Higher Education 4 (1975) 173–199

THE ASSESSMENT OF TEACHING IN HIGHER EDUCATION: A CRITICAL RETROSPECT AND A PROPOSAL

PART I: A CRITICAL RETROSPECT

HENRY C. JOHNSON, Jr.,
The Pennsylvania State University,

DENT M. RHODES and
ROBERT E. RUMERY,
Illinois State University

ABSTRACT

Evaluation of college and university teaching is considered in the context of growing demands for accountability of educational institutions and particular roles of faculty in achieving goals for which these institutions are believed to be accountable. In the first of two papers,* three contemporary approaches to evaluation of teaching in higher education are critically reviewed. All three of these approaches – assessment of learning outcomes, assessment of teacher characteristics and analysis of pedagogical behaviors – are found to be defective on logical, theoretical and empirical grounds. A common thread of deficiency is the absence of a coherent theoretical framework for analysis of teaching and phenomena associated with teaching. Specific defects are analyzed in each of the three approaches as well as in the ubiquitous methodology of rating of teaching performance. In analysis of evaluation by assessment of learning outcomes, teaching is shown to be neither necessary nor sufficient to subsequent learning outcomes. Assessment of teacher characteristics fails to identify those characteristics peculiarly indigenous to teaching as a generic activity. Analysis of pedagogical behaviors fails to distinguish critical teaching acts from more general teacher characteristics interpretable in terms of teacher personality. Furthermore, no adequate basis is provided for normative interpretation of data pertaining to pedagogical behaviors. The use of rating scales to provide data about teacher characteristics or pedagogical behavior rests on assumed rather than demonstrated validity of results. Evidence of validity or meaningfulness is replaced by evidence of consistency which is often spurious. The first paper concludes with an outline of requirements for a more constructive approach to the task of teacher evaluation. In a second paper, an outline of a theory of teaching is sketched which conforms to these requirements. Realization of this theoretical structure in teaching assessment report forms is described. Tentative conclusions from trial use of forms-in-development and recommendations for additional data sources are discussed in terms of their potential contribution to improvement in teaching.

* The second paper will be published in the next issue of this journal (August).

Introduction

Although education as a process of personal and social development in terms of some ideal culture has been central to western civilization since the Greeks, in modern times its role has gradually been transformed into a universal, institutionalized process of schooling. "Progressive" western societies have accepted the notion that both optimal personal development and the common good could be brought about by deliberately changing persons, at least in the sense of consciously assisting their development by formal means.[1] Recently, however, the spectre of personal, social and economic breakdown has laid greater and greater tasks on the schools and simultaneously created a growing disenchantment with them. Indeed, some critics now openly suggest that our fundamental belief in schooling is, if not antithetical to genuine human progress, at least seriously defective. While few as yet accept the validity of this radical critique of schooling, there are also few indeed who have not become increasingly skeptical of the school's ability to fulfill its promises. Even so, the vast majority of citizens and governments are reformist rather than revolutionary, preferring to reaffirm their faith in institutionalized educational development by calling for improvement, rather than abandon or curtail such an important instrument of social policy.

At the heart of this process of schooling is the elusive and faintly mysterious instrumentality loosely called "teaching," an activity in which we intervene in one another's lives, presumably effectively and justifiably, in order to bring about the socially and culturally determined changes which we desire. We have generally been inclined to consider teaching an arcane art or mysterious personal power, simply presuming that it must be responsible for whatever personal and social goals schooling does appear to accomplish.

Of late, however, as we have lost something of our confidence in schooling, we have begun to demand a more rigorous examination of teaching as an activity rather than rely upon our former superstitious acceptance. If teaching is the nub of the process of schooling, and if schooling seems to be failing, we are reluctant to commit unlimited resources to them and put absolute trust in them. There ought, we think, to be evidence for their unique claims and for belief in their utility. While much of this demand is frequently wrong-headed, it is undeniable that a new importance is now attached to the long-standing problem of defining and assessing not only the enterprise of schooling in general but teaching

[1] For a contemporary example, see Rowntree (1974), p. 5.

specifically, as a crucial activity within that institutional framework. Thoughtful and responsible teachers themselves are raising the question of how they may more effectively improve the performance of the activity from which they take their name. The result is a widespread and growing urgency to reopen questions which all admit to be serious and complicated.

Unfortunately – and this has been particularly true of what we have come to call higher education – there has been no generally accepted way of either defining or assessing the activity of teaching. Primary and secondary education have been characterized by a plethora of purported theories and specific technical and methodological proposals, but these are largely unrelated to each other and usually without articulation from one level to the next. Colleges and universities have left the issues almost entirely unresolved (Falk and Dow, 1971). Only the necessity of making sporadic institutional judgments, in the face of obvious shortcomings perceived by the public and the participants themselves, has forced even a little change to occur. This has led to confusion, indeed to virtual chaos. The simultaneous demands for action, coupled with the inadequacy of previous analyses, suggest then that the questions of assessing teaching within the context of educational institutions must be reopened, but reopened with great care for their complexity.

Because of the central role of teaching, the crucial question upon which we shall fasten in these two necessarily limited essays will consequently be the following: How can what may broadly be called "instructional activity," as it occurs in institutions of higher education, be improved? Considering the question of improving instructional activity or teaching will, however, require several carefully ordered steps. First of all, it will be necessary to begin by constructing a provisional definition of what teaching is. (This definition is "provisional" not just in the sense of "tentative" but also in the sense of enabling us to look ahead and to test.) Such a definition is necessary in order, next, to stipulate what will be relevant observations and to make them critically. Thirdly, these relevant observations of teaching must be formed into a coherent pattern, not merely in the sense of reporting descriptively the details of the activity, but in the sense of ordering the various possible details normatively. This task obviously entails the further difficult task of determining some principles of value and specifying their application. Only if these steps can be achieved in some systematic and critical fashion will it be possible, lastly, to create plausible prescriptions for the progressive development of instruction either considered as a general program or as an individual activity.

While we will not here follow this program seriatim, nor fully plumb

the depths involved in all these requirements, we will sketch our critique and proposals in such a way as to make these requirements evident and to encourage further discussion and critical development in respect to them. We shall begin by examining categorically existing practical programs for "evaluating" or "assessing" teaching, showing their deficiencies in respect to the requirements just noted. Then, in a second essay, we shall develop a positive proposal which we think more adequate both theoretically and practically.

General Critique of Current Approaches

Previous attempts at constructing programs for evaluating teaching or instruction have been largely unsuccessful and frequently harmful in their effect. These attempts at evaluating teaching are conceived and focused narrowly, indeed almost exclusively, in terms of isolated, individual classroom activities. They are usually separated from the educational context, with its social and cultural reference. Almost entirely, such attempts have been centered in single instruments administered episodically rather than as part of a cumulative program. Evaluation has consequently been realized in terms of solitary, discrete acts, largely (we should argue) because of the methodological form which has seemed most attractive and immediately useful in terms of institutional interests. Finally, each of what we regard as the three fundamental approaches to evaluation presupposes (but usually entirely unconsciously and hence uncritically) some "model" or definition of the teaching process which is ambiguous, unclear, or dubious, and which is frequently mis-educative – i.e., runs contrary to any defensible concept of education as a normative construct. Consequently such instruments are not only practically failing but are incapable of fulfilling the task we have asserted to be fundamental: the warranted improvement of instructional activity or teaching within the context of education and schooling.

The current approaches to the evaluation of teaching can be grouped in three broad categories: 1) measurement of learning outcomes presumed to be the result of teaching; 2) measurement of teacher characteristics presumed to facilitate learning or the attainment of other possible educational goals; and 3) analysis and measurement of relevant categories of pedagogical behavior. In the sections of this essay that follow, we will attempt to show that these three approaches to the evaluation of teaching have "reached a dead end" (Anderson and Hunka, 1963), not because they have been technically misapplied but because they are fundamentally misdirected. However, since all three of these approaches also share two

basic problems: 1) the absence of adequate theoretical development or integration and 2) the confusion of measurement with evaluation, we shall examine these basic problems before proceeding to the separate critiques.

In the last few decades, most published attempts at evaluation of teaching have been characterized by apparent uncritical acceptance of a naive empiricism virtually devoid of any coherent theoretical development or integration of previous results. A blue-ribbon committee charged with assessing the status of research into criteria of teacher effectiveness observed that "research too often proceeds without explicit theoretical framework, in intellectual disarray, to the testing of a myriad of arbitrary, unrationalized hypotheses. The studies too often interact little with each other, do not fall into place within any scheme, and hence add little to the understanding of the teaching process" (American Educational Research Association, 1952). Their statement is a clear admonition to ground programs or procedures for the evaluation of teaching in a coherent theoretical framework. The admonition has not been ignored but neither has it been the foundation for any substantial continuing course of action, as surveys of more recent practice reveal (e.g. Cohen, et al., 1973).

The requirements of a theory of instruction have been outlined repeatedly (e.g. ASCD Commission on Instructional Theory, 1968; Bruner, 1963; Gage, 1963; Thelen, 1951; and Travers, 1966), and the crucial distinction between descriptive theory and normative theory has also repeatedly been made (e.g. by Bruner and by Maccia, 1965). There does appear to be a substantial difference of opinion about the question of whether the theory should be descriptive or normative (prescriptive). Bruner has expressed the belief that a normative theory is essential while Smith (1960) has argued that the present state of knowledge about teaching and its connection with education goals is insufficient to support a normative theory. In spite of this confusion about the appropriate form of a theory of instruction, the use of the term "evaluation" seems to imply the requirement of a normative theory, since judgments of value are necessarily involved. But, as we shall show, the required grounding of a program or procedure for evaluation of teaching in a normative theory is no more evident now than it was in 1952. Even when excursions into instructional theory have been normative in form (e.g. Bruner), there is no general agreement as to which aspects of teaching performance should be included in such a theory. Finally, no extant attempt at formulating a theory of teaching has gone very far toward meeting requirements such as those set forth by Travers, Maccia, and others.

A second problem held in common by all three current approaches to evaluation of teaching is failure to make appropriate practical distinctions between description, measurement, and evaluation. The term

"measurement" commonly refers to representation in numerical form of some discernible aspect of things or events. As we shall see later, use of the word "represent" poses a number of logical, empirical, and theoretical problems, but, for the time being, it will suffice to say that this common use of the word "measurement" is essentially descriptive, not normative. The crucial aspect of the concept of evaluation, on the other hand, is the establishment of the relative worth of alternative outcomes or courses of action according to some explicit concept of value. In any specific instance of evaluation, the nature of the value principle employed can have determining implications about the observations and the form of measurement operations which are part of the evaluation process. There has also been a recent tendency to use "evaluation" as in fact equivalent to "description." In this usage an "evaluation" of programs eventually points out or describes their effects rather than assessing or judging their relative worth. Clearly this is inadequate to the task of determining what improved practice would be and contrary to the fundamental notion of evaluation. (See, for example, Parlett and Hamilton, 1972.)

To illustrate the importance of these distinctions, in an intercollegiate swimming and diving meet, the winner of the 200 meter freestyle event is determined by measuring the elapsed time between the starting gun and the finish of the event. But in the five-meter platform diving event, it would be odd to declare the diver who entered the water most quickly the winner of the event. In the swimming contest, the "value" of a performance is a function of elapsed time as measured by a timing device; in the diving contest, the "value" of a performance is a function of its similarity to an implicit or explicit standard of performance as measured by judgment of experts. To confuse the appropriate procedures of measurement and standards of evaluation is to make the whole enterprise nonsensical.

We shall now turn our attention to a separate examination of each of the three principal contemporary approaches to the evaluation of teaching.

Evaluation of Teaching by Measurement of Learning Outcomes

Perhaps the most obvious approach to the evaluation of teaching is by looking at student learning outcomes as direct results. The ease with which this approach can be stated, and its apparent common sense, no doubt largely account for its perennial attractiveness. (For a recent example, see Popham, 1973.) The approach begins with a conception of teaching as a ternary or triadic relation (Henderson, 1965). In this view,

teaching is something that occurs when a teacher, *A*, teaches some body of content or set of skills, *B*, to some person or group of persons, *C*. What is meant, of course, is that *A* gets *C* to learn some *B*. As Bantock (1961) puts it, this process represents "the conscious bringing about in others of certain desirable mental or dispositional changes by morally acceptable means." "Successful" or effective teaching, in this view, will then be assessed by the degree or quantity of change brought about. We thus measure the effectiveness of a teacher or institution by selecting some *B* and examining the *C*s in order to measure the change presumably induced as a consequent of *A*'s "teaching." Though some, including Bantock himself, have argued that no criteria for good teaching in general are possible (due to potential variance within *B*), it does appear at least indirectly normative in individual cases, since more is patently better. And, while Bantock and others may not like it, it has proved almost irresistibly attractive to compare *A*s, at least all the *A*s who teach the same *B*.

The evaluation of teaching by the measurement of learning outcomes as manifest in students has, however, met with considerable resistance, principally on pragmatic grounds; but, logical, theoretical and more rigorous empirical objections can be raised as well. Inadequacy of technical evaluation resources, inadequately specified or understood goals of instruction, and incomplete or unwilling faculty participation are some of the possible pragmatic obstacles to measurement of learning outcomes as an evaluation of teaching. But even when not adopted as the sole functional base for evaluation of teaching, assessment of learning outcomes is frequently regarded as the criterion against which data from other sources can be validated. For example, McKeachie et al. (1971), in attempting to validate student ratings of teachers, obtained correlations of these ratings with measures of student performance on achievement tests as well as on tests purporting to measure different aspects of thinking ability. Rodin and Rodin (1972) claim that what students have learned is the "objective criterion of teacher effectiveness" and contrast this with "subjective" student ratings. Others (e.g., Tyler, 1948; Cohen and Brawer, 1969; McNeil and Popham, 1973) have promoted measures of student attainment as "ultimate criteria."

The logical basis for the use of measures of student attainment as either proximate or ultimate criteria of teacher effectiveness seems to be represented by the hoary slogan, "if the student has not learned, the teacher has not taught," or, as it is sometimes succinctly stated, "no learning, no teaching." In this view, the paradigm example of the "least effective teacher" is obviously the teacher whose students learn nothing at all. These slogans pose a logical problem, however. This is illustrated by

reconstructing them in the form, "teaching has occurred if and only if learning occurs." Ordinary principles of logic specify exactly two conditions under which an argument of this form is valid. The first condition is that both of the component statements ("teaching has occurred" and "learning occurs") are independently established as true. The second condition under which this form of argument is valid is that both of the component statements are independently established as false. In order to establish empirically the validity of the argument, a minimum requirement is that a discrete set of events or activities be clearly identified as teaching and another set clearly identified as learning. Given these minimum conditions, the argument is valid only when learning is invariably preceded by events identifiable as teaching events. The existence of the phenomenon of incidental learning and countless examples of persons acquiring certain skills without instructional intervention (e.g., learning to play the piano "by ear") suggest strongly that the argument is essentially invalid. In the annals of educational research, such a stringent test of the proposition is virtually unknown. Instead, what appears to be the case is that the argument "teaching has occurred if and only if learning occurs" is assumed to be valid and the truth or falsity of the statement "teaching has occurred" is inferred from the truth or falsity of the statement "learning occurs," a clearly fallacious inference.

A less rigorous view of the relationship between teaching and learning which is intended to justify the measurement and evaluation of teaching through its results in the student is perhaps best grasped in the celebrated analogy developed by John Dewey (*How We Think,* 1910). Dewey appears to suggest that teaching and learning are strictly correlative and exactly comparable to the activities of buying and selling: learning is to teaching as buying is to selling. Thus, Dewey argues that it would be as absurd to say one had taught all day without learning occurring as it would be to say one had sold without buying occurring. There are, of course, a number of complex issues under the surface which have been widely and continuously discussed. However, we think a very simple test will expose the conceptual inadequacy of this analogy: while we can say that nothing which has been bought has not been sold, we cannot say that nothing which has been learned has not been taught – at least unless we wish to make the distinction of teaching from other conditions under which learning occurs entirely vacuous.

There is admittedly a plausible sense in which if we did not *intend* that students learn as an ultimate consequence of what we call teaching activity, then to engage in teaching would be odd behavior indeed. But buying and selling have a similarly loose sense. It is plausible to say that in the market there are buyers and sellers who cannot always buy or sell even

though this is what they intend. Both teaching-learning and selling-buying lack precisely that necessary connection which they must have if we are to make warranted inferences from one to the other, as evaluation by outcomes purports to do.[2]

Beyond these logical and conceptual difficulties, evaluation of teaching by assessment of learning outcomes puts inadequate emphasis on the contribution of the learner to the attainment of learning outcomes. As Rothkopf (1970) put it, "You can lead a horse to water, but only the water that gets into his stomach is what he drinks." His statement is a succinct summary of results of a long series of experiments reported by himself and his associates which established that learning in school-like settings depends far less on structural characteristics of learning situations (including instructional strategies) than on certain crucial activities of learners. Rothkopf coined the word "mathemagenic" to characterize such activities necessary to learning as attending, rehearsing, encoding, reading, etc. Anderson (1970) has suggested attempted control of these mathemagenic activities as central to the activity of the teacher. Without arguing whether such control might be desirable, it should be pointed out that two ubiquitous features of these activities make control difficult. The first feature is that performance of these activities is a matter of choice on the part of the student. In relevant circumstances, the student chooses whether he will pay attention in lectures, read assignments, or review what has previously been read; rarely are these activities the only ones available. Beyond the matter of choice is the matter of capacity. A student may choose to work on an assignment for a calculus class, but be unable to perform the required practice because he lacks essential prerequisite skills. For example, a calculus problem might require application of certain rules or concepts of analytic geometry which the student either failed to master in previous study or forgot.

Against the availability of choice in the performance of appropriate mathemagenic activities, it might be argued that the teacher must somehow "motivate" students to engage in these activities. In fact, motivation is conventionally represented as an essential activity of teaching (Eble, 1972, p. 177). Unfortunately, this argument poses further and more difficult logical problems. The first has to do with how the word "motivate" is being used. A straightforward etymological examination suggests that it is used to signify arousal or activation of motives. If this signification is accepted as a legitimate interpretation of common usage, then other logical problems arise. It is not clear whether this usage is

[2] This and much more is ably handled by Green (1971). See esp. pp. 135ff.

descriptive or explanatory, but in either case it is problematic. If descriptive, the problem is that motives as internal states are not empirically observable, but are rather inferences from action, or from observed conditions presumed to be isomorphic with the internal states. For example, if a student actively participates in discussion, completes assignments on time, or does work beyond that assigned, he is described as "motivated." In this example, "motivated", if used descriptively, is a synonym for a set of behavioral specifications. But in ordinary usage, "motivation" is commonly understood to refer to some causal *agent* which is a *reason* for the occurrence of these activities; and consequently a purely descriptive use of the term would appear to be trivial, at least for our purposes.

The use of "motivate" in an explanatory sense, implying causal agency, has its own difficulties. When loosely used, it is frequently "verified" by reference to the behavior it is alleged to explain. It thus becomes synonomous description, and hence insufficient as explanation. A more tightly constructed approach does not entirely dissolve these difficulties. In the first place, as R. S. Peters has suggested, causal explanations of the kind of purposive or intentional learning which is at issue in schools may be inappropriate because of an important distinction between responding and acting intentionally. The S-R framework for viewing human behavior need not be considered an exhaustive mode of analysis and its use to cover behavior which we wish to examine from other perspectives is not only unnecessary but may be ill-advised. Furthermore, the causal linkage which must be established in the case of schools and school learning is so complex that sufficient confirmation for a useful analysis seems virtually unavailable. And each link in the chain must be verifiable or the whole chain fails. Finally, in practice, explanation by motivation entails reference to other psychological constructs which raise their own problems.[3]

In the face of these difficulties, the proponent of measurement of learning outcomes as an approach to evaluation of teaching might, however, argue instead that all he intends to claim is that teaching may be considered effective to the extent that it increases the probability of occurrence of specified learning outcomes. So stated, the assertion that teaching produces learning rests not on logical grounds but rather on an interpretation of statistical inferences made from empirical data. The statement is in this case not intended to be a statement of *fact*, but a statement of *most plausible belief*. However, before we can regard it as the

[3] For a full analysis, see Peters (1960).

most plausible belief, alternative assertions must be shown to be less plausible or (as statisticians put it) less probable. Siegel and Siegel (1967), following the lines of Campbell and Stanley (1963) have very clearly shown the difficulties of verifying such assertions by means of statistical evidence because of the multiplicity of environing events which might lead to similar outcomes. A student might have achieved certain learning outcomes, for example, by cheating. Alternative explanations of learning outcomes are particulary troublesome in the nonexperimental circumstances usually encountered in school settings. The logic of statistical inference – indeed, the logic of scientific inquiry itself – requires that alternative assertions (i.e., hypotheses) both be made explicit *and* be shown to be less probable. Among the alternatives to be rejected as less plausible, at least the following are of interest: 1) that learning outcomes could have been achieved in the absence of any treatment identifiable as teaching; and 2) that learning outcomes represent prior knowledge of students. As Siegel and Siegel have pointed out, these plausible alternative hypotheses are rarely tested in any educational research, let alone in programs which attempt to evaluate teaching by assessment of learning outcomes.

Tests of these very hypotheses were, however, involved in one recent experimental study of teaching (Leicht and Rumery, 1973). Two results were of interest in this experiment. First, there were no statistically significant differences in test performance among groups of students assigned to four different instructors (although one of the instructors had been cited for superior teaching and another had a total of only one semester's teaching experience, and was not at the time even on the teaching faculty). Secondly, while the effects of prior exposure to material by reading and of hearing the material in lectures were substantially additive, the effect of reading was double the effect of exposure to lectures, no matter who the lecturer was or what the specific content of readings and lectures. These results at least suggest that alternative hypotheses about learning outcomes such as their relation to prior knowledge or actions are *more* plausible than a teacher causation hypothesis.

Finally, evaluation by learning outcomes entails dependence upon tests. The construction of such tests usually does not include adequate procedures to insure their content validity. The meaning most often attached to the term "content validity" is that the test be an adequate and representative sample of some universe of interest. As Cronbach (1970) has shown, however, to achieve genuine "content validity" is a complex problem requiring careful specification before learning can be inferred. But, even in the most refined and elegant testing procedures, the existence of an exhaustive table of specifications is seldom evident. We have little or

nothing of this sort for the subjects now taught in most schools. Even worse, the logically necessary step of testing the fit of actual test content to a table of specifications and of both to educational activities, either intended or realized, is still more rare.

Furthermore, the problem of content validity is complicated by an additional, almost universal failure to distinguish between general and specific content validity. Consider the situation frequently found in middle-sized to large colleges or universities where there are substantial numbers of introductory classes in rhetoric. It is conceivable that a table of specifications could be constructed to assess a set of outcomes common to all sections. Yet an individual teacher might intend and realize specific institutionally and individually valuable outcomes excluded from any such table of specifications and not intended or realized by other teachers. Obviously, the deviant instructor would be at a clear disadvantage when evaluated by student achievement on standardized tests.

While Tucker (1962) has proposed a complex solution to the problem of differing points of view about desirable educational outcomes, it is still easy to imagine that some set of goals intended by an individual instructor and a set of institutional goals could work at cross purposes. In the absence of a coherent normative structure, there is no basis for choice between individual goals and institutional goals, and institutions are not always right. Finally, given that a common test could be constructed which adequately represented institutional *and* individual goals, the risk remains that an instructor could choose to "teach the test" and consequently become identified as an instructor who is "effective." Such a strategy would be, to say the least, *educationally* dubious.

The Evaluation of Teaching by Measurement of Teacher Characteristics

This approach to evaluation of teaching attempts to show that teachers with certain characteristics (such as friendliness, fairness, humor, sensitivity, enthusiasm, or the appearance of competence, for example) are approved, valued, or accepted by individual students or groups of students. The efficacy of the approach appears to rest upon the notion that learning will be increased if students come to perceive their teachers as attractive human beings. Thus, teachers who possess the supposedly desirable characteristics will presumably be good teachers. Furthermore, teachers who possess more of them, or possess them to a greater degree, or appear to possess them in the eyes of a greater number of observers, will be better teachers than those who have them only to a lesser degree.

The logic of this approach appears almost identical to the previous

approach. In its strong form, advocates would have to show that learning occurs if and only if some critical set of teacher characteristics is manifest. In its weaker form (founded on the belief that the relation is not strictly causal but probable) the position would require likewise a demonstration similar to that required for the weaker variant of the previous approach. Consequently, the logical and empirical critiques of this position substantially parallel the arguments developed in the previous section. In order for the assertion to be true, it would have to be independently shown that learning occurs only in the presence of the specified set of teacher characteristics *and* that learning does not occur in their absence. In the educational setting, it would also have to be shown not only that learning does or does not occur in the presence or absence of these characteristics, but that *particular* learning occurs or fails to occur to an acceptable degree. There appears, however, to be equally slight warrant (either logical or empirical) for connecting any of the commonly enumerated characteristics with teaching at all, let alone with teaching effectiveness. On the contrary, they seem indistinguishable, except in their setting, from any list of pleasing characteristics generally found in one's acquaintances or preferred for office supervisors or factory foremen. Now, it may be that we want all teachers to be pleasing, but that is not all we want, and it is the crucial differences which go unexamined in this approach.

The empirical validity of this approach is presumed to lie in the application of the so-called "critical incident technique" (Flanagan, 1954). In a familiar application of the technique, subjects with "considerable educational experience" were asked to specify "the 'very best' and 'very poorest' teachers" they had had when in school and, further, "to describe some incident or something outstanding that was remembered over the years" about these teachers (Ryans, 1960, p. 79). Respondents were specifically cautioned to avoid descriptions of critical incidents which a) named or listed personality traits, b) reported behavior idiosyncratically important to the reporter, c) reflected general stereotypes about teachers, or d) reported incidents primarily important for their dramatic impact. The principal advantage of the technique is that it presumably replaces vague generalities with concrete instances of good or poor performance. But, as Cronbach (1970) has pointed out, the technique is not truly objective. The incidents are recalled within a conceptual framework – an "implicit theory of teaching," to paraphrase Cronbach (1955) – and a reporter is more likely to recall only that information which was conceived to be relevant at the time it was received. While the reporters in Ryans' study were mature adults, the information they were to recall was stored when they were children, and the information available from memory is controlled by the conceptions of teaching held by them as

children, not as mature adults.

Theoretical support for the teacher characteristics approach to evaluation of teaching appears to originate in Rotter's social learning theory (1954) and the body of research on effects of authoritarian versus democratic leadership, stemming from the original research of Lewin, Lippitt and White (1939) and White and Lippitt (1960). As Anderson (1959) has pointed out, results of the authoritarian-democratic studies have been ambiguous at best when either group productivity or morale are used as criteria of effectiveness of leadership. The ambiguity in these results is hardly surprising, for two reasons. First, the terms "authoritarian" and "democratic" are so laden with surplus meaning that realization of conditions univocally interpretable in these terms is difficult. Even if such realization were achieved, Anderson's critique continues, the consequences would not be highly generalizable since, in most situations, teachers would not be so harsh as to be characterized as "authoritarian," nor so nondirective as to be characterized as "democratic." More likely, they would try to be as nondirective as task requirements and situational demands allowed; hence any "type" characterization, let alone the extremes of "authoritarian" or "democratic," would be consistently appropriate in only an extremely small share of circumstances.

The role of pleasing characteristics in enhancing the teacher's value as a positive social reinforcer is not only an uncritical equation of reinforcement with reward, but, as a substantial body of research suggests, misrepresents the contribution of reinforcement to human learning. Estes and some of his associates reported a series of experiments in which two components of reinforcement, reward and information, were independently controlled (Keller et al., 1965; Humphreys et al., 1968). The results of their experiments supported the hypothesis that the information component enhanced performance on a verbal discrimination task but counter-indicated a reward hypothesis.

The Evaluation of Teaching by Analysis of Pedagogical Behaviors

It is because of difficulties such as those just outlined that attempts to focus upon various forms of specifically pedagogical behavior have generally replaced the crude use of personality characteristics. In this approach, teaching is seen as a generic activity, definable in its own terms. Attention is usually paid to one or more teacher "acts," "actions," "activities," or "behaviors," variously categorized as "logical," intellectual," "strategic," "linguistic," "performative," "expressive," "skill," "institutional," "managerial," "organizational," among others. Perfor-

mance of these behaviors can be further appraised as "successful," "effective," "good," "preferred," or some other evaluative designation. The pedagogical behaviors may be identified by the analysis of teaching-as-practiced (Reagan, 1965; Smith et al., 1967; Komisar, 1968; Gray, 1969; Green, 1971), or application of the critical incident technique (Hildebrand et al., 1971; Ronan, 1971).

Teaching, in this sense, can be viewed as consisting of some set of logical, linguistic, and/or psychological operations carried out by the teacher in a particular social context. The more proficient a teacher is in engaging in these operations, the better that teacher is said to be. In this case, students (or other observers) monitor these operations and make judgments on the extent to which they are present and/or how well they are performed. But this approach suffers from two principal defects: first, analysts show little if any agreement on which pedagogical behaviors are most significant, how they should be categorized, or in what combinations, if any, they may or ought to appear; second, formulations derived from the use of the critical incident technique lack a coherent theoretical basis and, as with teacher characteristics, are based on each student's particularized and implicit conception of teaching. In the absence of theory, consistency of response can be attributed just as legitimately to a collective student mythology of teaching as to any rigorously conceived model of teacher behavior.

Indeed, the emphasis on teacher behavior itself, rather than on personality characteristics or presumed results of teaching, may be more apparent than real. For instance, if by the teacher's giving a "good explanation" or "motivating" successfully, one means that students understand what is explained or that they act motivated, then the assessment is still being made on the basis of results achieved by students, not on proficiency in performing the act of explaining or motivating. It is quite possible that logical acts of teaching – e.g., inferring, defining, comparing, etc. – can be evaluated independently of their results (Green, 1971), but it is very doubtful whether students (or, for that matter, "peers") can be expected to have sufficient expertise to make judgments about the adequacy of such acts. Their capacity for judgment would again have to derive from the supposed results of a logical act of teaching itself, or from some antecedent condition.

The language used in most student response forms based on pedagogical behavior is usually sufficiently global and vague as to raise strong questions about whether personality characteristics are not in fact being called for and being reported. Even if they use other words, they tend to reduce purported teaching behavior to personality characteristics or psychological conditions and they call for observers to record what are in

effect their idiosyncratic, affective responses to those characteristics and conditions. The proposal by Hildebrand et al. is instructive. Their program of evaluation began by a process of having students nominate a group of teachers as their "best" or "worst." A five-item instrument was produced, based on student selection of those descriptors of aspects of teaching which were characteristic of the best and worst teachers they had previously named. "Thus," the authors maintain, "a short-form rating instrument was established that is quickly answered, yet is objectively known to be broad, balanced, and highly discriminating between effective and ineffective teachers."

The five items are as follows:

(1) Has command of the subject, presents material in an analytic way, contrasts various points of view, discusses current developments, and relates topics to other areas of knowledge;

(2) Makes himself clear, states objectives, summarizes major points, presents material in an organized manner, and provides emphasis;

(3) Is sensitive to the response of the class, encourages student participation, and welcomes questions and discussion;

(4) Is available to and friendly towards students, is interested in students as individuals, is himself respected as a person, and is valued for advice not directly related to the course;

(5) Enjoys teaching, is enthusiastic about his subject, makes the course exciting, and has self-confidence.

It is perhaps worth noting that the authors are at pains to state that the "scale" which they derive from item (5) and which they call the "Dynamism/Enthusiasm" scale, is the "most highly related" to the original "ratings of overall effectiveness." The scale they denominate "Organization/Clarity" is supposedly the second most closely related.

Now, clearly, of the five items in this instrument, items (3), (4) and (5), are essentially global descriptions of personality and the traits in items (4) and (5) seem but faintly restricted to teaching. Furthermore, each is complex, and the question of whether any instructor might embody these traits differently, or in differing degrees, is an obvious one. Item (4) appears particularly bothersome, due to the inclusion of "is valued for advice *not* directly related to the course" (emphasis supplied), a curious descriptor of pedagogical behavior indeed! Finally, all the items raise questions of what their key terms mean, and whether any meaning is sufficiently stable across the reports of the various students to offer any real information.

Items (1) and (2) are open to many of the same difficulties but they appear to be getting at a somewhat different object, something that at first sight at least resembles pedagogical behavior. Yet, item (2) ("makes

himself clear, states objectives, presents material in an organized manner, and provides emphasis") proves upon closer inspection an apt illustration of the collapsing of purportedly "pedagogical" behavior into personality characteristics. For example, two key emphases within the item are "clarity" and "organization." It requires very little practical experience or acquaintance with pedagogical theory to realize that what is to students only partly known and insufficiently understood quite easily appears to them as the fault of a confused and disorganized instructor. The result, we argue, is that students then necessarily answer in terms of their own idiosyncratic responses. The subject matter involved may indeed have been presented in a confused and disorganized manner, but the reported *perception* that it has does not entitle us to say that it in fact was. This approach can therefore *not* produce the claimed "objectively known" and "highly discriminating" distinction between "effective and ineffective teachers." In addition, since effectiveness is purportedly being measured, a valid judgment in this case could very likely be made only by someone who had actually learned as a consequence of the instructor's activity. By doing so, however, we are for all practical purposes returning to the first approach to evaluation and asking whether something has been learned. And, if we must abandon that approach, we are left with "evaluation" as nothing more than a recording of how the student feels about something he attributes in some way to the instructor.

Finally, there is the problem of establishing logical, empirical, and theoretical grounds for the choice of any particular set of pedagogical behaviors as the basis for evaluation of teaching. That a charismatic actor posing as a teacher can, through his behavior, deceive even experienced educators has been strikingly demonstrated (Naftulin et al., 1973). In this investigation, although the content of a lecture and discussion was intentionally "irrelevant, conflicting, and meaningless," those who participated rated the "teacher" quite favorably on such items as "Did he put his material across in an interesting way?" and "Did he present his material in a well-organized form?" If these items allow the raters to respond favorably without regard to the quality of the intellectual activities involved, then serious questions about both their substance and utility must be raised.

Educational researchers concerned with evaluation have rarely offered logical grounds for their choice of particular sets of pedagogical behaviors, but when they have, these grounds have been related to the potential enhancement of learning associated with the occurrence of these behaviors. This logic leads us again to the measurement of learning outcomes as a method of validation and hence into the thicket of logical problems already discussed. The sole empirical warrant for particular

choices of sets of pedagogical behaviors seems to arise from application of the critical incident technique with all of its difficulties. As for theoretical warrant, most evaluation forms relying on observation of pedagogical behavior have been almost totally atheoretical, depending upon the mathematical legerdemain of complex data reduction techniques to arrive at "meaningful" interpretations of their results. When study of pedagogical behaviors has stemmed from any theoretical base, that base has tended to be descriptive rather than normative, and the consequences of confusing description with evaluation have already been discussed.

Implied Teaching Models and Their Effects

We alleged earlier that any approach to the evaluation of instruction necessarily requires some conception or "model" of teaching itself in order to function. If space permitted, it would be highly instructive to examine a number of particular proposals and construct their explicit or (more often) implicit models. However, the multiplicity of proposals makes that unfeasible. More importantly, in almost every case, no model has in fact been consciously and critically developed. Indeed, careful examination suggests that the models involved usually flow not from clear concepts of what teaching is but from certain measurement techniques which are gratuitously presumed to be effective in locating evidence. A curious result follows: what is *now possible* to "measure" becomes uncritically accepted as what is *important* to measure. The whole process of measurement and evaluation is hence not only confused but in effect turned upside down.

As is already evident, however, the three categories of approaches to evaluation we have discussed do betray some general tendencies of interest. They tend in general to operate on an unsubstantiated but at least loosely "causal" model. In this model, teacher-caused learning becomes the ultimate criterion for teacher value. We have already stated that while we accept pupil learning as an ultimate *intent* justifying the educational process as a whole, we find the causal linkages defective (even in their milder forms). More importantly, causal models are generally inappropriate for the purpose of evaluating teachers in schools which claim to educate in any but the most trivial sense of that term.

Most unconsciously adopted models of "teaching" also tend to incorporate specific teaching techniques as a focus for their observations (e.g., "lecture," "discussing," etc.). So far as we can see, there is simply no adequate logical or empirical evidence for presuming the effectiveness of

any one, or any set, of these techniques in producing learning.[4] Furthermore, most of the models direct attention principally at the teacher or the student as individuals rather than toward the interactive process of teaching itself. Exceptions to this, such as classroom observation schemes which examine the "climate" of the school room or analyze logically the verbal interchanges found in class sessions, tend simply to describe the interaction or the conditions and thus (at least as presently developed) seem unsuited to the task of evaluating and improving teaching.[5]

It is also worth noting that the tendency to focus on "teacher-caused" learning assumes greater value for learning derived from a teacher than from other sources, a value assumption of considerable significance. This principle leads to an important consequence; practical attempts at evaluation based upon teacher-caused, or what we might call teacher-effect models, tend to ignore student initiated learning, even though (at least in our view) it much more closely approximates defensible educational goals.

Finally, the lack of explicitness and clarity with respect to models of teaching which afflicts current proposals for evaluation is, we think, made evident by the fact that where one would expect to find improvement as a consequence of their use, little if any is to be found (Centra, 1973). If effective models were governing the assessment of teaching activity, one could not expect such improvement necessarily, but it would be odd to find virtually nothing even resembling it. Yet, this is the case. Even mere change is seldom found, at least as any direct result of the enormous effort which has been devoted to evaluation in recent years.

Our conclusion is that because the implied models are almost always unconscious and uncritical, and usually inadequate upon closer inspection, certain unfortunate results have followed. The effect of the prevailing, loosely "causal" models of teaching has been uncritically to throw into prominence particular features of the teacher-student relationship – notably a preoccupation with student achievement and/or student acceptance either of the teacher's characteristics or certain features of his classroom manner or actions. This has led to a further preoccupation with rating scales derived from observations by students as the easiest, and apparently most appropriate, mode of making an assessment of the teacher.[6] These scales provide additional "models" of teaching and supply covert principles for its evaluation (again in an almost entirely uncritical manner)

[4] See, for example, Walle (1972), Kallos (1973), Dubin and Taveggia (1969).

[5] See, for example, Rosenshine and Furst (1973), esp. pp. 160–62.

[6] Representative examples of such rating scales and suggestions for their use may be found in Eble (1970) and Miller (1972).

which also focus attention heavily on the teacher rather than upon teaching as an interactive activity. The tendency toward the uncritical use of rating scales has been, on our view, nearly fatal to the whole enterprise of evaluation. Rating scales are certainly inadequate to the task imposed upon them and (when seen in their true effect) quite possibly contradictory to the very intent of education as a process. It is to the problems posed by the use of rating scales that we will now turn.

Analysis of Ratings

"The most serious fault in the application of ratings is that their validity is accepted on faith when investigation might show that the faith was seriously unjustified" (Guilford et al., 1962). In a test of this assumption these authors found that ratings of scientific workers on ability factors were uncorrelated with tests measuring these same factors even though raters were highly trained and familiar with the persons they were rating. Instead, ratings on ability factors provided more information about ratings on other, presumably distinct, factors than they did about test scores on corresponding factors. Guilford and his associates attributed these anomalous results to "constant errors" (Guilford, 1954) characterizing rating "styles" of individual judges. When students are called on to rate teachers, either in terms of teacher characteristics or pedagogical behaviors, the difficulty is exacerbated by the fact that students are neither trained in judging the required characteristics nor any more than nominally familiar with teachers they are rating. Theoretical analysis of ratings (Coombs, 1964) suggests even more fundamental difficulties.

In the application of rating scales commonly used in evaluation of teaching, it is *assumed* that response categories are ordered with respect to attributes identified in item terms. For example, according to conventional usage, a response "strongly agree" to the item "The instructor encouraged students to think for themselves" (ETS, 1971) would be interpreted as indicating a substantial amount of instructor behavior in support of independent thought on the part of students. On the other hand, a response "strongly disagree" would be interpreted as indicating absence of behavior with that intended consequence or as indicating behavior with opposite intent. Coombs hypothesizes that the "strongly agree" response should instead be interpreted as indicating that the degree to which the instructor encourages independent thought coincides with an implicit "ideal point" characteristic of an individual rater.

The hypothesis is graphically depicted in Figure 1. Two instructors differing in the degree to which they support independent thought are

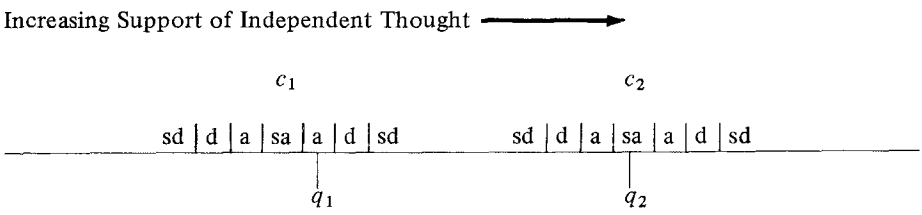

Fig. 1. Joint distribution of student ideal points (c_1, c_2); response category boundaries; and instructor points (q_1, q_2) on a J-scale representing increasing support of "independent thought."

represented by points q_1 and q_2 on a J-scale (a joint distribution of "ideal points" and points characterizing instructors). Ideal points of two student observers are represented by points c_1 and c_2 with the short bars on the J-scale representing boundaries of the response categories "strongly disagree" (sd), "disagree" (d), "agree" (a), and "strongly agree" (sa). The student whose ideal point is at c_1 strongly agrees with the statement in question as it applies to instructor q_1 and strongly disagrees with the statement as it applies to instructor q_2. The situation is exactly reversed for the student observer whose ideal point is represented by c_2. In scoring these responses, both "strongly agree" responses would be scored in the same way even though the two observers reverse the locations of the two instructors on this scale. By securing estimates of ideal points independently of ratings, Levinthal et al. (1971) verified that for scales on which the indicated response categories are implicitly evaluative, Coombs' theoretical interpretation is essentially correct.

Perhaps the most important implication of Coombs' hypothesis and its subsequent empirical support is that interpretation of numerical scale values obtained from scales of this type are problematic without independent knowledge of the ideal points of observers rating teaching performance. A similar conclusion is reached in Remmers' (1963) characterization of ratings as perceptual reports. If these reports are viewed as the product of inferences about instructor behavior, a theoretical analysis suggested by Sarbin et al. (1960), the validity of these inferences depends upon correspondence between perceptual reports and specific, objectively observable instructor actions. Elementary scientific and logical considerations require that such correspondence be demonstrated rather than assumed, a condition rarely, if ever, encountered. Although these conclusions are not sufficient to establish that ratings cannot be valid, sufficient doubt is cast to preclude their use as prima facie evidence of teacher effectiveness.

In place of the fundamental analysis required to establish adequate evidential status, what is commonly encountered in the literature on rating methods is reliance on "face validity" or on reliability (See e.g. Costin et al., 1971). But the substance of face validity is the assumption of validity to which Guilford et al have addressed their criticism.[7] The only basis remaining to justify use of rating methods as sources of evidence of effective teaching is the claim that ratings by multiple student observers yield data which are reliable. This claim, however, is both substantively and logically defective.

The substantive defect in the claim that ratings are reliable is that commonly used methods for assessing the reliability of rating data are inappropriate to the task. Typically, reliability estimates are obtained by one of several internal consistency procedures – most often split-half or procedures related to the Kuder-Richardson formulations. In these methods, what is regarded as systematic information is consistency among items or subsets of items. The defect here is that it is not the items in a rating questionnaire which are the primary instruments of measurement but, as Remmers has pointed out, the observers who use the questionnaires. What appears to be required are estimation methods involving repeated observations by a fixed pool of observers of a common group of instructors using a fixed set of items.[8]

The logical defect in justification of rating scale data on grounds of reliability lies in the implicit extension of the *claim* of reliability to the *argument* that if data from rating methods are reliable, then they must also be valid (i.e. useful or meaningful). This argument is logically false: while it is true that an unreliable measure cannot be valid, the fact of reliability does not guarantee validity in any of its forms. Consequently, justification of use of ratings on grounds that they are reliable can only be seen as an unwarranted substitution of consistency (which itself may be unsupported) for meaningfulness.

Some Conclusions

If, then, the three most common approaches to the evaluation of teaching betray serious deficiencies, and if the most frequently used

[7] For a critique of the concept of face validity, see Mosier (1947). For a more thorough discussion of validity, see Cronbach (1971). For discussion of relations between validity and meaningfulness, see Suppes and Zinnes (1963) or Coombs et al., (1970), Chapter 2.

[8] For detailed discussion of methods of estimation of the reliability of ratings, see Ebel (1951), Guilford (1954), LaForge (1965), Stanley (1961).

measurement method (the rating scale) further complicates the question, the task would seem to require our beginning afresh in the light of what may be learned from such a critical appraisal. That task should be to construct a comprehensive approach to evaluation as a program rather than simply to put together yet another instrument to "measure" and "evaluate" some facet of the process which has groundlessly been assumed to be a significant indicator of the total process of teaching. The basic lessons are clear:

1) Any evaluation of teaching, if it is to be interpretable or to lead to improvement in the general level of teaching practice, must be theoretically grounded.

2) Any theoretical formulations must be normative, rather than merely descriptive. Theoretical sentences must refer to the ideal case, rather than to the median level of teaching-as-practiced. Here, of course, there will be differences of opinion; but, in our view, the term "evaluation" (implying, as it does, some principle of value) necessitates a normative structure. Furthermore, measurements associated with realization of theoretical formulations must involve proximity to an ideal rather than magnitudes of certain characteristics or extent of certain behaviors, although we acknowledge that the two may be identical in particular instances.

3) Theoretical formulations must be grounded in empirical data from broad areas of human learning. We do not intend to suggest, as has been frequently but erroneously asserted, that teaching is either necessary or sufficient to student learning. On the other hand, it is clear that the intended purpose of nearly all teaching is the facilitation of student learning. We hasten to add, however, that our concern (in the context of education as a process) is not with learning as a generic activity, but with particular learning. Consequently the required theory must also deal with the relative *value* of various categories of learning and deal with factors affecting *choice* of alternative activities.

4) Adequate conceptions of the evaluation process must enable us clearly to distinguish the accountability of teachers from accountability of students and accountability of other components of educational institutions: administration, governing boards, funding agencies, etc. Furthermore, the particular demands of accountability must be specified in terms of genuinely educational goals.

We shall begin the attempt to construct a more adequate approach to teaching assessment by attempting first to answer the question of what teaching is. While we do not suggest that a perfect or exhaustive definition of teaching is being offered – or, indeed, is even available – a defensible one which gets at necessary aspects of the process must be constructed.

This definition must enable us to distinguish essential teaching activities from the host of accidental or insufficient activities in which teachers may engage. It must be usable as a means of observing teaching, whether by students, peers, or the teacher himself. If it meets these requirements, it will then be possible to measure at least selected necessary aspects of the process and then to map and employ them in such a way that cumulative development and improvement in teaching is possible.

This definition must also be normative, and not merely descriptive. It must allow for qualitative discrimination among outcomes and practices and not merely indicate their relative "effectiveness." In this way it will provide a basis for the assessment of individual practitioners in relation to one another and (of much greater importance) in relation to their own individual development. A normative definition of teaching, utilized as the basis of a continuous program of assessment, would provide the possibility of effective prescription in cases of demonstrable deficiency and could shape possible professional growth as well. Finally, a normative definition of teaching is necessary for the direction of new activities, including the development of rigorous theoretical formulations of teaching and programs of investigation and testing. It would also appear to suggest outlines for a productive regimen for the preparation of teaching personnel.

The task of laying out such a general conception of teaching, and indicating the assessment program which might flow from it, will be taken up in the next essay.

References

American Educational Research Association (1952). Report of the Committee on the Criteria of Teacher Effectiveness. *Review of Educational Research* 22: 238–263.

Anderson, C. C. and Hunka, S. M. (1963). "Teacher Evaluation: Some Problems and a Proposal". *Harvard Educational Review,* 33: 74–96.

Anderson, R. C. (1970) "Control of Student Mediating Processes during Verbal Learning and Instruction". *Review of Educational Research,* 40: 349–369.

Anderson, R. C. (1959). "Learning in Discussions: A Resume of the Authoritarian-Democratic Studies." *Harvard Educational Review,* 29: 201–215.

Association for Supervision and Curriculum Development, Commission of Instructional Theory (1968). "Criteria for Assessing the Formal Properties of Theories of Instruction". In Gordon, I. J., ed., *Criteria for Theories of Instruction,* pp. 16–24. Washington, D. C.: Association for Supervision and Curriculum Development.

Bantock, G. H. (1961). "Educational Research: A Critique." *Harvard Educational Review,* 21: 264–280.

Bruner, J. S. (1963). "Needed: A Theory of Instruction." *Educational Leadership,* 20: 523–532.

Campbell, D. T. and Fiske, D. W. (1959). "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin,* 56: 81–105.

Campbell, D. T. and Stanley, J. C. (1963). "Experimental and Quasi-Experimental Designs for Research on Teaching." In Gage, N. L., ed., *Handbook of Research on Teaching,* pp. 171–246. Chicago: Rand McNally.

Campbell, N. R. (1920). *Physics: The Elements* Cambridge: Cambridge U.P.

Centra, John A. (1973) "Do Student Ratings of Teachers Improve Instruction?" *Change* 5, (April) pp. 12–13.

Cohen, A. M. and Brawer, F. B. (1969). "Measuring Faculty Performance." *ERIC Clearinghouse for Junior College Information.* Washington, D.C.: American Association for Junior Colleges.

Cohen, A. M., Trent, J.,W. and Rose, C. (1973). "Evaluation of Teaching in Higher Education." In: Travers, R. M. W., ed., *Second Handbook of Research on Teaching,* pp. 1041–1052. Chicago: Rand McNally.

Coombs, C. H. (1964). *A Theory of Data.* New York: Wiley.

Coombs, C. H., Dawes, R. M. and Tversky, A. (1970). *Mathematical Psychology: An Elementary Introduction.* Englewood Cliffs, N.J.: Prentice-Hall.

Costin, F., Greenough, W. T. and Menges, R. J. (1971). "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." *Review of Educational Research,* 41: 511–535.

Cronbach, L. J. (1970). *Essentials of Psychological Testing.* 3rd edn. New York: Harper and Row.

Cronbach, L. J. (1955). "Processes Affecting Scores on "Understanding of Others" and "Assumed Similarity." *Psychological Bulletin,* 52: 177–193.

Cronbach, L. J. "Proposals Leading to Analytic Treatment of Social Perception Scores." In: Tagiari, R. and Petrello, L., eds., *Person Perception and Interpersonal Behavior,* pp. 353–370. Stanford, Calif.: Stanford U.P.

Cronbach, L. J. (1971) "Test Validation." In: Thorndike, R. L., ed., *Educational Measurement,* pp. 443–507. Washington, D.C.: American Council on Education.

Dawes, R. M. (1972). *Fundamentals of Attitude Measurement.* New York: Wiley.

Dubin, Robert, and Taveggia, T. (1969). *The Teaching-Learning Paradox.* Eugene, Oregon: University of Oregon Press.

Ebel, R. L. (1951) "Estimation of the Reliability of Ratings." *Psycho-metrika,* 16: 407–424.

Eble, K. E. (1972). *Professors as Teachers.* San Francisco: Jossey-Bass.

Eble, K. E. (1970). *The Recognition and Evaluation of Teaching.* Salt Lake City, Utah: Project to Improve College Teaching.

Educational Testing Service. (1971). *Student Instructional Response Schedule.* Princeton, N.J.: Educational Testing Service.

Falk, Barbara and Dow, Kwong Lee. (1971). *The Assessment of University Teaching.* London: Society for Research into Higher Education.

Flanagan, J. C. (1954). "The Critical Incidents Technique". *Psychological Bulletin.* 51: 144–151.

Gage, N. L. (1963). "Paradigms for Research on Teaching". In: Gage, N. L., ed., *Handbook of Research on Teaching,* pp. 94–141. Chicago: Rand McNally.

Gray, C. E. (1969). "The Teaching Model and Evaluation of Teaching Performance." *Journal of Higher Education,* 40: 636–642.

Green, T. F. (1971). *The Activities of Teaching.* New York: McGraw-Hill.

Guilford, J. P. (1954). *Psychometric Methods.* 2nd edn. New York: McGraw-Hill.

Guilford, J. P., Christensen, P. R., Taaffe, G. and Wilson, R. C. (1962). "Ratings Should Be Scrutinized." *Educational and Psychological Measurement,* 22: 439–447.

Henderson, K. B. (1965). A Theoretical Model for Teaching. *School Review,* 73: 384–391.

Hildebrand, M., Wilson, R. C. and Dienst, E. R. (1971). *Evaluating University Teaching.* Berkeley, California: Center for Research and Development in Higher Education.

Humphreys, M. S., Allen, G. A. and Estes, W. K. (1968) "Learning of Two-Choice, Differential Reward Problems with Informational Constraints on Payoff Combinations". *Journal of Mathematical Psychology,* 5: 260–280.

Kallos, D. (1973). "On Educational Scientific Research." *Report from the Institute of Education, University of Lund.* No. 36 (April).

Keller, L., Cole, M., Burke, C. J. and Estes, W. K. (1965). "Reward and Information Values of Trial Outcomes in Paired-Associate Learning." *Psychological Monographs,* 79 (Whole No. 605).

Komisar, B. P. (1968). "Teaching: Act and Enterprise." *Studies in Philosophy and Education,* 6: 168–193.

LaForge, R. (1965). "Components of reliability." *Psycho-metrika,* 30: 187–195.

Leicht, K. L. and Rumery, R. E. (1973). *Role of Teacher Structuring and Student Structuring of Learning Materials in Student Learning.* Washington, D.C.: Department of Health, Education and Welfare, Office of Education, Bureau of Research, Final Report, Project No. 001692, Grant DEG–5–41–0054 (50B).

Levinthal, C. F., Lansky, L. M. and Andrews, C. (1971). "Student Evaluations of Teacher Behaviors as Estimations of Real-Ideal Discrepancies: A Critique of Teacher Rating Methods." *Journal of Educational Psychology.* 62: 104–109.

Lewin, K., Lippitt, R. and White, R. K. (1939). "Patterns of Aggressive Behavior in Experimentally Created "Social Climates." *Journal of Social Psychology,* 10: 271–299.

Maccia, E. S. (1965) "Instruction as Influence toward Rule-Governed Behavior." In McDonald, J. B. and Leeper, R. R., eds., *Theories of Instruction,* pp. 88–99. Washington, D.C.: Association for Supervision and Curriculum Development.

McKeachie, W. J., Lin, Y. and Mann, W. (1971). "Student Ratings of Teacher Effectiveness: Validity Studies." *American Educational Research Journal,* 8: 435–445.

McNeil, J. D., and Popham, W. J. (1973). "The Assessment of Teacher Competence." In Travers, R. M. W., ed., *Second Handbook of Research on Teaching,* pp. 218–244. Chicago: Rand McNally.

Miller, Richard I. (1972). *Evaluating Faculty Performance.* San Francisco: Jossey-Bass.

Mosier, C. I. (1947). "A Critical Examination of the Concepts of Face Validity". *Educational and Psychological Measurement,* 7: 191–205.

Naftulin, Donald H., Ware, John E. Jr. and Donnelly, Frank A. (1973). "The Doctor Fox Lecture: A Paradigm of Educational Seduction." *Journal of Medical Education,* 48: 630–635.

Owens, M. S. (1971). "Evaluation of Teaching Competence by Three Groups of Educators." *Journal of Experimental Education.* 40: 77–82.

Parlett, Malcolm, and Hamilton, David. (1972). *Evaluation as Illumination: A New Approach to the Study of Innovatory Programs.* Occasional Paper No. 9, Center for Research in Educational Sciences, University of Edinburgh. (October).

Peters, R. S. (1960). *The Concept of Motivation.* London: Routledge; New York: Humanities.

Popham, James. (1973). *Evaluating Instruction.* Englewood Cliffs, N.J.: Prentice-Hall.

Reagan, G. M. (1965). "Toward a More Justifiable Theory for the Evaluation of Teachers and Teaching." Unpublished doctoral dissertation. Michigan State University.

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

Rodin, Miriam and Rodin, B. (1972). "Student Evaluations of Teachers." *Science,* 177: 1164–1166.

Ronan, W. W. (1971). *Development of an Instrument to Evaluate College Classroom Effectiveness.* Washington, D.C.: Department of Health, Education and Welfare, Office of Education, Bureau of Research, Final Report, Project No. 1-D-045.

Rosenshine, Barak and Furst, N. (1973) "The Use of Direct Observation to Study Teaching." In Travers, R. M. W., ed., *Second Handbook of Research on Teaching,* pp. 122–183. Chicago: Rand McNally.

Rothkopf, E. Z. (1970) "The Concept of Mathemagenic Activities." *Review of Educational Research,* pp. 325–336.

Rotter, J. B. (1954). *Social Learning and Clinical Psychology.* Englewood Cliffs, N.J.: Prentice-Hall.

Rowntree, Derek. (1974). *What Is Educational Technology?* (Monograph No. 1. The Open University Institute of Educational Technology) Milton Keynes.

Ryans, D. G. (1960). *Characteristics of Teachers: A Research Study.* Washington, D.C.: American Council on Education.

Sarbin, T. R., Taft, R. and Bailey, D. C. (1960). *Clinical Inference and Cognitive Theory.* New York: Holt, Rinehart.

Siegel, L. and Siegel, L. C. (1967). "A Multivariate Paradigm for Educational Research." *Psychological Bulletin,* 68: 306–326.

Smith, B. O. (1960). "A Concept of Teaching." *Teachers College Record,* 61: 229–241.

Smith, B. O., Meux, M., Nuthall, G. and Precians, R. (1967). *A Study of the Strategies of Teaching.* Urbana, Ill.: University of Illinois, Bureau of Educational Research.

Stanley, J. C. (1961). "Analysis of Unreplicated Three-Way Classifications, with Applications to Rater Bias and Trait Independence." *Psycho-metrika,* 26: 205–219.

Suppes, P. and Zinnes, J. L. (1963). "Basic Measurement Theory." In Luce, R. D., Bush, R. R. and Galanter, E., eds. *Handbook of Mathematical Psychology.* Vol. I. pp. 1–76. New York: Wiley.

Thelen, H. A. (1951). "Experimental Research toward a Theory of Instruction." *Journal of Educational Research,* 45: 89–136.

Travers, R. M. W. (1966). "Towards Taking the Fun out of Building a Theory of Instruction." *Teachers College Record,* 68: 49–60.

Tucker, L. R. (1962) "Factor Analysis of Relevance Judgments: An Approach to Content Validity." In Dressel, P. L., Chrmn., *Proceedings of the Invitational Conference on Testing Problems,* Princeton, N.J.: Educational Testing Service.

Tyler, R. W. (1958). "The Evaluation of Teaching." In Cooper, R. M., ed. *The Two Ends of the Log.* Minneapolis: University of Minnesota Press.

Walle, A. (1972). Beyond Teaching Methods: Educational Encounters in Need of a Theory." *Journal of Management Studies.* (October), pp. 274–90.

White, R. K. and Lippitt, R. (1960). *Autocracy and Democracy: An Experimental Inquiry.* New York: Harper.