on the Evaluation of Principals, Teachers and Students*. ERIC # ED 141 407. 1977.

Neely, John. In: Balzer, Levon, editor, *op. cit.*

Oldham, Neil. *Evaluating Teachers for Professional Growth*. ERIC # ED 091 846. 1974.

Painter, John, and Kent Granzin. "A New Explanation for Students' Course Evaluation Tendencies." *American Educational Research Journal* 10(2): 115-24; 1973.

Peck, Robert, and Gary Borich. *Personality Measures That Predict Teaching Performance*. ERIC # ED 093 946.

Popham, James. "Alternative Teacher Assessment Strategies." In: Andrews, editor, *op. cit.* (See: McDonald)

Reavis, Charles. "Clinical Supervision: A Review of Research." *Educational Leadership* 35(7): 580-84; 1978.

Robinson, John. "The Observation Report—A Help or a Nuisance?" *NASSP Bulletin* 62(416): 22-26; 1978.

Roper, Susan, Terence Deal, and Sanford Dornbusch. "Collegial Evaluation of Classroom Teaching: Does it Work?" *Educational Research Quarterly* 1(1): 56-66; 1976.

Rosenshine, Barak. "PBTE—Proceed With Caution." In: Andrews, editor, *op. cit.* (See: McDonald)

Shavelson, Richard, and Nancy Dempsey. "Generalizability of Measures of Teacher Effectiveness and Teaching Process." ERIC # ED 150 018. Far West Laboratory for Educational Research and Development, 1975.

Smith, Janice, and T. J. Brown. *Secondary Students Evaluations of Their Courses and Teachers and Attitude Toward Schools*. ERIC # ED 129 867. 1976.

Soar, Robert, and Ruth Soar. "Problems in Using Pupil Outcomes for Teacher Evaluation." ERIC # ED 150 187. Washington, D.C.: National Education Association, 1975.

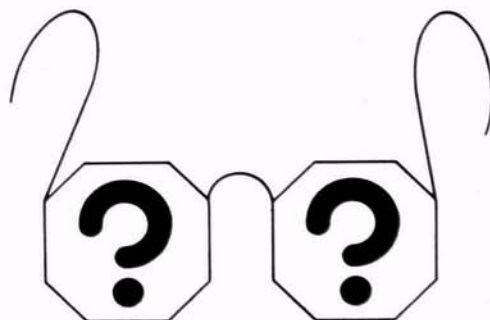
Swartz, Ned. *Divergent Perceptions of Teaching Effectiveness by Different Groups of Raters*. ERIC # ED 104 959. 1975.

Tuckman, Bruce. *Teacher Behaviour is in the Eye of the Beholder: The Perceptions of Principals*. ERIC # ED 137 928. 1977.

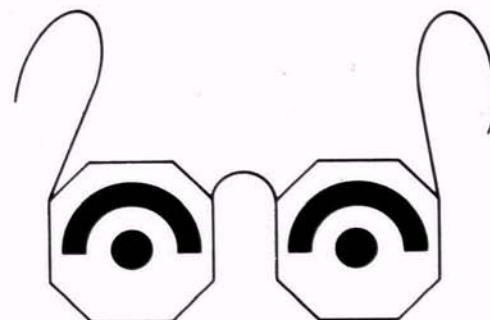
Wilcox, Ray. *A Comparison of Secondary School Teachers Judged Most Effective by Principals, Current Students, and Graduates*. ERIC # ED 139 103. 1976.

Wolf, Robert. "How Teachers Feel Toward Evaluation." In: Ernest Haus, editor. *School Evaluation*. Berkeley, California: McCutchan, 1973.

Wood, Peter. *Teacher Evaluation: Feedback to Improve Teaching*. ERIC # ED 139 733. 1976.



The new revised PRI Reading Systems will answer all of your questions about the management of reading instruction.



You'll be surprised how excellent PRI Reading Systems are . . . coming in Summer 1980.

CTB
The Testing Company

CTB/McGraw-Hill, Monterey, CA 93940

Copyright © 1979 by the Association for Supervision and Curriculum Development. All rights reserved.