Florida Journal of Educational Research, 1971, 13

PROBLEMS AND ISSUES IN CONTEMPORARY EVALUATION WITH EMPHASIS ON SYSTEMS OR PROGRAMS SUCH AS TEACHER EDUCATION*

Leonard S. Cahen

Educational Testing Service

Summary

The 1960s have been marked by major changes in education. These changes have included major social and curriculum innovations. Accompanying these innovations, we have observed a rebirth in the philosophy and technology of evaluation. The paper covers some general evaluation issues and questions including the relationship of evaluation and educational research. A plea is made for two stage evaluations of programs or systems such as teacher education. The first step includes the evaluation of the specific program to bring about changes in teacher behavior. The second stage looks at the effects of the change behaviors of teachers on pupils and the school environment.

The 1960's have passed into history. These have been dynamic years for education. Among the 10 major education events listed for the 1960's in the Education section of *Time* magazine (December 26, 1969), one notes the acceleration of high school curriculum reform by Educational Services Inc. (ESI), the birth of Project Head Start, and the far reaching implications of Congress passing the Elementary and Secondary Education Act of 1965. [1]

These three issues reflect important changes in social philosophy and the information explosion that have brought us through a decade of innovation.[2] Accompanying this wave of innovation in education, we have observed a rebirth in the philosophy and technology of evaluation.

This paper will sketch some of the issues and problems of contemporary evaluation. The issues are challenging. The thinking about evaluation develops slowly. Some evaluation topics, such as the need and desirability for stating goals of instruction in behavioral terms, are hotly debated. A decade ago the need of stating outcomes in specific behavioral terms was rarely, if ever, questioned

*Invited address, Florida Educational Research Association—NCME meetings at Jacksonville, Florida, January 23, 1970

In addition to the presentation of a general sketch of some evaluation issues and questions, my presentation will briefly discuss the evaluation of teaching plus a proposal for the development of an interactive system that will incorporate curriculum, teaching methods, individual differences of learners, and environmental dimensions. A final section of the paper will discuss the relationship of evaluation to the broader area of educational research.

Before turning to the list of issues and questions, let me provide a working definition of what I mean by evaluation. I define evaluation as a rational process of reaching decisions about the worth of something. The "something" may be the quality of a new version of a science program or curriculum, a film strip, a new model of the tape recorder to be used in a school room, Teaching Method A, etc. Evaluation is more than obtaining a set of measures reflecting outcomes or output of a specific program. Evaluation requires the application of value judgments in the interpretation of the data. A certain level of output may be interpreted as high-level performance by consumer A, a low-level performance by consumer B, and possibly irrelevant to the frame of reference of consumer C.

We are slowly learning to consider the role of values in evaluation. Values help determine what is included in the curriculum package or product, who is taught the curriculum, what goals are established, what measures are taken to assess performance, and lastly, how the assessments are weighted by the consumer in reaching decisions about the effectiveness of the materials or program.

One of the important developments in evaluation over the past decade was the identification of two separate stages of evaluation—the developmental or modification stage and the public release stage Michael Scriven (Scriven, 1967) labeled these roles as *formative* versus *summative* evaluation in his provocative monograph called "The methodology of evaluation." We will return to these two roles shortly.

Evaluation is learning to borrow skills and approaches from many disciplines within the behavioral sciences. Current thinking suggests we consider incorporating, where appropriate, field research—observational techniques from anthropology, the skills and methodologies of the econometrician in looking at cost-effectiveness of innovations, etc.

[1]Malcolm Provus (1969) cites the clause in the 1965 Elementary-Secondary Education Act which established evaluation as a necessary building block in the design of American educational reform and felt this evaluation requirement would eventually have greater impact on education than the program itself.

[2]Wayne Welch (1968) estimated that NSF contributed over $100,000,000. to major curriculum projects in the first 11 years following the initial grant.

The consideration of methodology beyond the scope of psychology and psychometrics makes it clearly evident that evaluation, as a decision-making process, is a complex enterprise. There are no simple questions (such as whether curriculum A is better than curriculum B), no simple answers. Methodological and conceptual understandings for meaningful and effective evaluation must be developed.[3]

Let us now move on to a partial list of evaluation questions and issues. The first issue deals with the roles, or stages of formative versus summative evaluation. This point was mentioned briefly earlier.

Evaluation Issues or Questions

1 Stages of Evaluation (Formative vs Summative)–Who evaluates?

2 Logical Analyses of Materials

3 Outcomes–Major Effects (Intended and Unintended) and Side-Effects

4. Short-Term Versus Long-Term Outcomes

5 Comparative Evaluation

6 System Evaluation

7 Methodology and Design

8 Behavioral Objectives

The *formative* evaluation stage attends to the development, tryout and sequential modification of materials or products. Measures are obtained on pupil performances, teacher opinions about outcomes, etc. This information is fed back to the innovation team and the product is modified and the cycle starts once again. The summative stage is reached when the product has gone through the necessary number of tryouts or formative stages. It is at this final stage that Scriven feels a large display of information and outcome data on the product must be made available to the

[3]See Bloom (1969), Cronbach (1963), Grobman (1968), Merwin (1969), Scriven (1967), Stake (1967), Stake and Denny (1969), R. W. Tyler (1969), and Welch (1969) for detailed analyses of the changing roles of evaluation.

potential consumers so that this information is available for a rational selection from among competing products. It must be pointed out that the distinction between formative and summative evaluation is not as clear as it first may appear to be. They do suggest different purposes and methods, however. If innovation and evaluation are to be continuous activities in our culture, a summative evaluation really serves as the first stage of a new formative evaluation.

A related question to issue or question 1 reflects on the question, "Who evaluates?" Scriven has suggested that staff or team evaluators be heavily involved in the formative stages while an independent body of evaluators, with no personal commitment to the product, come in and serve as the summative evaluators. One might consider the *Consumer Reports* approach as one illustration of summative evaluation.

The second issue involves the need for developing logical approaches for analyzing materials or products. The analyses must be made of the content or discipline underlying curriculum materials as well as the dimensions of how the materials are to be taught, the congruence between goals and the ways that the materials will be used in reaching the goals, etc. From logical analyses, clearer statements of the outcomes should be obtainable. This would allow a group of, say mathematicians, to look over a set of mathematics curriculum materials and identify the areas where there is agreement and disagreement about the logic of the discipline underlying the materials. From logical analysis, one should be able to obtain an explicit statement about what specifics of the discipline have been included in the curriculum package, which ones have been omitted and why.

The third issue or question involves the dimension of outcomes in evaluation. Over the past decade, we have become more and more convinced that a single achievement measure does a great injustice in reaching decisions about the effectiveness of a program or innovation. We are starting to move toward multidimensional outcomes. For example, if we were evaluating a mathematics curriculum package, we would want to know how well the students could handle simple computations, how well they could solve problems, how well they know basic terminology, how well they could solve a new piece of mathematics, etc. Too often in the past we have looked at a single achievement measure as our sole evidence in reaching decisions.

We also are learning to attend to the dimension of affect as an important evaluation area. We already know (and have really known for many years) that this is a very difficult area to conceptualize and assess. However, it is important to look at affect, as well as achievement, as an outcome. One would be very disturbed if we were raising the achievement level of pupils but at the same time making the students dislike the subject so much that they would refuse to elect courses in the field when this opportunity was available at later stages of education. You will note that I have made a simple breakdown of the different types of outcomes in Issue 3 of the table. We must look for the intended and unintended dimensions of output. The unintended dimension might be considered as a form of side-effect. It is also important that the evaluators and innovators specifically list the types of outcomes they would consider as having positive valence and the types of outcomes they would consider as having negative valence. This, in connection with looking at the unintended or side-effects, gives a much sounder base of information for decision-making. A problem arises when one raises the question of unintended or side-effects. If one could really anticipate side-effects, he would build the assessment of these dimensions into the systematic evaluation. As a precaution in identifying side-effects, it is very important that we do a great deal of observation in the process of the evaluation. Another option to consider is the systematic gathering and analysis of comments of teachers, parents, etc. Too often we play down this type of information as being too "soft." I think we are learning that "soft" information in the form of opinions and observations has a very real place in the evaluation scheme.

We have learned to utilize information about side-effects from techniques that have been developed in medical evaluation. For example, a certain drug may reduce the level of infection but make the patient so drowzy he cannot drive an automobile safely.

The fourth issue deals with short-term versus long-term outcomes. Many of the innovators of the large national curriculum projects in the past decade felt that many important outcomes would not become apparent until two or more years passed after the materials were introduced. Many evaluation studies will require long-term as well as short-term evaluation so that a more complete picture of achievement and attitude can be displayed.

The commitment of long-term assessment in evaluation implies that the funding agencies and the innovators are working in education over the long haul and are not developing "quickie" courses or materials that are in and out of the schools rapidly.

Issue number five involves the comparative aspects of evaluation. What are the issues underlying the commonly asked question, "Is method A (new curriculum) superior to method B (old curriculum)?" Lee Cronbach (Cronbach, 1963) felt that the comparison of an "old" curriculum versus a "new" curriculum defied logic. He expressed the opinion that the new curricula had different goals and expected outcomes and therefore made comparison logically impossible. The issue of comparative evaluation has been one of the major questions raised over the past decade. Scriven feels that comparative evaluation must be done and reported at the summative evaluation stage.[4] The question, I think, is not whether comparative evaluation should be done but what kinds of comparative evaluation yield meaningful information for evaluation. I would propose that comparative evaluations can be done on dimensions such as affect, teacher's liking of and ability to teach the competing curricula, the ability of students to apply their knowledge to new areas of learning (transfer of learning) etc. I do not feel that it makes a great deal of sense to compare curriculum A versus curriculum B on a dimension that has been included in one of the curricula but not the other. I am hopeful that we will learn more about the types of comparative evaluations that make sense and yield information for decision-making.

Issue number six deals with systems evaluation. A system is a combination of components. Going back to the examples of curriculum, the actual learning materials (texts, etc.), form one component of the system, teachers form another component, the environment and the community form another component, etc. One must consider outcomes under each of the different components of the system and separate evaluations must be made for each component. At a final stage, a total evaluation is made of the sub-evaluations and the components. Too often, as Henry Dyer (Dyer, 1968) has pointed out, curriculum packages have been developed independently of the teachers who must teach them. Dyer points out that almost everybody has searched for a "teacher free" curriculum.

[4]The Scriven Monograph reflects a viewpoint developed after extensive interaction about evaluation issues with Lee J. Cronbach. The reader will find the 1963 paper by Cronbach titled "Course improvement through evaluation," plus the Scriven Monograph as two of the best papers dealing with contemporary approaches to evaluation.

Issue number seven deals with methodologies and designs for evaluation. Many evaluators have criticized the general application of the experimental method to evaluation (Sufflebeam, 1968; Provus, 1969). Part of the problem involves the issues already discussed under comparative evaluation. From the work of experimental psychology and education and from classical statistical models, we have become accustomed over the years to the concept of a control group versus an experimental group. What does a control group really mean in an evaluation study? Is there really any group that does not receive any type of instruction? Rarely, I think. What is more appropriate as an approach, I think, is to conceptualize different manipulations for teaching a specific curriculum. The manipulations are tailored to the individual differences of the learner. (See Lindvall & Cox, 1969.) We now attempt to learn what type of modification of the innovation or curriculum package is appropriate for what type of student with what type of teacher. There is a great deal of discomfort among contemporary evaluators about what types of methodologies and designs are appropriate. It is agreed that outcomes must be multidimensional as was pointed out earlier. Current methodological problems also lead us to believe that some of the classical psychometric models for assessing individuals may not be the most appropriate models to use in formative evaluation. We are learning that a great deal of information lies in the assessment of group performance rather than individual performance. Techniques such as item-sampling may allow us to assess group performance over many different outcome areas. In 50 minutes of testing, it seems tragic to test all students in the group on the same items. We must also learn to obtain more information for evaluation studies by making micro-analyses of the types of errors students make. This is vital for feedback to the innovator.

The last issue or question deals with the topic of behavioral objectives. The interested reader should consult the papers by Atkin (1968), Bloom (1969), Eisner (1969), Grobman (1968), Popham (1969), Sullivan (1969), L. Tyler (1969). The basic issues involve whether innovators can specify objectives in behavioral terms and when the objectives should be specifically outlined in the formative evaluation stage. As a point of science, it is necessary to state objectives so one can determine whether they have been reached. Anyone dealing with evaluation over

the past 10 years probably has observed the reluctance of people from the disciplines (say from mathematics or science) to state their outcomes in behavioral terms. This is not difficult to understand and should not be disconcerting. These people have not been trained to talk about outcomes in the form of behavioral objectives. A well-trained evaluator can work with the innovator in the development of these objectives. It also has been pointed out that the whole innovation activity can grind to a halt when an evaluator hounds the innovation staff to state the outcomes in observable behavioral terms. The innovators have trouble doing this and the creative enterprise of innovation may never get off the ground. One technique that seems worthy of exploration is to ask the innovator to develop test items that complement developed materials. It seems reasonable that approaches can be found to tease out the behavioral objectives from analyses of these test items.

This list of eight issues or questions dealing with contemporary evaluation is certainly not a finite set. Other issues such as the study of innovation, how changes take place in school districts, process in the classroom, etc. are also worthy of attention. We will now proceed to the topic of the evaluation of methodologies intended to improve teaching.

Over the past few years we have seen interesting things taking place in the area of instruction or teaching. They have run the gambit from CAI (Computer Assisted Instruction) which some people conceive as a way of holding the method of instruction constant to a re-exploration of the teaching process through vehicles such as micro-teaching and the wide scope of programs under the heading of Training Teacher Trainees (TTT).

These programs also require evaluation. However, I wish to make one point at this time. The act of teaching must be conceived as a component in a learning system. That is to say, we must search for a combination of teacher characteristics (styles, techniques, teacher personality dimensions, etc.), characteristics of the learner, and the types of materials to be learned. This sketchy referrel to a model will be returned to later.

Donald Medley (1969) has listed three questions that are keys to research in teacher education. These questions are:

1. What are the behavioral skills a teacher must possess in order to be effective?

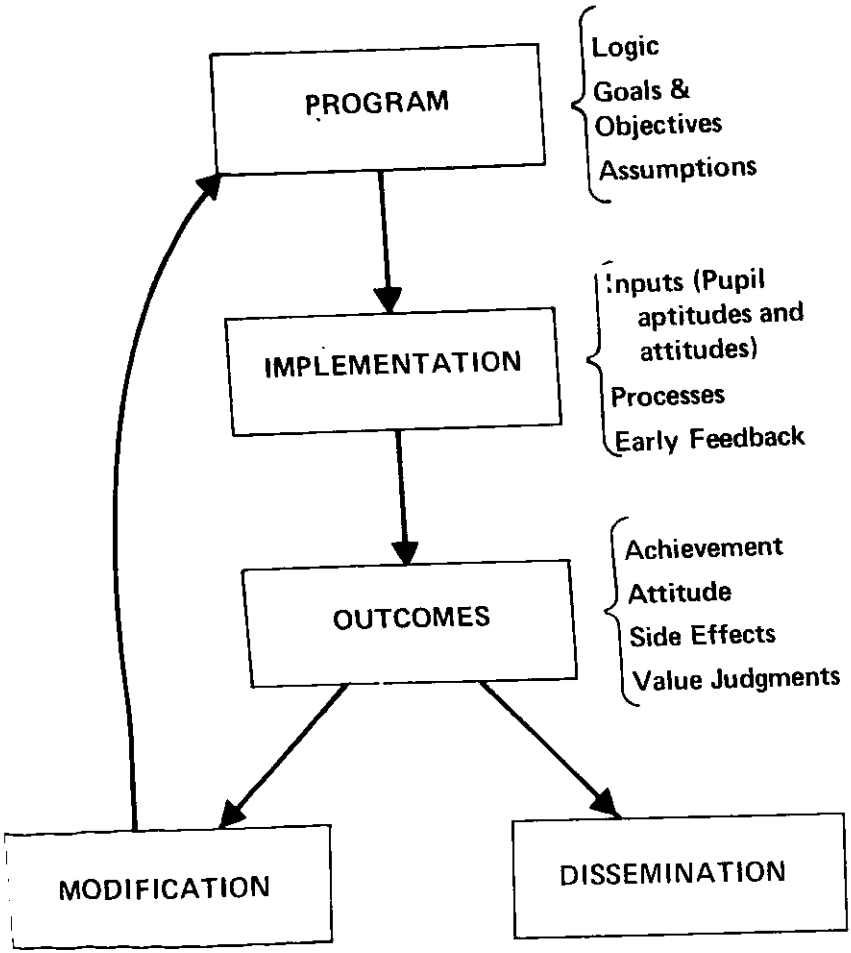
2. What are the characteristics a student must possess before he can acquire these skills?

3. What are the training experiences that will help the student acquire the skills more efficiently?

Medley points out that the failure to answer the first question has precluded any possibility of success in answering the other two. Medley also alerts us, correctly, to the fact that we have too frequently neglected the concept of individual differences among teachers. He also urges that we conceptualize teacher effectiveness in a multidimensional sense—i.e., there are many different kinds of teacher effectiveness and a particular teacher may be more effective in one sense than the other. This idea is consistent with the part of my earlier presentation where I stated that we were beginning to become more and more aware of the need to make assessments for evaluation over a wide range of outcomes.

Let us look at Figure 1.

FIGURE 1



Assume that the term "program" in Figure 1 represents a type of teacher-training activity. The program may be developed from theory, value systems, goals of the program developer which reflect how he wants teachers to be if the program is successful,

etc. The program then proceeds through a series of formative evaluations which include feedback and modification, a loop back to implementation, etc. Hopefully, after successive stages of formative evaluation, the method is considered adequate for dissemination. This is not to say, however, that all programs are generalizable and need to be disseminated. It is perfectly acceptable that a program be developed that attempts to train teachers for specific, or even unique, situations.

Let us turn now to Figure 2.

FIGURE 2

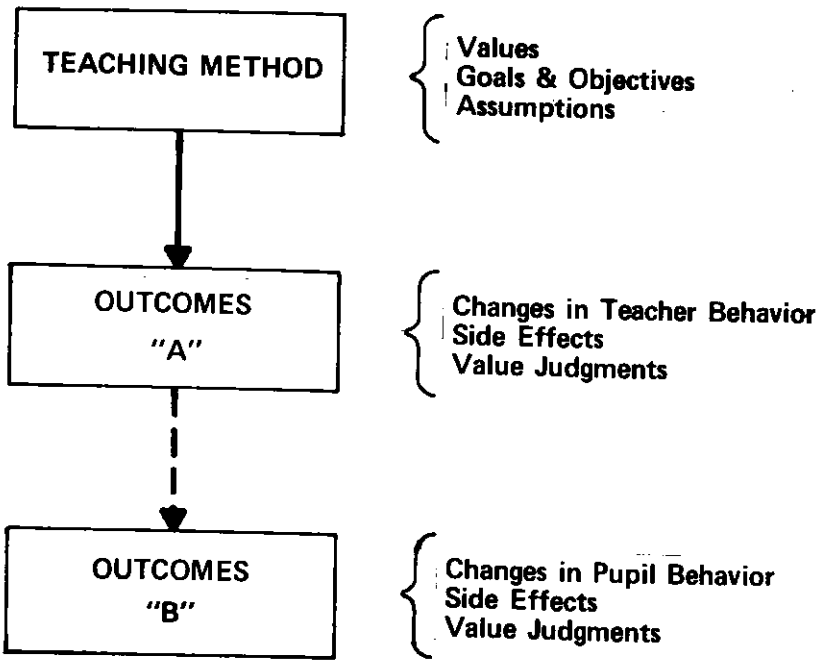


Figure 2 takes the schema of Figure 1 out one additional step. This step is a crucial one and is too often neglected in the development and evaluation of instructional methods. The additional step requires the program director and evaluator to look at the crucial question:

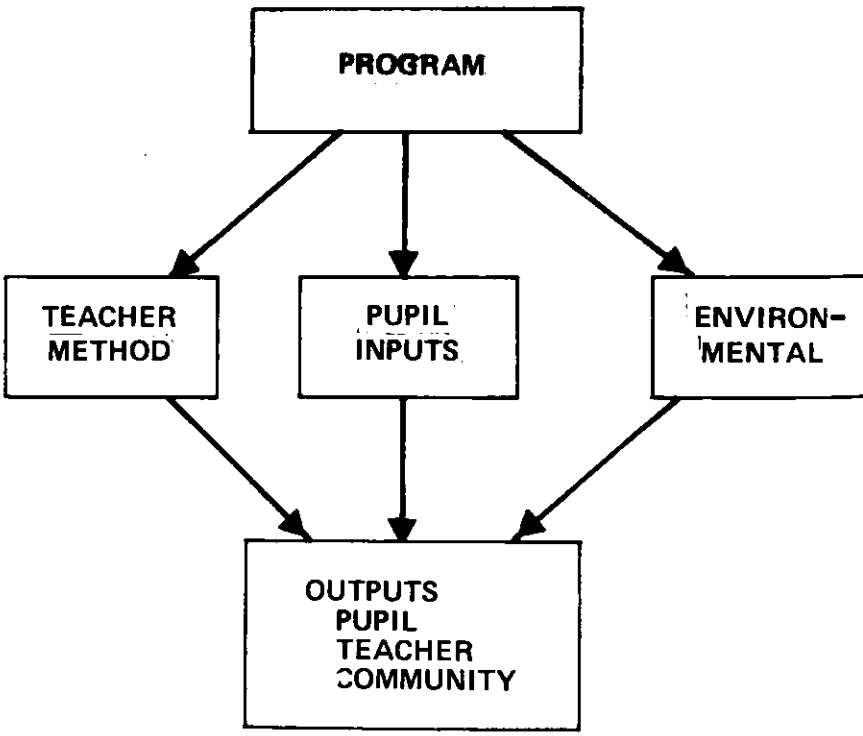
> "If I have changed teachers' behavior by the program (outcome A), are there gains in the positive effects on students (outcomes B)?"

Medley has stated that the ultimate objective of any teacher-education program must be measured in what happens to pupils. We often hear exciting comments about the effectiveness of a teacher-training method or technique (micro-teaching for example). The comments tell us that the behaviors of the teachers being trained by the method have been significantly changed for the better. However, acknowledging that the comments are true, I am urging that we ask the next question—

> "Will the changes in teacher behaviors result in positive changes in pupil behaviors when the teachers go into the classroom with real pupils?"

Figure 3 illustrates the integration of teacher-training methods, pupil inputs, and environmental dimensions into a systematic program. I am hopeful that future stages of innovation will take into account methods of training teachers, teacher characteristics, aptitudes and characteristics of the pupil including the student's history of learning, and environmental variables. By environmental variables I mean special situations such as urban education, etc.

FIGURE 3



An important dimension that must also be integrated into the system, which has been left off Figure 3, includes the development of different types of materials (say curricula) for different students that will integrate and mesh with teaching method, teacher characteristics, pupil characteristics and environment. For example, one might conceive of three types of a geometry curriculum. One type of material would be for students who have the aptitudes to learn geometry when the material is presented in a visual or spatial mode, verbal material for students who are highly verbal, etc. The concept of aptitude-treatment interaction is a method that attempts to incorporate and integrate pupil aptitudes with the material. I am encouraging us to include teacher style and teaching methods as part of this tailored package of instruction. Figure 3 alerts us to look again at multidimensional outputs which will give us a feeling about the

overall success of the program but will also provide us with information about the modification of the component of teaching method, pupil characteristics, etc.

The final part of my presentation will address itself to the relationship of evaluation to the broader area of educational research. Some interest and attention to this relationship has been given recently by professionals in the field. The interest has resulted in concerns about the nature of the evaluation and research processes. Added emphasis has come from an awareness of the shortage of competent professionals to staff the many projects and institutions requiring research and evaluation activities, and from the increasing concern about how to train evaluators. Universities such as UCLA, Illinois, Stanford, Minnesota, to name a few, are now offering specific training (and even majors) in the field of contemporary evaluation. Stake and Denny, (1969) and Hemphill (1969) have looked at some of the distinctions between research and evaluation. Stake and Denny take the position that evaluation does not have the responsibility for making its findings generalizable. This is to say that the principal difference between research and evaluation, in the opinion of Stake and Denny, is the degree to which the findings are generalizable beyond thier applications to a given product, program or locale. In their words the evaluators sacrifice the opportunity to manipulate and control (a basic in research endeavor) but evaluation gains relevance to the immediate situation.

Hemphill feels that research and evaluation share many characteristics of method and approach, but evaluation differs from research on dimensions of generalizability, the role of values, and the amount of control one places on the process of obtaining information.

Cronbach and Suppes (1969) make a distinction between conclusion-oriented inquiry versus decision-oriented inquiry. These terms are used as replacements for the often used terminology of "basic" versus "applied" research. Cronbach and Suppes include operational or institutional research, and product or developmental research, under the classification of decision-oriented inquiry and cite examples that evaluation, as conceptualized in this paper, can be considered as falling in the decision-oriented inquiry category.

It is dangerous to consider evaluation as a sloppy, non-rigorous enterprise without adequate controls. Good evaluation requires a high level of rigor.

Hopefully, we will learn more about the nature of evaluation and its relationship to research. It would seem reasonable that good evaluation will have payoff and contribution to educational theory. An educational system needs theory to systematize individual differences in teachers, learners, environments, and materials.

I have tried to cover many topics in this paper. A short list of some of the key issues and questions about contemporary evaluation has been presented. The plea for looking at the effectiveness of teacher-education programs as being reflected by pupil performance was made. I have also tried to encourage the concept of a system that would incorporate individual differences of learners, teachers, environments, and materials. Lastly, the relationship of evaluation to research was briefly discussed.

One hopes that during the next decade we will learn to utilize evaluation activities rather than give lip service to the concept. Evaluations of the major curriculum innovations of the past decade were either poor or nonexistent. We must first learn to accept the responsibility to evaluate, and then learn how to do it.

References

Atkin, J. M. Some evaluation problems in a course content improvement project. *Journal of Research in Science Teaching*, 1963, 1, 129-32.

Atkin, J. M. Behavioral objectives in curriculum design. *Science Teacher*, 1968, 35, 27-30.

Bloom, B. S. Some theoretical issues relating to educational evaluation. *Educational Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969.

Cronbach, L. J. Course improvement through evaluation. *Teachers College Record*, 1963, 64, 672-83.

Cronbach, L. J., & Suppes, P. (Eds.) *Researeh for tomorrow's schools: Disciplined inquiry for education.* A report of the Committee on Educational Research of the National Academy of Education. London: Macmillan Co., 1969.

Dyer, H. S. The art of unwrapping curriculum packages or how to be an educational string saver. *The Bulletin of the National Association of Secondary-School Principals.* 1968, 52, 141-158.

Eisner, E. Instructional and expressive educational objectives: Their formulation and use in curriculum. *Instructional Objectives:* AERA Monograph Series on Curriculum Evaluation, No. 3. Chicago: Rand McNally, 1969.

Grobman, H. *Evaluation activities of curriculum projects.* AERA Monograph Series on Curriculum Evaluation, No. 2. Chicago: Rand McNally, 1968.

Hemphill, J. K. The relationships between research and evaluation studies. *Educational Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969. Chapter 9, pp. 189-220.

Lindvall, C. M. & Cos, R. C. The role of evaluation in programs for individualized instruction. *Educational Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969.

Medley, D. M. The research context and the goals of teacher education. *Educational Comment.* Toledo, Ohio: University of Toledo, College of Education, 1969.

Merwin, J. C. Historical review of changing concepts of evaluation. *Educational Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969.

Popham, J. W. Objectives and instruction. *Instructional Objectives.* AERA Monograph Series on Curriculum Evaluation, No. 3. Chicago: Rand McNally, 1969.

Provus, M. Evaluation of ongoing programs in the public school system. *Educational Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969.

Scriven, M. Methodology of evaluation. *Perspectives of Curriculum Evaluation* (edited by Robert E. Stake.) AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967. pp. 29-83.

Stake, R. E. The countenance of educational evaluation. *Teachers College Record,* 1967, 68, 523-40.

Stake, R. E., & Denny, T. Needed concepts and techniques for utilizing more fully the potential of evaluation. *Education Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969.

Stufflebeam, D. L. Evaluation as enlightenment for decision-making. Columbus, Ohio: College of Education, Evaluation Center, The Ohio State University, 1968.

Sullivan, H. Objectives, evaluation, and improved learner achievement. *Instructional Objectives.* AERA Monograph Series on Curriculum Evaluation, No. 3. Chicago: Rand McNally, 1969.

Tyler, L. A case history: Formulation of objectives from a psychoanalytic framework. *Instructional Objectives.* AERA Monograph Series on Curriculum Evaluation, No. 3. Chicago: Rand McNally, 1969.

Tyler, R. W. Introduction to NSSE Yearbook. *Educational Evaluation: New Roles, New Means.* Sixty-eighth Yearbook, Part II, National Society for the Study of Education. Chicago: Univ. of Chicago Press, 1969.

Welch, W. W. The need for evaluating national curriculum projects. *Phi Delta Kappan,* May, 1968.

Welch, W. W. Curriculum evaluation. *Review of Educational Research,* 1969, 39, 429-443.