Research Information Service

JAMES RATHS AND HALLIE PRESKILL

Research Synthesis on Summative Evaluation of Teaching

If evaluation is taken to mean assessing the worth or value of a phenomenon, it is surprising to note how few of the authors who have written on the evaluation of teaching are actually concerned with "evaluation."

In preparing this review, we have studied more than 100 articles, books, essays, and research reports dealing with the evaluation of teaching. The activities characterized as evaluation in this literature were akin to coaching, teaching, supervising, counseling, or helping teachers become better teachers. In fact, the consensus of the field, summarized by McGreal (1982), is that effective practice in evaluation calls for reducing the judgmental components of the process for optimal impact on teaching improvement. While these efforts are salutary, do they represent what is commonly understood to be "evaluation"? Apparently, evaluation is considered valuable only if it works to improve (but not judge) instruction.

While summative evaluations may not be helpful to teachers as a way of improving instruction, they may have other useful purposes, such as those suggested by Bolton (1973): validating the schools' teacher selection process; rewarding superior performance; protecting students from incompetence; and supplying information that will lead to the modification of teachers' assignments, such as placement into other positions, promotions, and terminations. This review is intended to summarize the salient ideas found in the teacher evaluation literature as it pertains to making summative evaluations of teaching.

James Raths and Hallie Preskill are staff members of the Center for Instructional Research and Curriculum Evaluation at the University of Illinois, Champaign.

Ideas to Guide Considerations of Summative Evaluation

Our review of the literature suggests five important ideas about teacher evaluation that may serve to illuminate the recommendations we suggest at the close of this article.

Summative vs. Formative Evaluation. Summative evaluation is not a mere extension of formative evaluation (Scriven, 1967; Bloom and others, 1971; Bloom, 1980; Scriven, 1981a). The former is a bottom-line judgment concerning, in this case, the quality of teaching under review. A summative outcome is the grading label assigned to a teacher's performance, such as satisfactory, needs improvement, or outstanding. The summative assessment is not necessarily intended to be helpful to the teacher. The process is designed to contribute to other goals such as those cited above from Bolton (1973).

Formative evaluation is designed to provide the teacher with information, tips, advice, and suggestions that should serve to enhance the summative evaluation the teacher will ultimately receive. Formative evaluation is directed to the teacher, and is intended to be helpful. In this process, the teaching act is almost always dimensionalized by the evaluator or by the teacher. The data derived from observations focus on discrete acts of teaching such as questioning techniques, management skills, and discussion strategies.

It is not clear what the relationship is (if any) between the findings of a formative evaluation and the subsequent summative evaluation. For instance, assume that a teacher is rated "at risk" in a summative fashion and is provided with the services of a supervisor or a helping teacher with the goal of improving instruction. Formative procedures are instituted, using goal setting or clinical supervision models as their vehicles, and improvements are demonstrated as a result of these interventions. It is not clear how salient those improvements might be to the subsequent summative evaluation. As the saying goes, the whole is greater than the sum of its parts.

Criterion and Standards. To borrow heavily from the prestige that scientific approaches are accorded in our culture, procedures for evaluating teachers have made use of the notions of "criterion and standards." This language suggests a precision that is not often evident in the teacher evaluation effort. Nevertheless, it is helpful to consider the distinction between these terms to become aware of the difficulties teacher evaluators face. Consider the following definitions.

A criterion is a statement that describes or designates a variable of instruction or an attribute that is required to meet the criterion; a standard specifies the amount of the variable. For example, if height were a criterion for becoming a teacher, a standard might be set at 60 inches tall. If a candidate attained this standard, then he or she would be assessed as meeting the criterion. As is widely known, such explicit statements of criteria and standards are not available to teacher evaluators. Consider the widely accepted criterion "knowledge of subject matter." What is the standard the teacher has to meet on this important factor? Our judgments of teachers are, in the main, more subjective than objective. Decisions about a teacher's knowledge of subject matter are based on existential leaps from evidence that is likely to be unrepresentative and impressionistic to a conclusion that might have significant impacts on a teacher's career.

Styles of Perception. Research on

personnel evaluation suggests at least two different approaches marking the dispositions evaluators bring to their tasks (Ewing, 1977). The first style is characterized by attending carefully to the words and behaviors teachers use in the classroom. Evaluators ask questions about the purposes of the lesson, about the congruence between purposes and class activities, and about the evaluation techniques that were used in the lesson. Typically, the supervisor looks for patterns in teaching—using Flanders' (1970) analyses to assess the teacher's indirectness or the Joyce and Weil (1980) typology to classify the teacher's approach.

The second style appears to focus on indicators of intangibles. Evaluators using this approach are concerned with qualities such as warmth, caring, acceptance, cooperation, professionalism, and other important global attributes. While evaluators have difficulty defining these qualities or defending particular behaviors as being valid indicators of them, they are convinced that they "know them when they see them." In commenting on lessons, and in the absence of a search for patterns, supervisors in this style tend to cite behaviors that are related to the sought-after attributes of the great teacher. Thus, spelling errors and misstatements of fact are salient since they evidently pertain to "knowledge of subject matter." The availability of appropriate materials in the middle of a lesson is a concern since this factor relates to planning. Even a concern for the "window shades," long a rumored focus of supervisors, has relevance as it pertains to the teacher's commitment to orderliness and housekeeping details.

This approach is best illustrated by an anecdote. An experienced teacher had received the highest rating the principal could give, "outstanding," for five years straight. At the end of the sixth year, a new principal rated the teacher as only "excellent." When the teacher inquired as to the justification for the lower rating, the principal said: "John, you never seem to take any work home with you. Your arms are empty when you head for the parking lot each afternoon. A good teacher works harder at preparing his lesson by taking work home." John accepted the principal's judgment at face value. Each afternoon during the subsequent year, John took work home with him and pointedly reminded the principal that he was doing so. His evaluation at the end of the second year was still only "excellent." It seems evident that the principal made attributions to John based on his observations, namely that John was not fully committed to his teaching. Note that the teacher did not ask, as well he might, "What is there about my teaching that would suggest I am unprepared?" or "Are my lessons poorly planned?" or "Do I appear unready for my classes?" Such questions would have been asking for supervision more in the vein of the first style described above.

Discernment vs. Criticism. "Connoisseurship is the art of appreciation, whereas criticism is the art of disclosure" (Eisner, 1979). The connoisseur is a person who makes discernments reliably and with sagacity. The critic is able to communicate to others the bases of these discernments. The evaluator, as connoisseur, might be able to distinguish between an excellent teacher and a mediocre one. In addition, as a critic he or she is able to cite the factors that contributed to the judgment. A key ingredient in teacher evaluation, therefore, is not merely making discernments of who is or is not a good teacher. The important factor is being able to communicate the basis of that discernment to teachers, to

Highlights from Research on Summative Teacher Evaluations

Five dimensions of teacher evaluation are:

- **Summative vs. Formative Evaluation.** The purpose of the summative evaluation is to "grade" the quality of teaching. In contrast, formative evaluation provides the teacher with suggestions and information to improve teaching performance, which should lead to a higher rating on the summative evaluation.
- **Criterion and Standards.** The criterion is an objective statement that specifically identifies the aspect of teaching performance or characteristic to be evaluated. Most teacher evaluations are more subjective than objective.
- **Styles of Perceptions.** In general, evaluators look either at tangible factors of the teacher's classroom performance or for more intangible qualities. Perceptions of the first style are based on the teacher's behaviors, lesson planning, and techniques in the classroom. Perceptions of the second style are based less on definable patterns than on such qualities as cooperation, professionalism, and other general character traits.
- **Discernment vs. Criticism.** The discerning evaluator can identify which teachers are excellent and which are mediocre. The critical evaluator can identify the factors that make a teacher excellent or mediocre.
- **Arithmetic vs. Holistic Evaluation.** The arithmetic approach weighs each teaching dimension, totals the ratings, and arrives at an overall sum. This approach is useful when comparing teachers, but may not be justifiable for dismissing teachers. The holistic approach looks at the total effect of all teaching dimensions at one time. The two approaches are not necessarily exclusive.

Three methods can be used for making holistic judgments about teachers:

(1) **Paired-comparisons**—pairing teachers to be compared to one another in a process of elimination and selection to determine which teachers are most effective and which are least effective.

(2) **Intrinsic scoring**—rating each teacher as average, or 3, on a scale of 1 to 5 and then rearranging teachers symmetrically over the ratings until a spread is reached.

(3) **Performance-based procedures**—applying a five-point scale to individual, specifically defined dimensions of teacher performance. Teachers may receive different ratings on various dimensions. The result provides a defined basis on which to make summative evaluations.

school boards, and pessimistically speaking, to the courts.

Arithmetic vs. Holistic Approaches to Summative Evaluation. The arithmetic approach for arriving at summative evaluations makes use of a "weight and sum methodology" (Scriven, 1981b). This process entails listing all the relevant dimensions, weighting them in terms of their importance, rating teachers on each dimension, and then summing the products of the weights and ratings to arrive at a total score. Scriven argues that such an approach "can greatly improve on naive judgment" (p. 86). Some teachers use an analogous procedure in computing final grades. A final examination is prepared with each item on the examination given a weight of two points. A total score of 92 on the test is deemed worthy of an A while a total score of 88 is judged to merit a B grade.

Often arithmetic procedures such as this are used in research designs when hypotheses deal with comparisons of good or poor teachers. For instance, regression lines are computed based on the scholastic aptitude of classes (based on mean values) and the achievement performances of those classes. Teachers whose classes are above the regression line are termed "effective" and those below the regression line are classified as "ineffective." It is probably difficult to defend this process as a basis for dismissing teachers because of the arbitrary nature of the cut-off scores. However, similar approaches have been used to identify winners for "teacher-of-the-year" awards, an enterprise with more benign outcomes.

Another approach frequently used for evaluating teaching involves the judgments of individuals or groups of individuals that are rendered more or less holistically. In this approach, evaluators take into account the evidence concerning the teaching of a faculty member: students' comments, artifacts of teaching, and even observations of the teacher's performance in class. If the evaluation is to be made by more than one person, the group of evaluators may meet to hold discussions and share their judgments concerning strengths and weaknesses they have perceived. Teachers are evaluated on the basis of the total effect the information has made on the evaluators. The various dimensions of good teaching that might be used to translate the problem into one of "arithmetic" are not judged separately. Instead, all the criteria of good teaching that might be treated in an arithmetic fashion are assessed at once. The data that are used in defending the summative judgment are the "votes" of the group of evaluators.

It is likely that evaluators who use an arithmetic approach to arrive at a summative evaluation also render holistic judgments, at least in their own minds. It is a matter of human nature for the evaluator to make overall judgments based on his or her experiences with the teacher and the impressions made by information collected for the evaluation. The purpose of the arithmetic approach is to suppress and minimize these impressions, giving the appearance at least of being objective. Thus, many evaluators end up being "surprised" after implementing the arithmetic approach. A teacher thought to be superb on a holistic basis is found to rate in the middle of a distribution after applying an arithmetic procedure; or, conversely, a teacher judged to be mediocre based on intuitive bases is found to have the top score. The irony is that in terms of empirical studies, the holistic judgments rendered about teaching seem to have more predictive validity than do the arithmetic ones (Centra, 1980).

Dilemmas arise when the results of an arithmetic evaluation differ markedly from the holistic judgments of the evaluator. Which should dominate the decision? The dilemma is heightened with the knowledge that holistic judgments are often influenced by factors such as latent racism, sexism, and other biases that are onerous to all evaluators.

Aids to Holistic Judgments

Regardless of the approach, teacher evaluation is inevitably an exercise in rendering holistic judgments. The arithmetic approach entails making such judgments on a criterion-to-criterion basis while the holistic procedure calls for "overall" judgments. The following methods are useful in harnessing the holistic judgments of individuals or groups of individuals.

1. *Paired-Comparisons.* All the teachers in a department and school might be entered on a ballot—each name paired with all others. After studying all the evidence and completing all the required observations, the evaluator(s) consider the teachers in pairs and ask the question: "Of these two teachers, which is better?" The reliability of judgments is optimal in making comparisons between two teachers rather than trying to apply an abstract scale to teachers one at a time (Torgerson, 1958). A distribution is created by counting the number of times each teacher is selected.

Of course, if the number of teachers in the pool is large, the number of pairs would become too large to use this procedure. Accommodations can be made. For instance, paired comparisons could be implemented using subgroups of faculty—grade levels or departmental designations could be used to reduce the size of the groups. Runoffs between winners of the subgroupings could be carried out using the paired-comparison technique a second time. This procedure would not designate what the appropriate decision should be for an individual teacher. It would identify on a comparison basis which teachers are deemed superior and which are adjudged least effective. Holistic decisions would still have to be made concerning whether those receiving the lowest scores should be released, given inservice training, or reassigned to positions where they may perform more effectively.

2. *Intrinsic Sorting.* Huntley (1976) proposed a system for assigning grades to students in classroom settings that seems to have some applicability for making summative evaluations of teachers. His procedure involves making overall judgments of teachers on a comparative basis by placing all the names of those to be evaluated into a middle column of a five-column array. The columns are labeled as follows: (1) Column 1: Lower still; (2) Column 2: Slightly lower than the group norm; (3) Column 3: Performance of these teachers consonant with the expectations of the school and community; (4) Column 4: Slightly higher than the group norm; (5) Column 5: Higher still.

Evaluators first decide which teachers should be assigned to columns two and four from column three. Huntley suggests that a "symmetry rule" be imposed, and the number of names moved right or left not differ by more than one. Next, those assigned to column four may be reviewed, and those that seem to merit assignment to column five are determined. The same process is used to move persons from column two to column one. Again, nothing in this procedure requires that teachers assigned to category five receive a pay raise or that those assigned to column one should be terminated. It is a process for identifying teachers who are strong or weak in a summative sense using the holistic judgments of peers, supervisors, or principals.

3. *Performance-Based Procedures.* This process shares with Huntley's the premise that the average of the group deserves to be characterized as "average." It attempts to be more graphic about the qualities being rated and could be used as the basis of an arithmetic approach to the summative evaluation of teaching by applying it dimension by dimension; or it might be used as a process of harnessing the overall holistic evaluation of teachers. In this process, a five-point scale is commonly used. The third step of the scale describes in narrative the expected level of performance. The fourth level describes a better performance and the fifth level describes the best performance. In similar fashion, the second level depicts less than the expected performance and the first level describes the most unfavorable outcome thought likely. The narrative accompanying this procedure clearly states what is being surveyed. For instance:

1. The student teacher was not able to handle the class on his own. The cooperating teacher did not feel he was ready for the responsibility and only permitted him to teach one or two brief mini-lessons during the student teaching period.

2. The student teacher taught on his own for several periods of time but the cooperating teacher always stayed close or remained in the room to monitor his behavior and to be available in case he needed to be "rescued."

3. The student teacher taught on his own for the period of time specified in the student teaching agreements. His work was satisfactory and was seen as creditable by the cooperating teacher. On occasion, the cooperating teacher felt sufficiently confident to leave the room and even the school building while the student teacher was teaching.

4. The student teacher taught the class on his own for most of the semester. The cooperating teacher felt free to leave the student teacher to his own devices.

5. The student teacher taught the class on his own for most of the semester. The cooperating teacher felt that on a number of occasions the student teacher performed brilliantly—"often better than I usually do."

The problem in this case would be to decide which classification defined by the narrative best describes the student teacher's performance. Again, this procedure attempts to communicate the explicit basis of the judgment given of the teacher's performance. Performance-based scales such as these are used in the Teacher Assessment Project sponsored by the University of Georgia. Researchers there report astoundingly high coefficients of agreement in using these scales on 16 teaching skills ranging from "uses responses and questions from learners in teaching" to "shares and seeks professional materials and ideas" (Capie and others, 1979). A further advantage to the Georgia approach is to reduce the scale on which holistic judgments are made—from overall teaching to separate, fairly well-defined dimensions (Teacher Assessment Project, 1980).

Summary

This article is based on several important assumptions. First, summative evaluations are probably not much help to teachers in improving instruction, although they do serve other important functions in schools. Second, since teachers are evaluated mainly on the basis of criteria without standards, the process is by and large an impressionistic one requiring supervisors and principals to render holistic judgments. To make judgments such as these requires experience, good sense, and courage. The latter may well be fortified by involving a number of people in the decision process—using "triangulation" procedures to reach agreements and to protect teachers from capricious judgments. In the end, evaluators of teaching are called upon to prize the doubt *and* act on their best judgments. It is a difficult task perhaps made easier by some of the procedures we have suggested. ■

References

Bloom, Benjamin S.; Hastings, J. Thomas; Madaus, George F. *Handbook on Formative and Summative Evaluation of Student Learning.* New York: McGraw-Hill Book Co., 1971.

Bloom, Benjamin S. "The New Direction in Educational Research: Alterable Variables." *Phi Delta Kappan* 61 (February 1980): 382-385.

Bolton, Dale L. *Selection and Evaluation of Teachers.* Berkeley, Calif.: McCutchan, 1973.

Capie, W.; Ellett, C. D.; and Johnson, C. E. "Selected Investigations of the Reliability of the TPAI," Technical Report RPM 79-4. Athens, Ga.: University of Georgia, 1979.

Centra, J. A. *Determining Faculty Effectiveness.* San Francisco: Jossey-Bass, 1980.

Eisner, Elliot. *The Educational Imagination.* New York: Macmillan, 1979.

Ewing, David. "Discovering Your Problem Solving Style." *Psychology Today* 11 (1977): 69-73.

Flanders, N. A. *Analyzing Teaching Behavior.* Redding, Mass.: Addison-Wesley, 1970.

Huntley, John F. "Academic Evaluation and Grading: An Analysis and Some Proposals." *Harvard Educational Review* 46 (November 1976): 612-631.

Joyce, Bruce, and Weil, Marsha. *Models of Teaching.* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

McGreal, Thomas L. "Effective Teacher Evaluation Systems." *Educational Leadership* 39 (January, 1982): 303-305.

Scriven, M. "The Methodology of Evaluation." In *Perspectives of Curriculum Evaluation,* No. 1 in the American Educational Research Association Monograph series on Curriculum Evaluation. Chicago: Rand McNally, 1967.

Scriven, M. "Summative Teacher Evaluation." In *Handbook of Teacher Evaluation.* Edited by Jason Millman. Beverly Hills, Calif.: Sage Publishing, 1981a.

Scriven, M. "The 'Weight and Sum' Methodolog,." *Evaluation News* 2 (February 1981b): 85-90.

Stake, Robert E. "Objectives, Priorities, and Other Judgment Data." *Review of Educational Research* 40 (April 1970): 181-212.

Teacher Assessment Project. "Teacher Performance Assessment Instruments: Their Uses and Limitations." Athens: University of Georgia, 1980.

Torgerson, Warren S. *Theory and Methods of Scaling.* New York: John Wiley & Sons, Inc., 1958.