EVALUATION PERSPECTIVES: 1968

C. Robert Pace

CSE Report No. 8
December 1968

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

EVALUATION PERSPECTIVES: 1968

C. Robert Pace

It is easy to document the pervasiveness of evaluation in education today. It exists under the mandate of many large federal and state programs. The pervasiveness of evaluation is evident by the existence of the UCLA Center for the Study of Evaluation, the Evaluation Center at Ohio State, other university-based evaluation centers, the AERA report on curriculum evaluation, and the 1968 AERA presession training program on evaluation, which was attended by 90 people actively responsible for evaluating some significant educational enterprise. It is evident in the spirit of reform and innovation that one feels in many segments of education--new curricula, new technologies, new administrative patterns, and new clientele to be better served. It is evident in the general anxiety and uneasiness one feels in the presence of complex social problems and the desire for more effective solutions.

The diversity of activities that are thought of as comprising evaluation indicates that evaluation is a label which can be and is applied to a rather large assortment of problems and processes--so many, in fact, that the term itself has lost almost all precision and perhaps much of its capacity to communicate among teachers, administrators, and researchers. Many of the things that

are done under the label of evaluation could be called by another name.

The testing of products to describe their characteristics is called evaluation. Why not simply call it product testing? The accumulation of data about an institution's operation--its income, expenditures, costs per credit hour, faculty-student ratios, etc.--is called evaluation. Why not simply call it institutional accounting? The measurement of pupils' knowledge at the beginning and end of a course is called evaluation. Why not simply call it achievement testing? The diagnosis of pupils' present knowledge and skills and the assignment of pupils to individualized instructional treatments is called evaluation. Why not simply call it pupil diagnosis and assignment? Or, perhaps, instructional engineering? The study of procedures used to facilitate change or innovation and the willingness to modify plans as they are carried out is called evaluation. Why not simply call it sensitivity to group processes or openness to change and adaptation? The measurement of changes in pupils' interests, attitudes, values, etc., over time is called evaluation. Why not simply call it the study of personality development? The particular interactions between teachers and pupils and the discovery that certain approaches are effective with some students while different approaches work with other students is called evaluation. Why not simply call it the study of instruction? The collection of data and their review by an accrediting agency is called evaluation. Why not simply call it accreditation? The collection and use of information by administrators for decision-making is called evaluation. Why not simply call it the study of decision-making?

One could continue with further examples until nearly everything that has been called evaluation was eliminated. But then the most important characteristic of evaluation--namely, its pervasiveness-would also be missing. The fact is that all these activities are evaluative. The understanding of evaluation in all its forms calls for perspective, not for exclusion.

When we examine the diversity of concepts, practices, and methods in the field of evaluation--historically and analytically--we will find that the ways in which we think about evaluation and how we go about making evaluations are necessarily related to what we are evaluating and why it is being evaluated. The scope, content, method, and purpose of evaluating a set of frames in a programmed instruction sequence are quite different from the scope, content, method, and purpose of evaluating the consequences of a multi-million dollar investment in Head Start programs.

Historically, the word evaluation came into popular usage in education in the 1930's. The years following World War I were the years of tests and measurement, of individual differences, and selection and classification --the development of standardized achievement tests, group tests of intelligence, the measurement of interests, ability grouping in the schools, and psychometrics as a special field of knowledge and theory. Until sometime in the 1930's the word was measurement. Evaluation, as a term, came into being to express a broader concept. An article by Irving Lorge entitled "Evaluation: The New Stress on Measurement" seems to me to have symbolized the change. One of the most rigorous, tough-minded experts in measurement had taken notice. While Lorge used the word stress in the sense of "emphasis," in retrospect

he might better have used the word in the sense of "tension," for evaluation emphasized the inadequacies of measurement. Evaluation accepted and welcomed the use of observations, interviews, check lists, questionnaires, testimony, the minutes of meetings, time logs, and many other relevant means of assembling information. It included measurement and sought to extend its range, but it was more than measurement. It included psychometrics but held that psychometric theory was irrelevant in many evaluation activities. Moreover, evaluation freed itself from the arbitrary restrictions of the experimentalist's preoccupation with research design and hypothesis testing; for many programs and activities could not be accommodated by the rationale of the experimentalist.

Looking back at what we were doing in the decade before World War II, it seems to me that evaluation, as a new feature, had a more missionary emphasis than a broader scientific emphasis. Evaluation was seen as an instrument of reform. There was, of course, emphasis on balanced judgment of results, but there was also an emphasis on the process of evaluation. Evaluation was both an act and a result. Evaluation became related to group dynamics, action research, self-improvement, and to other "movements" concerned with the processes of change and betterment. The reason for evaluating any present activity or program was to improve it. Obviously, the participants in the activity or program had also to be involved in its evaluation because their very involvement would increase the likelihood that they would be willing to change in the light of the findings from the evaluation. They might reject the conclusions from someone else's evaluation, but they would act on the conclusions from their own evaluation. Thus, the process of carrying out an evaluation was directly related to achieving

the purpose of evaluation, namely, the purpose of change and improvement.

There was still another emphasis, preceding that upon group process; and this was the emphasis on objectives--specifically the objectives of educational programs. Evaluation was concerned with how well instructional objectives were being attained. This concern was codified in a standard procedure or process. Ralph Tyler outlined the process of evaluation as (a) identifying general objectives, (b) specifying these objectives in behavioral terms, (c) identifying situations in which the behavior could be observed, (d) devising and applying instruments for making the observation, and (e) relating the obtained evidence to be professed objectives. As this process was applied, it was evident that the clarity of objectives and the relevance of measures had a direct impact on the clarity and relevance of instruction. Thus, evaluation was a way to improve teaching.

In the years immediately after World War II the emphasis was on "self-study." For the most part, this was an emphasis on aims, an effort to clarify new goals and new directions for a new era and new challenges. In some circles the word "evaluation" was in disfavor.

From the mid 1950's to the present, there has been a re-emergence of evaluation. The contemporary stress on evaluation is a response to emergent technologies and social problems. With respect to technologies, there is a concern for evaluating computer-based instruction, instructional programming, TV teaching, new instructional products, and new instructional curricula. With respect to emergent social problems, there is a concern for evaluating the consequences of such large-scale programs as Head Start, Title I and Title III programs, and other activities which attempt to deal with equal opportunity, integration, the

disadvantaged, etc. There are also other current interests, such as new administrative and organizational aspects of the schools, and similar "innovations"; but these are mainly adaptive responses to the same basic events, namely, new technologies and new social problems.

I would like now to illustrate this history, primarily with examples from my own experience in the evaluation of higher education, from the emergence of the term evaluation in the early 1930's to the present day.

The difference between evaluation and traditional measurement was very apparent in the studies we made of the General College program at the University of Minnesota in the 1930's. We not only constructed achievement tests for the various General College courses; we also devised ways of getting at other sorts of outcomes. We looked at changes in the students' educational and vocational plans and judged whether they were more realistic. We looked at their personal adjustment, their attitudes toward home and family, and their opinions about contemporary social, political, and economic issues. We asked about their satisfaction with the college. We kept track of newspapers, magazines, and books they read in a specially devised reading room.

We thought of evaluation as contributing to program planning as well as measuring program outcomes. Consequently, we made intensive studies of the characteristics of the students and their backgrounds--their abilities, interests, problems, and experience--so that curricula, counseling services, teaching, and other aspects of the program could be related to student needs. We tried to analyze contemporary society and economic and occupational trends so that courses could be developed to deal with such issues. We studied the lives of young men and women

who had entered the university some years previously to see what we could learn from these young adults about problems our present students would shortly be facing.

No set of published tests and measurements available in the 1930's would have served our purposes adequately. We had to go beyond traditional measurement to get the data we wanted--to devising questionnaires and interview schedules, new attitude and opinion tests, check lists of behavior, rating scales, and many other kinds of inquiry which are common today but were uncommon thirty years ago. Because of our interrelated interests in program objectives, program planning, and program outcomes, the word "evaluation" was more appropriate to our activities than the word "measurement".

The concept of evaluation as intimately related to the objectives and improvement of instruction was most clearly illustrated by the work of Ralph Tyler and Benjamin Bloom when each was, successively, head of the Examiner's Office at the University of Chicago in the 1930's and 1940's. The faculty members attached to the Examiner's Office spent a great deal of their time defining the objectives of their courses. What were the professors meaning to teach? What were the students expected to do at the end of the course? What opportunities were the professors giving the students that would enable them to achieve the course objectives? If they were expected to acquire knowledge, was it knowledge of facts? of terminology? of methods? of principles? or of what? Should they be able to apply knowledge and principles to new problems? If so, then there must be items in the test which required this kind of behavior. This emphasis on the clarity of objectives and the corresponding relevance of test items subsequently led Benjamin Bloom and other colleagues around

the country to construct the Taxonomy of Educational Objectives, a taxonomy which was equally relevant for the classification of objectives and the construction of test items. As teachers reviewed the results of their course and comprehensive exams at Chicago, they could learn not only which students made the highest scores but also which objectives had been most fully achieved. They could then modify their teaching in an effort to increase the attainment of certain objectives. Evaluation was a cycle which involved clarifying objectives, measuring the attainment of objectives, and adapting teaching methods and materials to facilitate the better attainment of objectives. This cycle of continuous evaluation was a powerful method for the improvement of curricula, the improvement of instruction, and the improvement of testing.

The evaluation activities of the Commission of Teacher Education in the years 1939-1944 illustrate the emphasis on group process and self-improvement to which I referred earlier. Looking back over the book, Evaluation in Teacher Education, which Maurice Troyer and I wrote for the Commission, this emphasis is quite clear. The following quotes are pertinent:

> "Why do we evaluate? One very clear reason is in order to judge the effectiveness of an educational program. The unit for evaluation may encompass the total offerings of a college; it may be a single course, or it may be a fairly coherent aspect of a total program--such as general education, student teaching, or orientation and guidance. We undertake to evaluate the program because we hope thereby to improve it. By knowing its strengths and weaknesses we are enabled to plan more intelligently for its improvement. Similarly, we may evaluate the progress of an

> individual--ourself or someone else. And again, we do it because we hope thereby to advance progress, to attain greater success because we have found out what was holding us back. We know that knowledge of results aids us in learning new skills. So likewise, an evaluation of our status and progress helps us to improve the status and to make further progress. By analyzing our experience, resources, and programs we help to clarify them and to bring our efforts more directly in line with our purposes. Thus, evaluation is a technique that can and should lead to the continuous improvement of education" (p. 2-5).
>
> "Evaluation is of little worth unless the weaknesses it reveals are corrected. All evaluation reveals weaknesses as well as strengths. Who is to correct these weaknesses? Quite obviously, the students must correct deficiencies that apply to them, and the staff must correct deficiencies that apply to the educational program. But will they? They may not. They may produce an elaborate set of arguments to prove that the evaluation was untrustworthy, that the evidence it gathered was suspect and invalid. They are not likely to react in such manner to an appraisal which they have themselves carried out. That is why evaluation, to achieve its purpose, must be so conducted that confidence in the results is built up and readiness to change is fostered. Participation, making evaluation a genuine group enterprise, is one effective means of assuring that results will be put to good use" (p. 367-368).

It is important to note that the title of the book was Evaluation in Teacher Education. Our role in the Commission was not to make an evaluation of teacher education. But the very adoption of a consultant role by the Commission was exactly what led to the awareness of group processes and cooperative procedures throughout the Commission's work. The experience and the report of it helped to establish a connection between evaluation and reform, evaluation and self-improvement.

The emphasis on self-evaluation was continued after World War II. But it was not called self-evaluation; it was called self-study or self-survey. We had a rather large self-survey at Syracuse University in 1947-48. It involved a hundred or so faculty members, administrators, and trustees, organized into survey committees to appraise the present state of the university curriculum and instruction, graduate study and research, personnel services, faculty, library, plant, finances, administration, etc., and to recommend changes and directions of growth. One of the trustees had suggested that the university prepare a ten-year projection of its financial needs; whereupon, the Chancellor pointed out that one could make such a projection only in relation to the kinds of programs, services, and activities that needed to be financed. Hence, a broad look at the goals, resources, and operation of the university was really prerequisite to any long-range financial planning. Syracuse was one among many institutions which felt a need to "take stock" of itself in the postwar years. They were the years of rising enrollments and financial strains, of returned veterans whose perspectives and purposes were not quite the same as those of the typical undergraduates.

One year the Ford Foundation made grants to 21 colleges and universities to conduct self-studies. I was asked to help appraise what these grants accomplished. I visited a few of the institutions after they had completed their self-studies and then with others read the reports which all of the institutions submitted to the Foundation. We wrote a staff report and analysis (unpublished) for the Foundation.

In our analysis of the self-studies we concluded that self-study problems, procedures, and results were

in a broad sense interrelated. We felt that the extent to which significant change occurred depended on the breadth and intensity of participation in the self-study process. Participation of a level sufficient to give rise to strong feelings of personal involvement in the topic that was studied and of such a character as to make everyone who was concerned about the results not only aware of the progress being made but also a contributor to that progress had to be planned and built into the design of the self-study procedures. But it was difficult to obtain participation at this high level of involvement unless the group considered the problems to be sufficiently crucial and challenging--and capable of better solutions than those current on the campus--to warrant the time and energy of working on them.

Because of these relationships among topics, procedures, and results, we suggested that the total self-study process involved a series of choices and decisions, grouped around five phases of work, which occurred in a time sequence, as follows:

1. The decision to undertake a self-study
   a. What are the conditions of readiness for such a study?
   b. Who makes the decision to have a self-study and how is it made?
   c. How is the topic or focus for self-study chosen?
   d. What is expected to be accomplished?
2. Decisions in the planning stage
   a. Is there a conscious planning stage prior to active study?
   b. How sharply are the problems and purpose of the study defined?
   c. What methodology is decided on?
   d. How is responsibility for carrying out the study assigned?
   e. Is there a time-table for the whole enterprise?

3. Decisions in carrying out the study
   a. Are the various working groups coordinated?
   b. What provisions are made for extensive participation?
   c. Are the techniques of inquiry and analysis working out effectively?
   d. Are there progress reports?
4. Decisions in reporting the results
   a. To what audience is the report directed?
   b. Is it a single unified report?
   c. Is it a readable and persuasive document?
   d. Does it lay the groundwork for further activity?
5. Implementation and follow up
   a. What machinery is established to facilitate study and action on the self-study report and its proposals?
   b. Is the self-study a beginning or an ending?

It is quite evident from this analysis that our primary concern was with the process of self-study. This conclusion did not mean that we were unconcerned with the product; quite the contrary, it indicated an awareness of the fact that the validity of the product of self-studies was related to and in part dependent upon the validity of the process.

In 1954, the Committee on Evaluation and Measurement of the American Council on Education, of which I was a member, proposed to the Council that a publication on the topic of college self-evaluation should be produced. Such a publication, the committee thought, should meet the following objectives:

1. It should point out the widespread current activity in the field, relating such activity to its historical development and other influences.
2. It should discuss the nature of the self-evaluation process with its necessary concern for philosophy, measurement, and human relations.

3. It should illustrate good practices in self-evaluation that have been found effective, including techniques of data gathering, practices relating to effective group activity, and ways of clarifying goals or objectives.

4. It should summarize some of the conditions which are most likely to result in effective self-evaluation, drawing such generalizations from experience and research as appear to be valid.

5. It should make available a selective, annotated bibliography of relevant literature on college self-evaluation.

6. It should encourage thinking about the rationale and the broad design of effective self-evaluation so that whatever individual colleges may elect to do they may see their activity in some larger perspective.

In 1954 also, I submitted a modest request to the Behavioral Sciences Division of the Ford Foundation for support of a proposal entitled "Research and Development on Improved Designs and Methodologies for Institutional Evaluation and Self Study."

I suggested that the proposed project might have the following outcomes:

1. A thorough and critical appraisal and integration of previous experience and procedures in institutional evaluation;

2. A systematic drawing together of research and concepts from psychology, sociology, education, and related disciplines which bear upon the methodology, design, and productivity of self-studies;

3. The development of models and research designs for institutional evaluations;

4. The imposition of these designs upon previously conducted evaluations as a means of estimating their value;

5. The trial of new methods and designs in connection with currently active self-studies with judgments as to their efficiency;
6. Integration of all the above into one or more broadly applicable patterns for self-studies designed to enrich their usefulness, contribute to science and education, and provide methods and concepts with which future advances in the conduct of self-studies can be built.

I will not bother to apologize for the occasionally pretentious language of these proposals, but I will add two short footnotes: (a) the American Council did not obtain funds for the proposed book and (b) the Behavioral Sciences Division of the Ford Foundation was abolished.

My reason for citing these post-World War II activities in some detail is to give the basis for the generalization which I now want to make. Although what I earlier called a missionary emphasis has persisted and is still evident today, there was beginning to emerge a more scientific emphasis -- that is, a concern for cataloguing the evaluation process, for explaining why certain procedures were effective and others were not, and for trying to avoid the all too common ad hoc character of self-studies. In a sense one could regard this attempt to generalize and explain as no more than a concern for how to be a better missionary! But it was not that. We were really trying to make better, more systematic, and more reproducible evaluations.

During the past decade a good many events have reenforced an emphasis on science in our society. Let me comment briefly on two of them as they related to evaluation in education.

The first is the development of instructional products and technologies. The new curricula in math and

sciences and other fields have developed largely from analyses of the underlying structure of knowledge and concepts in these fields. What needs to be known at the most elementary level before knowledge at more complex levels can be meaningful? How are the essential concepts in the field built, one upon the other? The technology of programmed instruction, whether presented in a book, a sequence of pictures, a teaching machine, or a computer, is basically an effort to combine in the most efficient manner our knowledge of the psychology of learning and the hierarchical structure of what is to be learned. The emphasis on learning has its corollary emphasis on the specification of objectives in explicit behavioral terms. The emphasis of instructional materials and devices has its corollary emphasis on product testing to determine the effectiveness of these materials and devices. Consequently, the development of new instructional products and technologies has re-enforced the concern for experimental testing and evaluation. This kind of evaluation is basically a matter of hypothesis-testing in which adequate research design is essential.

In a quite different way, the need to evaluate large-scale social programs has also re-enforced the importance of science in evaluation. It might not seem so to the experimentalist. So-called field studies in naturalistic settings do not permit random assignment of subjects to treatments or many of the other standard procedures of the experimental psychologist and his counterpart in education. But the evaluation of such large-scale programs as Head Start or a total school district or other major federal and state activities does not require multivariate analysis of great complexity and scientifically objective estimates of consequences and

benefits rather than pious hopes for social betterment. Models and methods from the social sciences of economics and sociology are being applied to these large program evaluations, as illustrated in the reports of such analysts as Daniel Moynihan and James Coleman.

Thus, both the rise of educational technologies and the rise of large scale social problems demanding attention have, in different but complementary ways, given the concept and content of evaluation today a more objective social science emphasis, at least in comparison with the 1950's when the content and procedures of evaluation were largely influenced by the view of evaluation as a strategy for generating reform. Moreover, it is on the results, not the process, that today's scientific emphasis is placed.

At the beginning of this paper, I said that the way in which we think about evaluation and how we go about making evaluations are necessarily related to what we are evaluating and why we are evaluating it. The history I have reviewed provides one kind of perspective and illustration of this statement.

Another kind of perspective may be gained by attempting to classify the variety of evaluation activities one finds today in a way that acknowledges the validity of each and the validity of differences between them.

The most important classification, in the sense that its ramifications are extensive and obvious, is one that relates to the size, complexity, and duration of what is to be evaluated. Consider the following contrasting cases.

> <u>When</u> the unit to be evaluated is a small unit--small in size, limited in scope, and short in time--such as a half-hour film, a specific unit of instruction in a single course,

a particular method of teaching, or a programmed text,

Then, the following conditions are usually true:

1. The treatment (unit to be evaluated) can be clearly and explicitly defined;
2. The treatment can be compared with alternative treatments or control groups;
3. The requirements of experimental design involving random assignments of subjects to treatments can usually be met;
4. The assumptions for statistical tests of significance, appropriate in a hypothesis testing experiment, can usually be met.

Under these conditions, relevant evaluations can be:

1. Directly related to behaviorally defined objectives;
2. Designed as a hypothesis testing experiment;
3. Largely limited to the intended effects of the program or treatment.

In contrast:

When the unit to be evaluated is large, complex, and of long duration--such as a school system, a total institutional program, or higher education in the U. S.,

Then, the following conditions are usually true:

1. The treatment (unit to be evaluated) cannot be clearly and explicitly defined because it is not in fact a unitary phenomenon but is, instead, made up of many units interacting with one another

in varied ways and having varied purposes;

2. Gross differences between treatments can sometimes be found and compared, but control groups in the usual experimental sense do not exist;
3. Random assignment of subjects to treatments is impossible except occasionally in some small segment or limited part of the larger treatment;
4. Treatments are constantly undergoing change (no collegiate institution could or would freeze all of its procedures and programs for four years so that the conditions for an evaluation of them would remain stable).

Under these conditions, relevant evaluation:

1. Must consider a broad range of educational and social consequences;
2. Should never be limited by or confined to the stated objectives or intended effects of the program or treatment;
3. Should look for but may not always find contrasting conditions in natural settings for comparative analysis;
4. Must employ complex multivariate methods of treating data.

Also, as the unit or program to be evaluated becomes larger, the contexts within which the program operates--contexts such as organizational and administrative conditions, the relation to other programs within the school or system, the nature of the clientele and the community, the financial

> resources and their allocation, the atmosphere of the school--have a greater opportunity for influence; and it becomes crucial to include a range of such potentially relevant contextual variables in one's evaluation design.

As size, scope, and duration change--from small to large, simple to complex, short to long--there are corresponding changes in the nature and the procedures of evaluation, relevant to the differing conditions.

One can also classify and suggest the implications of different concepts of the role of the evaluator and the purpose of evaluation. Again, consider the following contrasting cases.

> When the evaluator is basically a teacher, reformer, or staff officer to the practitioner and the purpose of evaluation is to improve or change a program or practice,
>
> Then, the process of evaluation is characterized by:
>
> 1. A client-centered orientation--in that the clients specify the objectives (usually with help from the evaluators);
> 2. A cooperative mode of inquiry--in that the clients or practitioners, in addition to the evaluators, plan, conduct, and interpret the inquiry.
>
> The intended result is decision and action.
>
> But when the evaluator is seen as a neutral social scientist and the purpose of evaluation is information and analysis,
>
> Then, the process of evaluation is characterized by:
>
> 1. An independent orientation--in that the range of inquiry includes but is

not limited to the client's intended objectives;

2. A collaborative mode of inquiry--in that expertise from relevant disciplines is brought to bear on the design, conduct, and analysis of the inquiry.

The intended result is the provision of more complex bases for informed judgment.

Thus, the characteristics of good evaluation differ depending on what is being evaluated, why, and by whom. Evaluation cannot be described by a single set of rules. Evaluation is, indeed, pervasive. To see it in all its variety requires a perspective that puts different purposes and procedures in some proper arrangement. Historically, evaluation has been regarded both as a process and as a result. Today there remains a similar difference in orientation. To the extent that the current importance of science in our society has influenced thinking about evaluation, its influence has been toward independence, objectivity, and results rather than toward processes of cooperative participation. In my own perspective, this more scientific emphasis is long overdue.