are recognized as the earmark of any profession. To be adequate for this purpose the program would include such activities as these:

1. The school would set an example of the experimental point of view by presenting evidence as to the effectiveness of its own program, and also of the procedures the teacher is expected to use. Critical consideration would be given to the manner in which recommended procedures have been or can be tested and proved, as reported in the professional literature.

2. Competence in devising and testing new and more effective procedures would be accepted as an important outcome of the program. Teachers would be prepared to appraise the effectiveness of all untested procedures.

3. Finally, the program of preservice education would explicitly prepare the teacher for effective membership in his professional organizations. Such active membership is necessary not only to provide opportunities for the teacher to learn about, and to report, new developments in effective practices, but also as stimulus to continued activity and growth. It is this function of the professional association, rather than an economic one, that brought together the informal groups that eventually grew into the modern professional organization.

By way of summary, we recognize that the education of the teacher at the preservice level, at its best, must be limited. It is impossible to predict in advance the problems the teacher will encounter. Even if they were known, the time is not available for a program of preparation designed to meet the need. The alternative is to prepare a teacher who can deal adequately with existing problems, and who can develop whatever competence is required to meet future problems.

GALE ROSE

Toward the Evaluation of Teaching

Major methods of appraisal of teaching are analyzed here. Also considered are problems of objectivity and of values.

TO CONSIDER the evaluation of *teaching* is at once to focus on a process, a complex of acts, certain patterns of behavior, rather than on the person performing them or on the consequences of his behavior. The consequences deserve the most careful study, but they should be clearly distinguished from the behavior and circumstances which produced them. Likewise, the teacher personality, and other factors which interact with environment to produce certain teaching acts, may be thought of as causes or conditioners of performance and should be identified separately from the teaching itself.

These discriminations, which were pointed out in 1952 by the special AERA

GALE ROSE is secretary and research director, Utah School Merit Study Committee, Salt Lake City, Utah.

Committee, (1) seem to be essential before a scientific attack can be made on this problem. Failure to clarify these distinctions has resulted in the past in much research on the periphery of teaching without ever getting to the heart of the act itself.

Since teaching is a complex, on-going, emerging process, it is necessary, if one wishes to study it systematically, to learn how to make it hold still, to sit for its portrait, or more exactly, to succumb to measurement. But precedent to measurement must come definition and description. How can teaching be described so that it is susceptible of measurement? The most useful idea to appear in relation to this need has been the concept of role. When one begins to study teaching in terms of roles the teacher performs, whether toward known objectives or not, clarification of function becomes possible. Function-analysis can then develop within a relevant framework.

Methods of Appraisal

At this point it may be helpful to describe the major methods of appraisal which have been used by those who evaluate teachers or teaching. Appraisal types might be classified into three main categories in terms of the method used for the collection, arrangement, and reporting of data: (2)

1. RATINGS. One of the outcomes, or concomitants, of the general scientific movement in psychology in the early 1900's was the development and elaboration of methods by which judges can rate things and people. Two major types of ratings have been used:

a. *Direct Comparisons,* in which individual items, or people, are compared directly with other individuals, by

(1) *Rank Order*—a listing from one described extreme toward another (best to worst, for example).

(2) *Pairing*—each individual is paired and ranked with every other one, then all the rankings are handled statistically in order to arrive at values.

These methods have known deficiencies or difficulties and are no longer widely used for appraisals. The comparisons are personal rather than to a standard, steps between those rated are not equal, and each group must be considered as a universe of its own: ratings between groups are not comparable. In addition to that, the paired method is unwieldy, involving 190 comparisons for a group of only 20 persons. The methods of equal-appearing intervals and successive intervals have been developed to shortcut the paired comparison method.

b. *Scales.* These are expressions of linear relationship, in a number of steps or units out from a given point, or between two given points (excellent—unsatisfactory, for example). A scale may be a visual continuum with reference points along the way, or it may have fixed response points with no opportunity to mark in between them. A scale may be marked off in letter grades, number grades, by descriptive words or phrases, or even by names of persons or known objects which typify the various scale positions, or some combination of these. Scales may have any number of steps from 2 up, though 15 is about the maximum number ordinarily used; most have from 4 to 9 steps, with 5 probably the mode, and odd numbered ones predominating because of the normal curve concept. Scales may range between a zero point and a maximum, or they may range in two directions away from a zero point.

with opposite items at the extremes, or toward negative and positive sides of one dimension.

The number of items on which a person may be rated can be of any amount; for teaching there are usually from about 5 up to 50. These items can be of at least 3 major types:

(1) *Traits, Characteristics*, or *Qualities* of the person. These are usually words or phrases which describe some personality element, such as Cooperativeness, Potential, Efficiency, and Initiative, or physical aspects such as Voice and Appearance. This type sometimes also includes such other aspects centered in the person as Intelligence, and Professional Training, even though more objective measures are available.

(2) *Performance Standards.* These are categories of job behavior and, for teaching, might include such items as Teacher-Pupil Relationships, Teaching Methods, Committee Service, and Parent Conferences.

(3) *Results, Effectiveness.* In teaching these might include Pupil Progress, Parent Attitudes, and Staff Attitudes.

This classification of types of items on which a person can be rated by a scale method is a general differentiation which can be used with other methods of appraisal too. It makes clear the place of focus of the appraisal: the person, his performance, or the results. In practice, many appraisal methods combine or mix up these elements. The Cumulative Personnel Record system, often developed in large schools or districts, is a combination of such elements which the local group has decided are important. (3)

The whole scaled rating method is subject to defects which can be minimized and controlled to some degree by refinements of technique and training of raters, but never completely eliminated. Such difficulties include the well-known "halo" effect (letting a judgment on one item influence the rating on other items); the skewed distribution (tendency of individual raters to judge generally high or generally low); the fact that items may differ materially in the reliability with which raters use them; the problem of weighting, by which different items are given relative importance; the effect on the rater of the order in which items are listed; the reluctance of many raters to use the extremes; the usual dependence on memory when giving ratings; the fact that in effect the rater is trying to do two things at once: he is both recording-reporting and measuring-judging at the same time; this is desirable efficiency but of doubtful validity. Richardson, who developed the forced choice technique below, has this general comment on rating scales:

"The truth is that graphic or similar rating scales present an almost impossible task to the person who attempts to fill them out. The rater must, almost simultaneously, think of all instances of job behavior that come under each 'trait,' sort out all the significant or critical instances, evaluate them, set up norms or standards to describe just where behavior of zero value lies on the scale, and at the same time implicitly compare the man he is rating with other men on the same or similar jobs. In addition, he must make sure that all the favorable or unfavorable instances of job behavior are added (or subtracted) in a manner that will be fair to all the men, in the sense that they are really being rated on the same basis. At the same time, the rater must free his mind from bias (usually unconscious) for or against the man rated. With judgments so complicated, it is not surprising that good ratings are not thereby achieved." (4)

Scale methods are in wide use in many types of measurement, and new techniques are being developed. However, these frequently involve technical

statistical treatment and are not safe ground for the untrained user. Both to develop and to use valid and reliable rating scales requires more expertness than is available to many of those who wish to employ them, both in industrial and school administration.

2. STATEMENTS. Instead of trying to judge individuals directly in relation to each other, to group norms, or to objectified standards by the rating methods described above, it is possible simply to write out comments or to check specific statements which the appraiser believes to be true (or untrue) about the individual. These two methods are:

a. *Comments,* which the appraiser writes in his own words, as (1) unguided, free form, essay type, or (2) guided, structured, precategorized in which the areas of comment are specified. These are simply written opinions. (The *unwritten,* usually unsystematized collection of opinions is perhaps the most prevalent type of appraisal, which, in spite of its obvious inadequacies, is often the sole basis for important personnel decisions).

b. *Specific Statements,* which are checked by the appraiser as being true or untrue. There are usually about 50 such items to which a reaction is necessary, and two forms are used:

(1) *Forced Choice.* This technique was developed during World War II to avoid the errors and complexities of the typical rating method. There are 5 possible responses to each statement, 2 of which would appear to favor the person, 2 of a negative nature, and 1 neutral. However, the appraiser does not know what the effect of his checking any responses will be. He does not know which ones are weighted or in which direction. He is "forced" to make a critical judgment in relation to statements which are most or least descriptive of the person. The instrument used in this procedure is like a psychological test and requires a trained psychometrician close to the actual work situation for its proper development. It is scored and interpreted by a central personnel office.

(2) *Yes—No,* is a simpler response method of the same type as the original Forced Choice. The appraiser reads an item such as "often picked to supervise special school projects," and responds that it is True about this person, or Not True at present.

Just as do other methods described above, the Statement method assumes that the appraiser is in direct contact with the person being appraised and knows his work intimately. Where this is true, the appraisal process is often used as a meeting ground for the two or more parties and is intended as providing a basis for mutual growth and development.

3. PERFORMANCE RECORDS. All the methods described above imply that the appraiser has seen the individual in action and has had enough contact with him to have had an adequate sampling of his work. However, they do not require what is the essence of the Performance Record method: that is, that a specific, direct, planned observation of the individual performing on the job is made, and a record or report of this observation prepared. Many degrees of rigor in the observations and the recording are possible. Following are the major patterns:

a. *Anecdotal Records* may take two forms:

(1) *Unguided,* in which the observer writes down what he sees and hears that seems pertinent to him.

(2) *Guided,* in which the types of activity and items for record are predetermined and required. Within these

however, the observer would write the description of behavior as he sees it. The *Ohio Teaching Record* (5) is an example of this type: it includes 8 sections containing 58 observation guide items in addition to introductory and summarizing sections. This is a 30-page document designed for in-service development.

b. *Rating Records.* This type, of which Beecher's *Teaching Evaluation Record* (6) is the best example, contains a predetermined list of behavioral items. Beecher includes 32 items descriptive of both teacher activity and pupil response which the observer looks for and rates as occurring consistently and involving most of the pupils, or seldom, etc., in the total pattern of teaching which he is observing. Sample evidences for each item are suggested. Items which are not observed during the rating period are not involved in the final score. Beecher stresses the use of two independent, trained observers over two half-day periods a week or more apart, in order to obtain reliable records.

c. *Continuous, Complete Records.* These might take two forms:

(1) *Complete recording* of verbal and/or non-verbal behavior as it occurs during the observation period.

(2) *Coding of behavior,* as it occurs, in predetermined categories.

The first type would emphasize the obtaining of a complete record during the observation period which would be coded and interpreted later. The second would require the observer to know a code and make his record in its terms. The verbatim record is not a new idea but the development of codes for analysis or making of such records of teaching in terms of role theory may be a special contribution of the current Utah studies.

Major technical problems in the Performance Record systems, as in Rating and Statement systems, have to do with the validity and adequacy of item (or code) selection, objectivity and reliability in both recording and scoring the data, and adequacy of sampling in obtaining the records. Item selection in the Utah studies has involved several techniques based on large-scale teacher responses, the Flanagan critical incident method, and behavioral record analysis. Validity criteria are conceived in terms of the literature on conditions for learning, correlations with judged good and poor teaching, and consensus judgment of relevance, at the present level. Objectivity, reliability, and adequacy of procedure are all being handled as part of the total effort.

Objectivity and Values

There are now, undoubtedly, or will be developed, other ways of looking at teaching. One major choice an investigator has to make is in terms of the degree and quality of objectivity he will seek. Presumably, the steadily increasing ability being achieved to obtain objective pictures of learning will be equally useful and desirable in teaching. It is difficult to see how teaching can become truly professionalized until we can say what it is in precise, behavioral terms, and can take "photographs" of it as it actually occurs. It is surely true that teaching is a highly complex function responding to complex needs, but we must no longer throw up our hands and claim that it is an art too difficult to comprehend. Recent and current investigations are rapidly invalidating this assumption.

Learning how to take a picture of teaching does not, of itself, however, solve the value question. The problem remains of discovering what kinds of acts and patterns of acts are best for pupil

development and for achieving all purposes of the school. As function-analysis methods develop, this next step will not be too difficult to take, though it will have to be taken very carefully and over an extended period of time. The criterion measures would be conceived in terms of kinds of response occurring in relationship to kinds of teaching behavior demonstrated. This would involve a more comprehensive view of response than the ordinary pupil achievement score provides and will require extensions and refinements of measurement technique to encompass other important goals. Significant concern with goals beyond traditional subject matter achievement is seen in the Frenkel-Brunswick, et al., measures of attitude, social distance, and intolerance; Withall's and Wrightstone's measures of classroom climate; Anderson's measures of dominance-submissiveness; and the several sociometric and projective techniques. There is no shortage of competent researchers who can do this work. It is only a matter of their giving their attention to it. (7)

Available research in the fields of interpersonal relationships, group formation and action, perception, concept formation, and attitudes, is already substantial and steadily increasing. It is certainly indicative of function-response relationships, though not as yet conclusive or comprehensive enough. One of the aspects of the present Utah studies should make a significant contribution in the direction of bringing together results of such research as they relate to the role concept in teaching. (Dr. Marie M. Hughes is consultant for this effort and the code-construction effort.)

The values which can result from the intensive pursuit of research down these paths are evident. Kinney (8) has pointed very clearly to needs for diagnosis of teaching expertness as a basis for relevant in-service training programs and other personnel operations within the school system, as a basis for preservice selection and training of professional teachers, and as a basis for communication with the public in terms of the real service they are buying with the tax dollar. Specific, current problems, such as the teacher supply, certification and renewal requirements, fifth year training programs, intensive teacher training for liberal arts graduates, merit salary, supervisor-teacher-administrator relationships, class size, and differentiated teaching functions, all cannot be dealt with in a fully professional, that is, accurate and relevant, manner until teaching has been defined, measured, and evaluated with precision, reliability, and validity. One way to do this has been suggested above in elaboration of an earlier outline. (9) The effort to do it is already well advanced.

This statement should not close without reference to five other implications of a thoroughgoing approach to the evaluation of teaching:

1. The scope of a definition of teaching will be at least partly a function of local views and expectations of what the teaching job is. To the extent that the purposes of schools differ and the expected services from teachers differ, the definitions must recognize this. Classroom teaching functions are probably universal expectations, but not uniformly conceived; but then there are other functions related to curriculum committee work, extracurricular activities, reporting to parents, hall or lunch supervision, and so on. A significant effort of the organized profession to begin thinking through and agreeing on some of these things has employed the leadership of Kinney, Rosencrance, and others in recent years.

(10) It would seem that this effort should go on, while local school districts and teacher groups also give their attention to it. The full impact of all available media of communication and of possible specialization in teaching roles should find a place in these considerations. In time, both general and specific concepts of the professional roles appropriate for teaching in the American culture should emerge and find professional and public acceptance.

2. The significance of other related research will hinge on the success with which teaching is measured. Once this middle problem has been solved it will be possible to move out to meaningful studies of the effects of teaching and to studies of the conditions which have produced the teaching. Lacking objective pictures of teaching, investigators have had to relate their studies of teacher training and characteristics to such inadequate criteria as supervisory ratings. Likewise, studies of pupil learning and response have been related to variables not clearly central in the teaching process.

3. Although teaching is here considered in functional terms, the problem of the *content* handled in the teaching situation must not be ignored. That is, to the extent that the teacher is responsible for the content which is brought into the situation, another problem of evaluation exists. How relevant, accurate, and usable is this material? A number of intriguing questions are suggested which would require considerable space to elaborate.

4. The evaluation of teaching, whether by rating or other objectified procedures, is a highly complex, technical task, and it is a gross error for legislators, school board members, administrators, or teachers to expect it to be done easily and cheaply, and by untrained people. On the other hand, studies of the most efficient ways to proceed should be part of any total investigation.

5. If evaluations of teaching are to have lasting beneficial effects, it will be because the individual teachers whose work is evaluated have learned to view their own efforts somewhat dispassionately and scientifically. Evaluation programs will contain varying amounts of threat for each individual, depending on a whole group of associated conditions. To minimize the threat, and to learn how to make professional use of the objective data obtained, constitute a challenge to the profession as a whole and to each individual in it, which, if successfully met, can raise the general quality of educational effort to a plane now seen in only the exceptional classrooms and schools.

References

1. H. H. Remmers and others. "Report of the Committee on the Criteria of Teacher Effectiveness." *Review of Educational Research* 22:238-63; June 1952. See also the Committee's Second Report in the *Journal of Educational Research* 46:641-58; May 1953.

2. Many references on rating and scaling are available. See (a) Frederic M. Lord, "Scaling." *Review of Educational Research* 24:375-92; December 1954. (b) J. Wayne Wrightstone, "Rating Methods." *Encyclopedia of Educational Research*, Revised edition, 1950; p. 961-64. (c) Industrial Psychology, Inc. "Development of IPI Merit Rating Series." *Industrial Psychology, Inc., Notes.* 1953, 7 p. (d) Bert F. Green, "Attitude Measurement," *Handbook of Social Psychology*, Vol. 1. Cambridge, Mass.: Addison-Wesley Publishing Co., 1954.

3. William C. Reavis and Dan H. Cooper. *Evaluation of Teacher Merit in City School Systems.* Supplementary Educational Monographs, No. 59, January 1945, 138 p. The University of Chicago Press.

4. M. Joseph Dooher and Vivienne Marquis, editors. *Rating Employee and Super-*

visory Performance. New York: American Management Assn., 1950.

5. College of Education, Ohio State University. *The Ohio Teaching Record*, second revised edition. Ohio State University Press, 30 p.

6. Dwight E. Beecher. *The Teaching Evaluation Record*. Syracuse University Press, 1953, 16 p.

7. Two recent examples of studies in this direction are: (a) Harold E. Mitzel, *A Behavioral Approach to the Assessment of Teacher Effectiveness*. Paper delivered to AERA meeting, February 19, 1957. Office of Research and Evaluation, Division of Teacher Education, the College of the City of New York. (b) Carleton W. Washburne, *The Function of a Research Office in a Teacher Education Program*. Paper delivered to AERA meeting, Feb. 20, 1957. Brooklyn College, N. Y.

8. Lucien B. Kinney. "What Is a Good Teacher?" *The Role of the Administrator in the Improvement of the Instructional Program*. Logan: Utah State University; School of Education Bulletin, June 1954.

9. Gale Rose. *Outline of an Approach to the Evaluation of Teaching*. Utah School Merit Study Committee, April 1955, 3 p.

10. National Commission on Teacher Education and Professional Standards (NEA). Reports of Special Groups D(1953), A(1954), and A(1955). Washington, D.C.: the Commission.

GEORGE MANOLAKES

Needed Research in Reading

Research in reading instruction can profitably take several directions indicated briefly by this author.

THE COMPLEXITY of the reading act has always provided a wide range of research possibilities. The concern for the effects of reading upon individuals and groups, and for the ways in which reading is affected by their functioning has extended the scope of reading research investigations into psychology, physiology, and many other related areas of study. However, research in reading for those most directly concerned with the effectiveness of teaching and learning must necessarily be defined in terms of the needs for the improvement of instruction in the elementary and secondary schools.

Our attempts to apply related research findings have often resulted in distortions that have limited their usefulness, and our efforts to emulate these studies have further extended the fragmentizing of the instructional process. In accepting the responsibility for research concerned with the instructional program, we not only shall be functioning in the area of our greatest competence, but further will be contributing to a research function that only education can assume.

The need for this research function was dramatically demonstrated by the anxieties that resulted from the highly publicized criticisms of the teaching of reading when research findings proved to be a most effective deterrent to the regressive tendencies that developed in the face of challenge.

The present status of the reading instruction program suggests three pur-

***GEORGE MANOLAKES** is associate professor of education, New York University, New York, N. Y.*