R E P O R T R E S U M E S

ED 019 724 EA 001 277

DEVELOPING A COUNTY PROGRAM FOR EVALUATING TEACHING IN ELEMENTARY AND SECONDARY SCHOOLS.

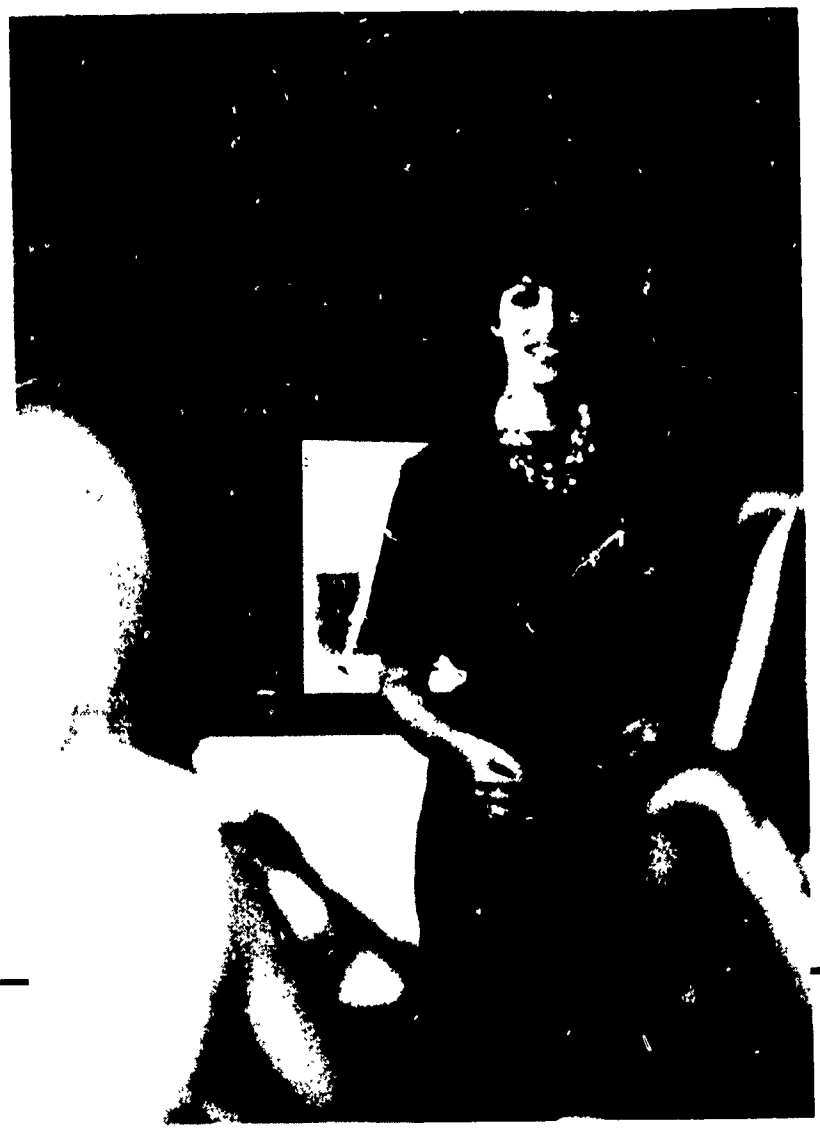
BY- CHRISTIAN, FLOYD T.
FLORIDA ST. DEPT. OF EDUCATION, TALLAHASSEE

PUB DATE SEP 67

EDRS PRICE MF-$0.25 HC-$0.64 14P.

DESCRIPTORS- *COUNTY SCHOOL SYSTEMS, *PROGRAM DEVELOPMENT, ELEMENTARY SCHOOLS, SECONDARY SCHOOLS, *TEACHER BEHAVIOR, TEACHER RATING, STUDENT TESTING, *TEACHER EVALUATION, INSTRUCTIONAL IMPROVEMENT, *EVALUATION CRITERIA, EFFECTIVE TEACHING, TALLAHASSEE,

TEACHING IS CONSTANTLY BEING EVALUATED BY TEACHERS, PUPILS, ADMINISTRATORS, AND PARENTS. NOT ALL PERSONS WHO EVALUATE TEACHING ARE LOOKING FOR THE SAME THING. IF THE PROCESS OF EVALUATION IS TO LEAD TO IMPROVED INSTRUCTION, THE TEACHER WHOSE WORK IS BEING EVALUATED MUST COMPREHEND THE FRAME OF REFERENCE FROM WHICH THE EVALUATION IS INSTITUTED. A COUNTY EVALUATION PROGRAM MUST FIRST DEVELOP A SET OF GENERAL POLICIES. FOUR GUIDING PRINCIPLES CAN HELP THE EVALUATION PROGRAM COMMITTEE TO AVOID MUCH OF THE CONFUSION ABOUT TEACHER EVALUATION--(1) CRITERIA AND EVIDENCE ARE THE TWO ELEMENTS ESSENTIAL FOR EVALUATION, (2) CRITERIA FOR USE IN EVALUATING TEACHING ARE THE PRODUCT OF A VALUE JUDGMENT WHICH CANNOT BE OBJECTIVELY VALIDATED, (3) THE NATURE OF THE EVIDENCE REQUIRED FOR EVALUATING TEACHING IS DICTATED BY THE CRITERIA SELECTED, AND (4) A SOUND EVALUATION PROGRAM SHOULD PROVIDE INFORMATION ON TEACHING WHICH IS RELEVANT, RELIABLE, AND INTERPRETABLE. THE PROCESS OF COLLECTING EVIDENCE NEEDS TO BE SEPARATED FROM THE PROCESS OF COMPARING IT WITH CRITERIA BECAUSE THERE IS A PROBLEM IN DETERMINING WHETHER DISAGREEMENTS IN EVALUATIVE JUDGMENTS RESULT FROM DISAGREEMENTS OR FROM DIFFERENCES IN THE EVIDENCE SELECTED AND BECAUSE INFORMATION ON SPECIFIC BEHAVIOR IS QUITE EFFECTIVE IN HELPING TEACHERS TO MODIFY THAT BEHAVIOR. (HM)



SEPTEMBER, 1967

DEVELOPING A COUNTY PROGRAM FOR EVALUATING TEACHING

IN ELEMENTARY & SECONDARY SCHOOLS

STATE DEPARTMENT OF EDUCATION
TALLAHASSEE, FLORIDA
FLOYD T. CHRISTIAN, State Superintendent

CONTENTS

SUGGESTED PROCEDURES
for developing a county program for evaluating teaching

Some people say that teaching cannot be evaluated. Others say that teaching can be and should be evaluated but maintain that such evaluation is not taking place.

It seems apparent, however, that teaching is continually being evaluated by teachers, by pupils, by administrators, and by parents. A teacher, after teaching a particular lesson once, makes his decisions regarding how to teach the lesson the next time on the basis of his evaluation of the first lesson. A high school pupil advises his friend to elect or avoid classes taught by a particular teacher on the basis of his evaluation of the teaching done by that teacher. A principal recommends a teacher for continuing contract on the basis of his evaluation of the teaching done by that teacher. A parent who selects or rejects a certain neighborhood because of good or poor schools bases his choice upon his evaluation of the teaching which takes place in those schools.

The plain fact is that evaluations of teaching constitute the primary basis for making virtually all decisions about schools and about teachers. This is manifest. Teaching is the mission for which teachers and schools were invented. Teaching is constantly evaluated.

WHY DOES THE EVALUATION OF TEACHING APPEAR TO BE A PROBLEM?

The problem is not that teaching cannot be evaluated or is not being evaluated. The basic difficulty stems from the fact that all persons who evaluate teaching are not looking for the same thing. The teacher may be looking for certain specific verbal or written responses from pupils. The pupils may be satisfied if the teaching holds their interest and may be dissatisfied if it does not (regardless of what they may or may not have learned). The principal may feel that if the classroom is orderly, the teacher is poised, and the pupils are attentive, the teaching is effective. The parents may evaluate the teaching favorably if a substantial portion of the pupils score above the seventy-fifth percentile on standardized tests. Consequently, the teaching rated as superior by one evaluator may be viewed by another evaluator as only average.

The evaluation of teaching has been recognized as a problem because different evaluators view teaching with different frames of reference.

HOW CAN A COUNTY BEGIN?

If the process of evaluating teaching is to lead to improved instruction, the teacher whose work is being evaluated must comprehend the frame of reference from which the evaluation is instituted. Moreover, there should be assurance that the particular frame of reference is justified in terms of the educational objectives for the specific grade, subject, and type of pupils being taught by that teacher.

The first thing needed in developing a county evaluation program is a set of general policies. It is suggested that a broadly based committee be appointed for the purpose of developing such policies. This committee would be composed of teachers, principals, supervisors and other personnel. It would be responsible for overseeing the development of the county program for evaluating teaching.

This committee would not be able to devise the complete program, however. It would not be presumed that this broadly based committee would possess the competencies necessary to develop sets of criteria and procedures appropriate for evaluating teaching within the many different curricular areas and levels included in the school program. Special committees would be required. These committees would be composed of teachers and supervisors within the specific area involved, with the possible addition of one or more principals or other "generalists." The special committees would, in effect, be subcommittees of the general committee.

ACTIVITIES OF BROADLY BASED COUNTY-WIDE COMMITTEE

1. Prepare a statement of purposes which the evaluation program is intended to fulfill. Such a base to work from is essential for the committees which will develop the criteria and procedures. For example, if the evaluation program is intended to help teachers to improve their instruction it would be necessary for committees to provide for conferences or other means through which the teachers can obtain the necessary feedback.

2. Subdivide teaching assignments within the county by subject and/or by grade and/or by characteristics of pupils served. It might be that subdivisions would be along grade lines of the elementary level and along subject lines at the secondary level.

3. Appoint special committees for each of the subdivisions. Set forth such additional guidelines as are appropriate for keeping committees on the right track. Set up a timetable including the items listed below. A period of 4-8 weeks should be allowed for each of these steps.

   a. Deadline for completion of the statement of the general type of criterion to be employed (viz., whether it will be based upon teaching processes, teaching products, or a combination of the two). This statement should also include an explanation of the rationale for selecting the particular type of criterion.

   b. Deadline for completion of an explicit statement of criteria to be applied and procedures to be followed in applying them.

   c. Deadline for completion and evaluation of a trial application of the evaluation program.

   d. Deadline for presentation of a "validated" evaluation program for transmission to the county superintendent.

4. In general, the committee should keep informed of activities of each of the special committees.

ACTIVITIES OF THE SPECIAL COMMITTEES FOR DEVELOPING EVALUATION CRITERIA AND PROCEDURES

The deadlines discussed above comprise an outline of the activities of the special committees. The basic role of each special committee is to use whatever resources it has available to devise an evaluation program for a special area or grade which will fulfill the purposes established by the general committee and which will also meet the criteria for evaluating evaluation programs which are discussed later in this paper. The committees might wish to make use of consultants from colleges and universities or from other school systems.

It is proposed above that each evaluation program be "validated." This means that the evaluation procedures should be tested in as many situations as possible before they are submitted to the county superintendent. It is likely that a valid indication of the relevance,

reliability, interpretability and equity of a program can be obtained only through field testing. A "validated" evaluation program is one which appears to meet those criteria.

IS A "VALIDATED" EVALUATION PROGRAM SUBJECT TO REVISION?

It is unlikely that even the persons who devise a set of evaluation procedures will be fully satisfied with them. The art of evaluating teaching is still in a rudimentary state. Thus, any evaluation program which will be developed will be subject to revision, no matter how carefully it has been "validated."

Teaching is the mission for which teachers and schools were invented.

. . . Teaching is constantly evaluated.

2 GUIDING PRINCIPLES

A failure to answer fundamental questions seems to be one probable cause for much of the confusion which exists concerning the evaluation of teaching.

If evaluators could agree on some basic premises, it is much more likely that they could reach some agreement on their evaluations. Basic questions such as those discussed below must be answered clearly and satisfactorily.

It should be noted at the outset that research on teaching has been of very little help in providing principles for guidance in evaluating teaching. Many studies have been conducted which attempt to isolate the factors which make an individual teacher effective or ineffective.[1] Unfortunately, however, these studies have yielded little knowledge with practical applicability. This fact has been attested by several writers. Remmers concludes that reports of research on teaching contain little information "that a superintendent of schools can safely employ in hiring a teacher or granting him tenure, that an agency can employ in certifying teachers, or that a teacher

education faculty can employ in planning or improving teacher education programs."[2] Turner and Fattu state that, "Seventy years of research on teacher effectiveness have not added much to our systematic knowledge, and it is difficult to see how another seventy can do any more if the same procedures are followed."[3] Other writers expressing dissatisfaction with research results and methodology include Barr and Jones,[4] Mitzel,[5] and Ryans.[6]

Thus, it can be safely affirmed that past research provides little evidence as to what constitutes effective teaching. Because of this, individual school systems must rely mainly upon the best practical judgment which they have available in order to devise a framework for evaluating teaching. Teachers, supervisors, and administrators, with possible assistance from consultants from colleges and universities or from other school systems, are the ones who must answer the basic questions discussed in the subsequent paragraphs.

WHAT IS EVALUATION?

Evaluation is the act of assigning a value. It is what the meat inspector does when he grades meat. It is what the used car dealer does when he sets a price on a car. It is what the teacher does when he assigns marks to pupils. It is what the music critic does when he writes his review. Likewise, it is what is done when a pupil, teacher, administrator, or parent makes a judgment about the quality of teaching in a given classroom.

In all of the above cases there are two ingredients which are essential if evaluation is to take place. The first is a set of criteria or standards against which the thing being evaluated can be measured. In each of the above examples the evaluator has in his mind a set of criteria or standards which serve as a model to which he relates his observations. If two persons evaluating the same thing do not agree on their evaluation, it could be that they are using different criteria. Stated differently, when two persons evaluating the same thing do not agree, they may be using a different model. (The terms "criteria," "standards," and "model" are intended to connote basically the same concept).

The second essential ingredient for evaluation is evidence. The information which the evaluator relates to the model constitutes evidence. The meat inspector might study the color of the meat and the distribution of fat. The used car dealer's evidence will include such things as engine condition, make and model of the automobile, odometer reading, and the age of the vehicle. The teacher will consider such things as daily work, test results, and special projects. The music critic will consider such things as intonation, phrasing, technique, and pacing. When the pupil, teacher, administrator, and parent evaluate teaching, they will collect whatever evidence they deem relevant to their criteria. If their criteria were the same, their evidence should relate to the same aspects of teaching.

Disagreements on interpretations of evidence can be resolved by collecting additional data—presuming, of course, that prior agreement has been reached on the criteria or standards. When agreement is not reached as to which criteria are to be considered relevant, any subsequent agreements on the matter of evaluation are purely coincidental. This includes agreements on the nature of the evidence which should be collected, agreements on the interpretation of the evidence which is collected, and agreements on the final evaluation.

The first guiding principle can be summarized as follows:

> *Principle I: Criteria and evidence are the two elements which are essential in order for evaluation to be possible. Evaluation takes place when evidence is compared with selected criteria (i.e., the model). Unless agreement can be reached as to which criteria should be applied, any agreement on the final evaluation is purely coincidental.*

WHAT IS GOOD TEACHING?

How can good teaching be recognized? Or, more precisely, what constitutes acceptable criteria for evaluating teaching? Are some criteria more worthwhile than others?

To answer these questions it is necessary to make one or more value judgments. Teaching which is effective does not exist independently

but is an artifact created when an independent or collective value judgment is made. Robinowitz and Travers assert that, "No teacher is more effective than another except as someone so decides and designates . . . . The ultimate definition of the effective teacher does not involve discovery but decree."[7] Ryans[8] agrees that no type of criterion of effective teaching possesses intrinsic goodness. He states that the worthiness of any given set of criteria is dictated by the values of the specific culture which the teaching is intended to serve. The wisdom of judgments on criteria of effective teaching is certain to be enhanced if the influence of knowledge and experience are brought to bear.

The people who are best qualified to make judgments as to what constitutes good teaching in any given situation are those with (1) knowledge of the objectives which the teaching is supposed to fulfill, (2) knowledge of the situation in which the teaching will take place, and (3) knowledge of ways in which the teaching objectives can be accomplished.[9] Criteria which are established without consideration of the realities of teaching objectives, teaching situations, and teaching methods are likely to be capricious.

The place to begin in developing criteria for evaluating teaching is with the goals which the teaching is expected to accomplish. The teaching which contributes to the attainment of these goals is considered effective. It is the job of those assigned the responsibility for developing criteria to determine what type of teacher behavior is most likely to achieve these goals and/or to determine what kind of pupil behavior might constitute a valid index of the contributions of the teacher to these goals. Because specific goals vary between subjects, between educational levels, and with different types of student populations, it is probable that several different sets of criteria will be required. The number of sets of criteria to develop and the composition of the groups to which they will apply must be decided before groups are assigned to prepare statements of criteria for possible adoption.

The second guiding principle is stated as follows:

Principle II: Criteria for use in evaluating teaching are the product of a value judgment which cannot be objectively validated. However, once the objectives of the teaching program have been identified, the persons with the greatest familiarity with the situation in which the teaching is to take place are the ones who are best qualified to define the criteria.

WHAT KIND OF EVIDENCE SHOULD BE COLLECTED?

The type of evidence collected is dictated by the criteria which have been established. It is unreasonable to set out to determine specifically what kinds of evidence should be collected until after the criteria have been defined. On the other hand, it is unwise to devise criteria without having in mind the general type of evidence which would be required to employ them.

The types of evidence which have been collected in the past can be placed into three classifications. The first type consists of traits possessed by the teacher such as "amount of education," "honesty," or "pleasant personality." These can be called status variables as they describe characteristics of the status of the teacher. It is possible to collect this type of evidence without ever seeing a teacher in a teaching situation. It is worth noting (although it may seem facetious) that through the use of status traits, it would be possible to evaluate the "teaching" of someone who has never taught. While status characteristics may be of some value as predictors, there is no reason to use them when evidence obtained during or following an actual teaching situation can be obtained. Thus, criteria which imply evidence based upon status should be seriously questioned.

The second classification includes those things which occur during teaching. Examples include "asks open-ended questions," "arrests pupil attention without relying on authority," or "states assignments clearly." These are called process variables. Evidence of this type must be collected in a teaching situation.

The third classification includes those things which occur following teaching and presum-

ably as a result of teaching. Examples include "can spell correctly all words in the lesson," "can explain clearly the purposes and organization of the Federal Reserve System," or "can write a poem." These are called product variables. Evidence of this type can be collected both during and after the teaching situation.

Both process and product variables are appropriate for evaluating teaching. Many researchers have declared that criteria calling for product measures—namely, changes in the behavior of pupils—constitute the ultimate criteria of teacher effectiveness. On the other hand, many persons maintain that factors other than the influence of the teacher contribute significantly to changes in pupil behavior. Thus, they feel it is not possible to evaluate the work of a teacher solely in terms of the achievement of his pupils. In the domain of the local administrator or other instructional leader (as contrasted with the domain of the researcher),[10] criteria calling for process evidence are of particular significance.

The third guiding principle summarizes the above ideas.

Principle III: The nature of the evidence required for evaluating teaching is dictated by the criteria selected. The evidence used can relate either to the process of teaching or the product (results) of teaching. Status characteristics of teachers (which can be measured without observing the teaching or the results of the teaching) do not constitute appropriate evidence for evaluating teaching.

HOW CAN EVALUATION PROGRAMS BE EVALUATED?

What are the characteristics of a sound evaluation program? How can a good program be distinguished from a poor one?

There are four factors which should be considered in evaluating an evaluation program. The first is the relevance of the criteria and the evidence. Relevance refers to the extent of the relationship existing between the criteria for evaluating teaching and the goals of the educational program. If these goals formed the initial basis for developing the criteria, the criteria should be relevant. If the criteria evolved in some other manner, the matter of relevance should be studied very closely.

The second factor is interpretability. Interpretability refers to conditions which allow the evidence collected to be organized and analyzed in ways which will yield information that can be used for desired purposes. Section 231.29(2) of the Florida Statutes states that evaluation will be conducted "for the purpose of improving the quality of instruction, administrative and supervisory services." An evaluation program will be likely to yield interpretable data if the persons who are to use the information from the evaluation program — namely, the teachers and administrators — are involved in its development and understand clearly the purposes for which the information will be used. Because of variations between teaching objectives and teaching methods at different levels of instruction, it is probable that several different sets of criteria will have to be developed to insure interpretability.

The third factor is reliability. In this case, reliability refers to the consistency between evidence collected and behavior observed. If two evaluators observe the same teaching situation, their observations and their evaluations should display a high level of agreement. This, of course, is much less of a problem if all evaluations are to be made by one individual. The problem then is only for him to be consistent from one observation to the next. Multiple observers compound this problem and it is usually necessary to conduct several training sessions to obtain reliable results.

The fourth factor is equity: the evaluation program must be equitable. The criteria must not discriminate against a person with one particular teaching style unless it is agreed that his style is one which is not appropriate for accomplishing the objectives of the educational program. Normally, the problem of equity can be handled by providing for diverse representation within the group which develops the criteria.

The fourth guiding principle, which relates to the evaluation of evaluation programs, is as follows:

Principle IV: A sound evaluation program should provide information on teaching which is relevant, reliable, and interpretable. It should also be developed in a manner which will allow it to treat all persons whom it affects in an equitable manner. Beginning with educational goals and insisting upon the involvement of the people who are familiar with the situations in which the evaluation program will be applied and who will make use of the information should contribute to the attainment of these characteristics.

3 TECHNIQUES FOR COLLECTING EVIDENCE and comparing it with criteria

It has been pointed out that criteria of effective teaching must be accepted or rejected on the basis of value judgment.

While this judgment is more likely to be valid if careful consideration is given to the teaching situations in which the criteria will be applied, it remains that the judgment is to a certain extent a matter of personal preferences. Once the criteria have been determined, however, the development of procedures and techniques for collecting evidence is primarily a technical problem. Considerable work has been done in the development of observation and examination techniques. Hence, persons in county school systems who are responsible for prescribing the methods by which evidence will be gathered should become familiar with various techniques including rating, categorizing, and testing.

WHAT IS RATING?

Rating is a process whereby an observer collects and analyzes evidence and compares it with criteria without making any record of the evidence itself. In other words, the observer simply records his value judgment. Take as an example an evaluation criterion stipulating that it is desirable to give encouragement to students. If a rating technique were employed, the evaluator might simply indicate, using a five-point scale, that the teacher encouraged the pupils either "always," "much," "some," "little," or "never." If a system other than rating were used, the observer might record (in some type of "shorthand") instances in which the teacher gave encouragement.

Rating scales are by far the most widely used devices for evaluating teaching performance for both research and administrative or supervisory purposes. At least in the case of administrative and supervisory situations, this condition is likely to persist. The evidence which must be reviewed to determine whether or not teaching is effective is invariably extensive and subtle with numerous complexities which are difficult to catalog in advance.

In most cases, it has not been deemed practical to utilize objective procedures for reducing the vast amount of data. Thus, raters have been required to reduce the data to that which is significant and, in the same operation, to compare this evidence with the relevant criteria. When this occurs, summarizing and processing of evidence takes place entirely within the mind of the observer and only his conclusions are available for scrutiny. Hence, neither the data reduction process nor the evaluation process can be examined. If a case developed in which two "experts" evaluating the same teaching provided different evaluations, it would be a matter of speculation as to whether the discrepancy resulted from their selecting different evidence to process or from their applying different criteria in evaluating (unless, of course, the evaluators were available for questioning).

Fortunately, the evaluation process employing rating scales need not be so mercurial as the foregoing implies. The stability of results obtained with these scales can be controlled by controlling both the type and quantity of information to be processed and the processing itself. This can be done by providing sufficient descriptive material with the rating form to orient the user, by constructing a rating instrument composed of specific rather than general scales, and by constructing the individual scales carefully. Discussions of technical considerations in rating scale development and the literature relating to their use are presented by Guilford[11] and Remmers.[12]

An obvious technique for improving the reliability of ratings involves the training of raters. Such training could consist of a thorough orientation into the type of evidence which is to be considered significant and the type of criteria which are to be employed in analyzing it. This would be followed with practice in employing the scale including opportunities for comparing and discussing the ratings assigned. Practice sessions can be repeated until the desired level of reliability is reached.

CAN THE COLLECTING OF EVIDENCE AND THE PROCESS OF COMPARING IT WITH CRITERIA BE SEPARATED?

There are at least two reasons why consideration should be given to the possibility of separating the process of collecting evidence from the process of comparing it with criteria. The first relates to the problem of determining whether disagreements in evaluative judgments result from disagreements on criteria or from differences in the evidence selected. The second reason is that information on specific behavior is quite effective in helping teachers to modify that behavior. If the teacher can know what he did, as well as what the evaluator thought of what he did, he is in a much better position to modify that behavior (if modification is needed). A dramatic example of this is the effect of providing a teacher with a complete record of his teaching in the form of a videotape playback.

It was stated earlier that research has been of little help in identifying criteria for use in evaluating teaching. However, there are a number of writers who express optimism for the future of research on teaching. This optimism is based primarily upon recently adopted techniques for analyzing the dimensions of teaching and learning and for collecting data on classroom processes. Many of these observation techniques are also applicable for evaluating teaching in ongoing school programs.

The most widely used of the newer observation systems are the Observation Schedule and Record (OScAR)[13] by Medley and Mitzel and the interaction analysis system developed by Flanders.[14] Either of these systems might be used as examples for developing procedures for collecting evidence which is appropriate for the criteria selected.

The general procedure for developing a category system is to determine first which aspects of teacher or pupil performance are relevant (on the basis of the criteria adopted). The second step is to categorize those elements so that they can be objectively reported by an observer. An alternative, of course, is to locate a category system already in existence which can be adapted to the evaluation program.

Either of two different types of observation schedules can be used.[15] The first is called a category system. With this approach a list of relevant categories is devised. Normally these categories will relate to a specific dimension of behavior (such as verbal interaction). This list is presumed to be exhaustive from the standpoint that every unit of behavior which is witnessed by the observer can be placed in one of the categories. The completed observation record shows the total number of behavior units observed and the number classified in each category. The Flanders interaction analysis system is an example of the category type of observation schedule.

The second approach to constructing an observation schedule is called the sign system. With this system, a list of behaviors which may or may not occur is compiled. The observer then tallies those behavior units observed which meet the category definitions. It is not assumed that all behaviors which occur during the process of observation will be recorded. An example of a sign system is included in the *Teacher Practices Observation Record.*[16]

The category approach offers the advantage of accounting more thoroughly for behavior along a given dimension. To employ it, however, the number of categories must be limited so that the observer can keep them all in mind simultaneously and categorize observed behavior instantly. On the other hand, the sign system allows for a wider range of behaviors to be included. It does not, however, provide information as to the relative frequency of the behavior. Both systems are applicable to programs for the evaluation of teaching employing either process or product measures, provided of course that the relevant behaviors are defined and included in the list of categories used. The training of observers is necessary with sign and category observation systems just as it is with rating systems.

CAN PUPIL TEST SCORES BE USED TO EVALUATE TEACHING?

If product measures are to be used to evaluate teaching, the place of testing is obvious. Testing is a procedure which is used universally by teachers and administrators for assessing pupil learning. These educators, however, have been generally unwilling to use the results of such assessments as a basis for evaluating teaching. Their reasons might be summarized with two statements: (1) there are many factors which act before, during, and after a teacher's teaching which affect the amount of learning which takes place within any given individual; and (2) the tests which are available may not represent the full range of objectives toward which the educational program is directed.

The second criticism has been answered to a certain extent by the two volumes of the *Taxonomy of Educational Objectives.*[17] While the necessary evaluation instruments may not be immediately available, it appears that it would be possible to develop them if the specific objectives can be articulated.

The first objection raises some technical problems which are even more complex. Before these technical problems can be handled, however, a rationale must be developed which will serve as the basis for developing an equitable approach for using test scores. It does not seem reasonable to assume that the best teacher is the one whose pupils earn the highest scores; it could be that scores earned by some classes would be higher before the term begins than the scores which might be earned by other classes at the end of the term. There are also considerable problems in using gain scores (i.e., differences between pre-test and post-test scores earned by pupils).

It is difficult to say whether an average gain of ten points earned by a class whose initial scores were well below the mean is comparable to an average gain of ten points earned by a class whose initial scores were well above the mean. In order to cope with this problem, many modified approaches have been suggested for deriving measures of pupil gain.

If test scores are to be used to evaluate teaching, the basic need is to develop a rationale and a method which would provide an index of the amount of gain which each individual pupil would normally be expected to make in each of the areas deemed significant in a given class. It would then be possible to compare the observed gain with the expected gain. A teacher whose pupils gained more

than they would normally be expected to gain would be considered "above average." One whose pupils gained less than they would be expected to would be "below average." One whose pupils gained the amount which would normally be expected would be "average." It would be necessary, of course, to take numerous factors into consideration when calculating the expected gains for each pupil. These factors would include such variables as general aptitude, special aptitudes, prior knowledge, and motivation. It might be concluded that arriving at an objective estimate of the expected gains of each pupil is a problem even more complex than the problem of evaluating teaching. Nevertheless, this seems to be the only reasonable basis under which gain scores could be used as the primary basis for evaluating teaching.

If the teacher can know what he did, as well as what the evaluator thought of what he did, he is in a much better position to modify that behavior . . . if modification is needed.

FOOTNOTES

1 S. J. Domas and D. V. Tiedeman, "Teacher Competence: An Annotated Bibliography," *Journal of Experimental Education*, Vol. 19 (December, 1950), pp. 101-218; J. E. Morsh and E. W. Wilder, *Identifying the Effective Instructor: A Review of Quantitative Studies*, 1900-1952 (Research Bulletin No. AFPTRC-TR-55-44, San Antonio, Texas: United States Air Force Personnel and Training Center, 1954).

2 H. H. Remmers, *et al.*, "Second Report of the Committee on Criteria of Teacher Effectiveness," *Journal of Educational Research*, Vol. 46 (May, 1953), p. 657.

3 Richard L. Turner and Nicholas A. Fattu, "Skill in Teaching, Reappraisal of the Concepts and Strategies in Teacher Effectiveness Research," *Bulletin of the School of Education, Indiana University*, Vol. 36 (May, 1960), p. iii.

4 Arvil S. Barr and Robert E. Jones, "The Measurement and Prediction of Teacher Efficiency," *Review of Educational Research*, Vol. 28 (June, 1958), pp. 256-264.

5 Mitzel, "Teacher Effectiveness," *Encyclopedia of Educational Research*, ed. Chester W. Harris (New York: The Macmillan Company, 1960), pp. 1481-1486.

6 David G. Ryans, "Theory Development and the Study of Teacher Behavior," *Journal of Educational Psychology*, Vol. 47 (December, 1956), pp. 462-475.

7 William Rabinowitz and Robert M. W. Travers, "Problems of Defining and Assessing Teacher Effectiveness," *Educational Theory*, Vol. 3 (July, 1953), p. 212.

8 David G. Ryans, *Characteristics of Teachers: Their Description, Comparison and Appraisal* (Washington, D.C.: American Council on Education, 1960), p. 16.

9 This argument is developed in some detail by Cemal Yildirim in "An Analytic Model for Evaluation of Teacher Competence" (unpublished Ph.D. dissertation, Indiana University, Bloomington, 1963), pp. 95-102.

10 William J. Ellena, Margaret Stevenson, and Harold Webb (eds.), *Who's a Good Teacher?* (Washington, D.C.: American Association of School Administrators, 1961), pp. 5-6.

11 J. P. Guilford, "Rating Scales," in his *Psychometric Methods* (New York: McGraw-Hill Book Company, 1954), pp. 263-301.

12 H. H. Remmers, "Rating Methods in Research on Teaching," *Handbook of Research on Teaching*, ed. N. L. Gage (Chicago: Rand McNally and Company, 1963), pp. 329-378.

13 Donald M. Medley and Harold E. Mitzel, "A Technique for Measuring Classroom Behavior," *Journal of Educational Psychology*, Vol. 49 (April, 1958), pp. 86-92.

14 Edmund Amidon and Ned A. Flanders, *The Role of the Teachers in the Classroom: A Manual for Understanding and Improving Teachers' Classroom Behavior* (Minneapolis: Paul S. Amidon and Associates, 1963).

15 For a more complete discussion see Donald M. Medley and Harold E. Mitzel, "Measuring Classroom Behavior by Systematic Observation," *Handbook of Research on Teaching*, pp. 288-303.

16 Bob Burton Brown, *Teacher's Classroom Behavior* (Gainesville, Florida: Teacher Competence Research Project, College of Education, University of Florida, undated). (This is a group of instruments for use in evaluating a teacher.)

17 Benjamin S. Bloom (ed.), *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain* (New York: David McKay Company, Inc., 1956); David R. Krathwohl, *et al.*, *Taxonomy of Educational Objectives, Handbook II: Affective Domain* (New York: David McKay Company, Inc., 1964).

DEVELOPING A COUNTY PROGRAM FOR EVALUATING TEACHING

IN ELEMENTARY & SECONDARY SCHOOLS

STATE DEPARTMENT OF EDUCATION
TALLAHASSEE, FLORIDA
FLOYD T. CHRISTIAN, State Superintendent