DOCUMENT RESUME

ED 117 190 95 TM 005 050

AUTHOR Severy, Lawrence J.
TITLE Application of the Experimental Method to Program Evaluation: Problems and Prospects. TM Report 47.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-47
PUB DATE Nov 75
CONTRACT NIE-C-400-75-0015
NOTE 38p.
AVAILABLE FROM ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, N.J. 08540 (free while supplies last)

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS Formative Evaluation; Information Dissemination; Program Development; *Program Effectiveness; *Program Evaluation; *Research Design; *Research Methodology; Research Problems; *Scientific Methodology

ABSTRACT

The purpose of this paper is to: (1) discuss the reasons for applying the experimental method to program evaluation; (2) review the basic elements of the experimental method; (3) illustrate refinements and variants of the experimental method with examples from program evaluation in the social sciences; (4) describe ways in which funding directives can be adapted to the experimental method; and (5) discuss potential problems connected with the experimental approach. Thus, the intent is to present a comprehensive, but nontechnical, discussion of the issues surrounding program evaluation and the experimental method to help administrators, educators, graduate students, novice researchers, and program and project directors conceive of ways to effectively apply program evaluation to their own programs and other endeavors. (Author/RC)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 47 NOVEMBER 1975

APPLICATION OF THE EXPERIMENTAL METHOD TO PROGRAM EVALUATION: PROBLEMS AND PROSPECTS

Lawrence J. Severy

Introduction

In the United States today, an extensive network of social programs touches on virtually every aspect of our lives. Innumerable federal, state, and local health and rehabilitation services and education and welfare programs operate in such a way that they have some impact on all of us. In education alone, administrators continually face decisions that will affect hundreds of lives: Which programs should be altered or done away with? Which new and innovative programs should be implemented? How should they be run? How should they be funded? The business world faces equally difficult decisions. For example, because of today's energy problems, the auto makers in Detroit face decisions regarding programs aimed at the production of small economy models versus programs designed to help stimulate the sale of larger models (accomplished either through advertising programs or alterations in the current models). At any rate, in all of these programs, those in positions of authority must make decisions as to when to start, when to change, and when to terminate any program. Such decisions are often difficult because they usually affect the lives of many people.

Whenever a school administrator is faced with such a decision, he or she wants, of course, to choose the "right" option. Should the local school board opt for one busing plan over another, implement a head-start program rather than a follow-up program, move the unit for educable retarded citizens from one school to another, or endorse a math program or a reading program for middle-school children? Whatever the nature of the question, those who must decide want as much information as possible before making a decision. The process of collecting data for use in answering questions regarding programs is known as program evaluation. More specifically, when social scientists go about collecting information regarding the effectiveness of a particular program, or variety of models of a program, they are engaging in program evaluation. It shall be the purpose of this paper to: (1) discuss the reasons for applying the experimental method to program evaluation; (2) review the basic elements of the experimental method; (3) illustrate refinements and variants of the experimental method with examples from program evaluation in the social sciences; (4) describe ways in which funding directives can be adapted to the experimental method; and (5) discuss potential problems connected with the experimental approach. Thus, the intent is to present a

comprehensive, but nontechnical, discussion of the issues surrounding program evaluation and the experimental method to help administrators, educators, graduate students, novice researchers, and program and project directors conceive of ways to effectively apply program evaluation to their own programs and other endeavors.

Why the Experimental Method?

Consider a well-known event in this country's history. In 1954 the United States Supreme Court declared in Brown v.s. Board of Education that segregated schools were unconstitutional and that separate but equal schools were by definition unequal. Presiding Chief Justice Earl Warren stated that separating black children from other children "generates a feeling of inferiority as to their status in the community that may affect their hearts and minds in a way unlikely ever to be undone." Many have felt that this decision led to the most far-reaching and significant social experiment in the history of the country. Would forced integration--that is, forced interaction between members of a majority and minority group--have the desired effect of providing the optimal educational environment for all concerned? As is well known, that question has been argued ever since the decision was made and, for a variety of reasons, the answer has always been somewhat equivocal (11,47).

Although, the Supreme Court's decision may be one of the most important legal actions in our social history, it does not stand as an example of the scientific method applied to the evaluation of a program. If social scientists, in the role of program evaluators, instead of judges, jurists, lawyers, and so forth, had been charged with confronting this particular social problem, they would have proceeded in a vastly different fashion. This is not to say that the decision itself was an incorrect one; however, the Court's decision-making method by no means exemplified the scientific method of evaluation. How, then, would a group of social scientists proceed in a program evaluation of this kind?

First, they probably would try to define the nature of the problem as precisely as possible. In other words, they would attempt to identify and clearly state the number and types of factors adversely affecting either the environment, performance, or learning of the individuals involved. For example, it has been suggested that elementary school children might respond to forced integration in a fashion completely different from high school students, who have been in separate schools for a long period of time. Consequently, the social scientists might try to differentiate integration programs according to grade level. Similar delineation would continue until all relevant factors potentially affecting outcome are identified, and the resulting information would probably be used in the design of an experiment.

The social scientists would also concentrate on factors they hope to change by using different programs. Just what is meant by optimal educational environment? Is one only interested in the precise amounts of knowledge or content acquired by the students, or is one interested in other things as well? It is possible that in attempting to completely delineate and define the nature of this particular social issue (desegregation of schools), one might decide that the amount of information acquired by the children from the minority and majority ethnic groups is of only partial importance, and that the nature of the interpersonal behavior and interpersonal attitudes between these groups is of equal importance. Behaviors

and attitudes may reflect racial tension. Social planners might feel that the most effective program would be one that not only increases academic achievement but also reduces racial tension and prejudice.

The social scientist begins by attempting to identify as precisely as possible any factors that might be causing certain problems and then states as precisely as possible the nature of the required changes or the exact results he wishes to obtain.

Secondly, the social scientist probably would conceptualize a variety of approaches or programs and might implement all of them to allow for a comparison that would show which approach is best for creating the kinds of social changes required. A special case program would include a control group--a group subjected to no change whatsoever. In other words, while the social scientist implemented a variety of programs aimed at social change, he would also monitor changes that came about spontaneously. At any rate, the social scientist would probably schedule the implementation of various programs at particular times. This approach would allow the scientist to carefully analyze the effectiveness of each program and systematically vary certain aspects of each program. The scientist always tries to put himself or herself in a position to state clearly and accurately which factors are causing which changes. A most crucial component of the social scientist's approach, then, would be the design and implementation of the program. In addition, the scientist would carefully choose the appropriate statistical procedures to be used to determine the effectiveness of the various programs.

Finally, to ensure that the study results would have the greatest impact on the greatest number of people, the social scientist would probably bend over backwards to disseminate the findings to persons who are in a position of authority and have the power to make decisions about programs.

As previously stated, this was not the approach taken by the judges on the Supreme Court. They had to make an important decision and could not take the time to develop a variety of approaches and then look at the results and force school districts to adopt the best one. They were obliged to choose one position based solely on moral, ethical, and legal considerations. This is not usually the case, however, and when circumstances permit, program evaluation and the scientific method should be employed. In fact, the position to be taken in this paper is that the scientific method should be viewed as an ideal; by approximating this method as precisely as possible, one would be in the best position for choosing the most effective program. There is no question that this is not always possible; there are some inherent problems connected with the experimental method that will be discussed later. But the initial approach should be to understand how the experimental method can be applied to education and the social sciences in general; hold it up as a model or an ideal; see how well it fits; try to implement it as fully as possible; and then, only if necessary, turn to other methods.

Experimental Approaches

The importance of the experimental method to educational reform cannot be overstated. Its significance has long been recognized. As Campbell and Stanley (9)

point out, as early as 1923 W. A. McCall published a book entitled How to Experiment in Education. Enthusiasm for experimentation in the field of education increased greatly in the Thorndike era and perhaps reached its apex in the 1920s. However, disillusionment did come. Campbell and Stanley feel that several important factors explain this disillusionment. First, "claims for the rate and degree of progress which would result from experimental approaches were grandiosely optimistic and were accompanied by an unjustified appreciation of non-experimental wisdom." In other words, many people felt that the technology of teaching would be speeded up very quickly simply through adoption of the experimental method. In fact, experimentation takes time, and when the field didn't progress as fast as expected, instead of being disillusioned with teaching and its technology, experimenters and participants became disillusioned with experimentation. At any rate, in their 1966 work Experimental and Quasi-Experimental Designs for Research on Teaching (originally published in 1963 as a chapter in the Handbook of Research on Teaching), Campbell and Stanley reaffirm a committment to the experimental method as "the only means for settling disputes regarding educational practice, as the only way to verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of faddish discard of old wisdom in favor of inferior novelties" (page 2).

Before describing the ways in which social programs can be conceived and implemented by means of the experimental method, it is advisable to review some of the basic principles of the experimental method itself. This will be done in the context of discussions about two aspects of this method: first, the basic experiment; and, second, the quasi-experimental approaches.

Basic Experimental Method

A hypothetical teacher in a town in the Rocky mountains has a problem. She teaches mathematics to tenth-grade students in one-hour classes. In some of her classes, she has absolutely no behavioral disruptions from the students. She is able to proceed with her lesson plans, has very satisfactory interaction with the students, and generally has good rapport with them. On the other hand, there are some classes that she simply cannot contend with. She does not seem to be able to "stay on top of the class" and finds that she cannot teach the students very much. This bothers her a great deal, and as she thinks about it, she decides that there are a couple of factors that seem to determine the amount of disruption in these classes.

One factor seems to be the number of students in the classroom. Some of the classes have a large number of students, lets say 45, and others have only 20 students per class. She thinks, but is not sure, that the crowded classes have more disruption than the uncrowded classes, but that isn't always the case. The second factor she thinks that might be affecting behavior is, believe it or not, the temperature of the classroom. Recall that our hypothetical teacher lives and teachers in a hypothetical town on top of the Rockies. It's winter time, and at this high elevation tremendous swings in temperature occur throughout the day. On clear days, when the blue sky shines down, the temperature at the middle of the day gets well above 50 degrees. However, when the sun is down, or when it's a cloudy day, it's quite cold outside. In fact, at night the temperature goes well below zero. As the day starts out, before the furnace is fully on in the

school building the classroom is very cold. As the furnace chugs away, things finally get quite comfortable. On sunny days, when both the furnace and the shining sun are operative, classrooms get extremely warm. She's not sure exactly how temperature is affecting behavior, but somehow it seems to.

Because the teacher is aware of these two factors and is aware that they may be affecting behavior, she has the basic ingredients for performing an experiment. She has the hypothesis that the temperature of the classroom and the extent of crowding are related to the amount of disruption in the classroom. Experimentalists would say that temperature and crowding are the independent variables in this situation and that the dependent variable is the amount of disruption. That is, the amount of disruption hypothetically depends on the level of the temperature and the extent of the crowding. Experimentalists would go about setting up a plan for varying (or manipulating) the temperature and the extent of crowding and then note as precisely as possible whether or not there are changes in the dependent variable--in this case, the amount of disruption. To recapitulate: in experiments, scientists attempt to manipulate or vary independent variables and then note whether or not these manipulations create changes in the dependent variables.

Let's suppose that our teacher has set up such an experiment so that she has six classes--three that are crowded, and three that are not crowded. In each of the two sets of three classes (see Figure 1), one of the classes can be seen to take place with what we might describe as a low room temperature, another at a normal temperature, and a third at what may be described as a high temperature. If the teacher has one class representing each of these conditions, she would be said to have a complete factorial design (one factor being temperature and the other factor being crowded versus noncrowded); every level of one variable is completely crossed with every level of the other variable. She has "manipulated" these conditions and now sits back and counts or notices, as precisely as possible, the amount of disruption that occurs in each of the six conditions.

TEMPERATURE

DEGREE OF CROWDING	Cold	Normal	Hot
Not Crowded	20 Students Below 65°	20 Students 66° - 85°	20 Students Above 85°
Crowded	45 Students Below 65°	45 Students 66° - 85°	45 Students Above 85°

FIGURE 1. Factorial design for the study of disruption as a function of crowding and temperature in the teacher-researcher's six classes.

There are several possible results with this type of setup. To start, the teacher may find that there are absolutely no differences among any of the six classes. It might be that she has to conclude that temperature and crowding do not affect the amount of disruption. Another possibility is that either certain temperatures or crowding or both may be found to increase disruption. For example, it might be found that the warmer it is, the more problems there are, or, the more people there are in the classroom, the more problems there are. Such a result would be known as a main effect. In other words, a main effect for temperature would mean that regardless of crowding or any other conditions, the warmer it is, the more behavioral disruption there is. If both factors were shown to be important, it is very likely that the two factors would combine in such a way as to demonstrate that temperature and crowding were additive in nature. In other words, if both increasing temperature and increasing crowding were shown to affect behavior, it is likely that less disruption would occur in cold, noncrowded classrooms than in hot, crowded classrooms.

However, another possibility exists for a significant interaction when the two factors are combined. It might be that looking only at the main effects of temperature and crowding does not provide an answer as to what is happening. It may be that one needs to incorporate both factors in a precise description of the teacher's classroom problem. Suppose that temperature does not produce a main effect, and there are just as many problems in cold rooms as there are in hot rooms. And suppose there are just as many problems in crowded rooms as non-crowded rooms so there is no crowding main effect. But when the two factors are combined, an interesting interaction may occur. Consider the following possibility: It is possible that in four of the six classes our teacher has no problem. Whenever the temperature is normal--not too hot and not too cold--it does not matter if the classroom is crowded or not crowded. However, when the classroom is cold, it's possible that the crowded class gets angry and the non-crowded class does not. Further, when it's extremely hot, the crowded class may create problems and the noncrowded class may not. In other words, just looking at temperature alone will not provide the answer. In such cases, the interaction of the two variables is what creates the effect. Before looking at some of the problems and alternative explanations for the results described in our hypothetical example, let's consider another example.

A teacher is working with young, retarded citizens who have been described as educable. She is attempting to work with a reading program and is getting a bit frustrated because she cannot keep the students' attention. She feels that if she can't get their attention, she cannot teach them the content she is interested in getting across. She feels that unless she can get their attention, she won't make readers out of them. She searches her mind for various ways of getting their attention and considers the possibility that certain drugs may be beneficial for her purposes. She wonders whether one of several different drugs might have the effect of calming the students down so they will listen more closely to what she has to say. She has in mind several drugs, and after consultation with pharmacological behaviorists and physiologists, she has decided that there are three different drugs that might have the effect she is looking for.

There is a question in her mind, however, as to whether or not the age of her students would somehow interact, as described above, with the effectiveness of the different drugs. If she were to conduct an experiment, then, she would set about to evenly divide her students into several different groups. One of the factors used for dividing the students would be age. Possibly she would have children younger than 10 in one group and children 11 or older in the other. She would then divide each of the two groups into four different subgroups. She would want to try three different drugs or possibly three different dosages of one drug. She would be said to have three treatment groups and one control group. The important distinction between this example and the first one is that, whenever possible, experimenters like to include in their experiments a standard or baseline against which to compare the effectiveness of their treatments or their manipulations. In this case, then, our experimenter has designed an experiment in which the independent variables are age and drug treatment (either method or dosage) and the dependent variable is to be the length of the attention span for the children. This particular teacher could then assess whether or not age alone accounts for differences in attention span; whether one drug or another, or particular amounts of the same drug, can be seen to affect attention span; or whether age and the drug dosage combine or intereact to produce differences in the attention span.

Both of these examples should help make it clear that one needs to define the independent variables as precisely as possible and manipulate them accordingly. However, it is just as crucial to note as precisely as possible what is to be considered the dependent variable. In our second example, the teacher was assuming that attention span might vary with drug dosage and age and that this attention span would be directly related to how much she could teach the students in her reading program. Suppose she found that there were no differences in attention span under any of these conditions. In that case, the suggestion could be made to her that what she is really interested in is the amount of material that she is teaching the children. Therefore, as a double check, it might have been advisable to measure the amount of learning that occurred under the various conditions. She would have had an alternative dependent variable. Suppose she found that a particular dosage of a certain drug did seem to allow more learning even though it didn't affect attention span. This result is not entirely impossible and, of course, would have created a variety of new hypotheses that could be researched.

By this time, the reader has probably thought of a variety of alternative and rival hypotheses for the preceding examples to show why the hypothetical results could have been obtained and how the conclusions that the teacher-researchers wanted to draw might have been completely wrong. Let's examine the types of rival hypotheses that could be generated and consider how they could or should be carefully eliminated by means of employing experimental designs.

As was mentioned earlier, the basic purpose of the experimental method is to put oneself in a position to state that the independent variables cause changes to occur in the dependent variables. If there is any other possible explanation for the outcome, resulting from the experimenter's design, then the researcher cannot state that the independent variables are causing changes in the dependent variables. For this reason, researchers attempt to control all other factors

that have the potential of affecting the dependent variable. In other words, if there are other variables that can be seen to affect the dependent variables, researchers attempt to make sure that they exist in the same proportions or same distributions in each of the experimental conditions. There are a number of possibilities here, and we shall discuss each in turn.

Consider the first example. It is possible that temperature in this hypothetical mountain town is directly related to time of day, and that when the classroom was cold, students were still a little groggy early in the morning. They were not actively attending to the task at hand. As they woke up, the temperature became normal, and only towards the end of the day, when they got tired and grumpy, did it become hot. In order to distinguish, then, between the effects of temperature and the time of day, our teacher-researcher would have had to make sure that there was not a direct relationship between the temperature and time of day variables. This could have been done by taking the measures of behavioral disruption on days that were overcast and cold so that temperature would not be related to the time of the day. A competent researcher would proceed in a similar fashion to identify as many factors as possible that could influence the dependent variable.

A definitive case can be made for controlling extraneous factors in experimental design in a discussion of random assignment of subjects to the various conditions. Consider the second example. If it had been found that the control group learned much more than any of the drug or treatment groups, the researcher might have been persuaded that drugs were not the answer and that she had not found the solution to increasing learning. Imagine the problems that would have been created, however, if subjects had been assigned to the various experimental and control conditions based on the living units within the institution. What if mongoloid children were in one unit, autistic children in another, brain-damaged children in yet a third, and so on. The problem is that the researcher would not really know if the different types of retardation were influencing the amount of material learned or the attention spans, or if the drug treatments were influencing learning and attention span. In this case, one would say that the drug variable was completely confounded with type of retardation. Similarly, in the crowding and temperature study, if students of high socioeconomic levels were assigned to the noncrowded condition and other students were assigned to the crowded conditions, crowding and socioeconomic status would be completely confounded, and the researcher would then have to change her conclusions substantially. In other words, it would appear that higher SES students were not affected by temperature while lower SES students were affected by both cold and warm temperatures and engaged in behavioral disruption.

The main point of this discussion is that the researcher must be very careful to eliminate all rival hypotheses and must also be sure to control for the systematic influence of any variable which is likely to influence the dependent variable. He or she must be very careful not to confound the desired manipulation with any other. Because it is not always possible to conceptualize all of the potential influences on the dependent variable, researchers adopt a procedure known as random assignment of subjects to conditions. (This procedure

might also be adopted in the interest of expediency, for it is often too time consuming to get pre-experimental manipulation checks on all of the potential variables that could be matched throughout an experimental design.) At any rate, the point is that by randomly assigning subjects to the various conditions, there will be no systematic bias throughout the experimental design. When a subject is just as likely to be assigned to the control group as any one of the experimental groups there is very little likelihood of a systematic bias in the design.

This, then, is the basic experimental approach. Researchers identify those variables that they conceive of as the independent variables. They design experiments in which they can manipulate these variables in certain conditions. They then note changes, if any, in the variables they conceive of as the dependent variables. The hope is to draw conclusions about cause and effect. In order to do this, researchers must try to control all other potential change-producing factors either through the design itself or randomization of subjects in the various conditions. Because such procedures are not always possible in the real world, offshoots of the experimental method known as quasi-experimental procedures have been developed.

Quasi-Experimental Approaches

A discussion of quasi-experimental designs can become quite technical, complex, and cumbersome, and only the basics of such approaches will be discussed here. The reader interested in the more comprehensive presentations should refer to at least one of a number of sources. The first, Experimental and Quasi-Experimental Designs for Research by Campbell and Stanley, was mentioned earlier. Many illustrations come from educational research, and this short volume can be described as a classic in describing quasi-experimental designs. A second source is entitled Quasi-Experimental Approaches, edited by Caporaso and Roos (10). This volume contains a number of chapters concerned specifically with the quasi-experimental approach. A third source is a very short but effective paper in the ERIC series, TM Report 30 entitled "Evaluation Designs for Practitioners," by Maurice J. Eash, Harriet Talmage, and Herbert J. Walberg (18).

Unfortunately for educators and social practitioners, it is often simply impossible to implement the basic experimental method in certain situations. However, it is sometimes still possible to schedule certain treatments and measurements and analyze data that would be useful in evaluating program impact. A true experiment may be completely untenable, but an approximation of an experiment may still be very worthwhile, especailly if appropriate pretests and posttests, comparable to those used in the basic experimental method, are used. Close approximations of the experimental design have been labeled quasi-experimental designs. The characteristic difference in quasi-experimental designs appears to be the absence of randomized assignment of individuals to different treatment groups. Quasi-experimental designs are usually used in situations in which it is simply unfeasible, illegal, or illogical to randomly assign subjects to different treatment groups. In addition, sometimes there is no possible way to manipulate the stimuli or the treatment groups, according to the traditional factorial design but the experimental stimuli may occur naturally with no active intervention on the part of the researcher. The quasi-experimental design is now often used in

such situations. The purpose of this approach is clearly stated by Caporaso. He suggests that all quasi-experimental designs "attempt to approximate or stimulate manipulation to provide controls for confounding variables, and to probe the data for causal dependencies" (10).

Rather than specifically delineating all of the quasi-experimental designs, this paper will briefly review the more popular approaches as described by Riecken and Boruch (48), to familiarize the reader with the scope of this kind of approach.

Interrupted Time Series Designs: Whenever one encounters data that is collected at definite time intervals, one can then employ time series analysis designs. In what is known as a single interrupted time series design, the researcher checks to see if the pattern or trend of measurements on a particular dependent variable changes across time or is altered because of some experimental manipulation or treatment. In effect, the question becomes: Does what appears to have been happening before treatment appear to change as a result of some manipulation or programmatic intervention? For example, if arrests for speeding in a particular state appear to be increasing year by year, one question might be: Would the passage of a new law that makes the crime more serious have the effect of changing that particular trend? To answer the question, the law would have to be implemented and arrest trends would have to be studied. A modification of the single interrupted time series design involves the comparison of one series or trend with a second series or trend. For example, one might compare the trends in two different states resulting from the passage of a particular law. Again, the researcher would look at the trend before and after the intervention, and could now also compare one treatment group with another or with a control group.

Pretest-Posttest Designs: These very popular designs involve the comparison of a measure of the dependent variable before manipulation with a measure of the dependent variable after manipulation. In essence, this design is simply a short form of the interrupted series designs. Instead of having the luxury of analyzing multiple points across time for the dependent variable, one has only two data points--namely, the points before and after manipulations have occurred. As with the time series designs, the researcher considers a single experimental group measured at two different times and hopes to find changes that are caused by the manipulation. A second approach involves a comparison group in the pretest-posttest design. The comparison group design has the advantage of allowing a comparison between one group that has received the treatment and another that has not. Unfortunately, with the quasi-experimental approach, one does not have the opportunity to randomly assign subjects to the two groups. As a consequence, any significant change in the treatment group might simply be caused by differential selection or the criteria used for forming the groups. However, this method is more effective than just looking at the results for a single treatment group before and after manipulation.

Posttest Comparison Group Designs: Sometimes researchers can only measure the dependent variable after a program's treatment. In such cases, one would want to compare such scores to those of a second "control" group. Unfortunately, in these designs, individuals are not randomly assigned to different groups, and one does not know what the dependent scores might have been before manipulation or treatment. All the researcher can do is document differences in the measures after the manipulation has taken place. Clearly, this is one of the weaker approaches.

Correlational Approaches: Because of their capacity to handle many variables at one time, correlational techniques are becoming more popular all the time. Yet, with the correlational approaches, the researcher is always uncertain about causality. However, there is reason to believe that correlational techniques can be successfully utilized to help bring into focus relevant variables for later experimental manipulation and analysis. There have been attempts to at least suggest preponderant causality with correlational techniques such as cross-lagged panel designs, but these designs, too, have their problems (10). In effect, cross-lagged correlational approaches involve analyzing whether the diagonal correlation between X1 and Y2 is stronger than between Y1 and X2, given that synchronous correlations are stable. If that is the case, the argument goes that X is causing Y. Again, however, cross-lags should only be used as an approximation and an indicator of variables that should be looked at experimentally.

This brief review of quasi-experimental designs shows that with each approach of this kind there are elements that the researcher would like to have more control over but simply cannot. These approaches are merely approximations of the experimental method; as was stated earlier the experimental method should be considered the ideal and should be used whenever possible. For those who would like to try the basics of the experimental method for program evaluation, a discussion of some of the concerns and principles of program evaluation follows.

Program Evaluation

As the senior U.S. senator from Minnesota, Walter Mondale, so aptly put it, "we must design methods for filling the gaps in our information and methods to process...information systematically. We must develop a coherent set of problem definitions, goals, and solutions...In short, planning and evaluation must proceed at national and local levels..." (43). Program evaluation is not really a new thing Time and motion studies have been popular in industry for years, and social programs have been evaluated since before World War II--more than 30 years ago. Perhaps the best way to begin a discussion of program evaluation is with some questions such as: What is it? What is its intent? What are its characteristics? Weiss (72) suggests that "the purpose of evaluation research is to measure the effects of the program against the goals it set out to accomplish as a means of contributing to subsequent decision-making about the program and improving future programming." Within such a definition, she suggests that there are four key features. First, there is the idea that a research methodology be used to measure the effects. Secondly, there is emphasis on outcomes in relation to specific goals (rather than on efficiency, honesty, morale, or

adherence to rules, for example) as evidence of the effects of the program. Third, Weiss suggests that the comparison of effects with desired goals stresses the use of criteria for judging the effectiveness of a program. And, finally, she emphasizes the subsequent decision-making and future improvement of programming as the social purpose of the program, and therefore of the evaluation.

Hyman and Wright (38) suggest that the basic method of program evaluation has five major components: (1) the conceptualization and measurement of the objectives of the action program and of unanticipated relevant outcomes; (2) the formation of a research design of the criteria for proof of the effectiveness of a program; (3) the research procedures themselves, including procedures for estimating and reducing errors in measurement; (4) procedures for dealing with problems concerning index construction and evaluation of effectiveness; and (5) procedures for interpreting the findings on effectiveness or ineffectiveness.

This attempt to characterize program evaluation by describing the components of the process is a helpful strategy. In another comprehensive approach provided by Weiss (71), the author suggests that there are five basic stages in evaluation research: (1) discovering, identifying, or delineating the goals of a program; (2) translating the goals of the program into measurable indicators of goal achievement; (3) collecting information (data) on these indicators from persons who have been exposed to the program (participants); (4) collecting similar information (data) from an equivalent group that has not been exposed to the group (control group); (5) comparing the information from the program participants with the information from the control groups in hopes of finding differences in terms of goal criteria, and therefore in success of the program.

Each of the foregoing listings describes procedures aimed at measuring the success of programs. It has been clearly stated by Suchman (64) that underlying all of these efforts at program evaluation are three important assumptions: First, man can change his social environment; second, change is good; and third, change is measurable. "Thus, social problems are viewed as amenable to deliberate intervention, while the success or failure of such intervention is subject to demonstration through scientific, evaluative research studies." In other words, as has always been true, in the final analysis, the social scientist tends to be an optimist; further, that optimism can be justified and demonstrated by means of scientific methods.

Before proceeding further, several statements are necessary for clarification. First, this section of the paper may leave the impression that there is only one thing that can be accomplished through program evaluation--namely, a final analysis of the action program. Although this is an important purpose of program evaluation, it is not the only one Scriven (54) has termed this activity summative evaluation. This type of evaluation provides decision-makers with information about the effectiveness of a particular program so they can plan for the future. On the other hand, Scriven identifies another type of evaluation called formative evaluation. Formative evaluation is designed to produce information that is fed back into the program for improvement during its development and thereafter. Similarly, Fox (25) speaks of <u>critical</u> evaluation, which determines whether a program is continued, and <u>ongoing</u> evaluation, which identifies strengths and weaknesses. Formative or ongoing evaluation should

allow clever experiment designers to incorporate past information into new manipulations. Formative evaluation clearly has more potential for immediate impact than summative evaluation.

Another clarification is provided by Suchman. He identifies four factors to be considered in an evaluation: (1) effort (the amount of action); (2) effect or performance (results of effort); (3) adequacy of performance (is effort sufficient for the total need?); (4) process (how an effect was achieved); and (5) efficiency (effects in relation to cause) (23).

At this point, it should be obvious that when one speaks of evaluating a program, such a statement clearly calls for further delineation. Program evaluation is a highly complex process.

A question that readers may be asking at this point is: How can evaluation research be distinguished from basic or nonevaluative research? First, the evaluation researcher typically has his subject matter given to him. He does not formulate his own hypotheses; they come from program goals(33). Another distinction is that there are very few variables over which the evaluator can exert control. However, the most significant difference is one of purpose (intent) and not of method. Both types of studies attempt to utilize research designs for data collection and analysis based upon the logic of the scientific method. Evaluative study applies this model to problems that have administrative consequences, while nonevaluative research is more likely to be concerned with theoretical significance. But the validity of both types of studies rests equally on the degree to which they satisfy the principles of scientific methodology (65). In other words, it is the purpose of the research that distinguishes evaluative from nonevaluative research (often described as basic theoretical research). At this point, then, it would be useful to identify some of the purposes of program evaluation.

Weiss (71) suggests that "unless and until the evaluator finds out specifically who wants to know what, with what end in view, the evaluation study is likely to be mired in a morass of conflicting expectations." However, defining precisely who wants to know what is no longer so great a problem because of new pressures for accountability for expenditure of the government dollar. (We will later return to the way in which approaches such as management-by-objectives can lead directly to a clear statement of program objectives.) According to Weiss (71), however, historically evaluations of the success of programs have been undertaken so authorities can decide whether to (1) discontinue the program; (2) improve its practices and procedures; (3) add or drop specific program strategies and techniques; (4) institute similar programs elsewhere; (5) allocate resources to competing programs; or (6) accept or reject a program approach on the basis of theory.

Weiss has also suggested that certain goals in program evaluation can lead to the utilization of evaluation research findings. These goals are: the delineation of the theoretical premises underlying programs; the sequence of linkages that lead from program input to outcome with the specific description of processes through which results are supposed to be obtained; and last, an

analysis of the effectiveness of the parts of a program rather than total go or no-go assessments. Further, effective utilization of evaluation research is likely to be a result of early identification of potential users of evaluation results, involvement of administrators and program practitioners, prompt completion of evaluation, and early and effective methods for disseminating the findings. If the above purposes of program evaluation are adhered to, and the goals of evaluation research are actually followed as closely as possible, there will very likely be the following kinds of secondary benefits as suggested by Suchman: (1) redelineation by program administrators of the objectives and underlying assumptions of their programs; (2) delineation of the "what" and "how" of a program to identify the essential aspects of it; (3) delineation of appropriate target populations and situational contexts for programs; (4) re-evaluation of the theoretical bases of programs, with special attention given to the way in which principles are translated into practice; (5) development of new hypotheses designed to provide bridges where gaps in theory currently exist; (6) a new awareness among personnel that leads to the questioning of existing programs and a search for alternatives; and (7) an increase in staff committment and improved morale.

One would be politically naive not to recognize that program evaluation can also be used to scuttle a program that has real merit. Or, alternatively, one could strongly promote a poor program. There are a variety of ways in which this kind of thing can be done, and again, it is Suchman that provides us with some interesting terminology in this area. He suggests that an eyewash is an evaluation that seeks to justify a weak program by deliberately evaluating only the good-looking surface aspects. (Appearance replaces reality.) A whitewash is an evaluation designed to cover up program failure by avoiding objective investigation. (Vindication replaces verification.) A submarine is an evaluation designed to "torpedo" a program regardless of its effectiveness. (Politics replace science.) Posture is the use of evaluation as a "gesture" of objectivity. (Ritual replaces research.) And postponement is an evaluation designed to delay needed action by pretending to seek "facts" (Research replaces service.)

Thus, the purposes of program evaluation can be many. Perhaps the simplest way of envisioning the entire evaluation process is to think of modern pressures for accountability. Whatever the funding source may be--local, state, or federal government, or industry, all sponsors want to achieve as much as possible with the available money. It is the purpose of program evaluation, whether summative or formative, to indicate the ways in which programs or parts of programs are effective or ineffective and to indicate the findings in as scientific a manner as possible. This is not to say, however, that program evaluation can solve all problems. There are a number of obstacles that can make the task of program evaluation very difficult. We shall return to a full discussion of such problems later, but the following brief review will give an indication of the kinds of problems that are encountered:

1. Evaluation research deals with people and programs in real life action environments; therefore, the program is considered the primary activity, and research must become secondary. When it comes to choosing between perfect research design and meeting the needs of the program, priority is usually given to the program.

2. The goals of programs are rarely simple or clear-cut. Not only do goals vary, but programs themselves can differ in scope, size, duration, clarity, specificity, complexity, time-span, innovativeness, and so forth. It is little wonder, then, that it is hard to evaluate a program or to compare programs. As stated before, however, and as will be discussed in more detail later, because funding for programs have become more intricately tied to accountability for achieving goals, delineation of goals will probably become more clear-cut.

3. Program staff may be reluctant to cooperate with evaluators. Researchers generally depend on them for data, but they are often concerned about losing their jobs and worried about what the data may show. People simply don't like to have others sit in judgment of their work.

4. Control groups for evaluation research are often difficult if not impossible to fund. When funding exists for programs, it usually is not intended for use in procedures that require random assignment of some groups for treatment and intentional exclusion of others.

5. The traditional experimental method of evaluation only shows how well the program has achieved its goals after the completion of the program. This does not allow for the formative type of evaluation discussed by Scriven; but often program directors and staff want feedback earlier so they can make changes along the way.

6. Evaluation research is meant for immediate and direct use in improving the quality of social programming. A review of evaluation experience suggests that evaluation results have generally not exerted significant influence on program decisions. Unfortunately, decision makers often respond to other information besides the program evaluation. Their decisions (often political decisions involving funding) operate to negate the importance of the program evaluation report (73).

Because of these and other problems, program evaluation obviously is not a simple thing to accomplish. However, as previously suggested, it seems clear that one might as well begin by approximating the ideal method as closely as possible. It is with this goal in mind that we now turn to an interplay between the experimental method and the goals of program evaluation in several different settings. We shall investigate the successive phases of program evaluation by using hypothetical examples from several projects that the reader can view as models of sound scientific program evaluation.

Experimental Program Evaluation

The above discussion has attempted to delineate some of the characteristics and goals of program evaluation. It is now appropriate to discuss more thoroughly the process of experimental evaluation. A number of writers, such as Fairweather (21,22), Griessman (30), and Welch (75), have attempted to concentrate on process.

For purposes of this discussion, a logically derived framework incorporating ideas from many writers will be adopted. In order of presentation, topics to be covered include: (1) forming the program evaluation team; (2) developing contacts with staff personnel and persons in position of authority; (3) implementation of experimental procedures; (4) feedback; and (5) dissemination and follow-through.

Forming the Program Evaluation Team: There are several factors to be considered in forming the evaluation team. The first consideration should be the academic training and background of the various members of the team. The use of the word team suggests the presence of individuals with a variety of skills. This is not always the case, especially when projects are very small. Under such conditions, it is of course advisable to recruit an individual with the best evaluation and experimental design skills available. This person should also be as familiar with program content and materials as possible. Often, however, more than one individual is needed and, if this is the case, a variety of perspectives is usually desirable. For these reasons, during the past few years funding agencies have been interested in interdisciplinary, multidisciplinary, and transdisciplinary teams. Although the differences in these team compositions are somewhat subtle, the important point is that evaluation methods and theories are usually enhanced by a variety of perspectives arising from different disciplines.

A second factor to be considered in formation of the team is organizational structure. Caro (13) speaks of two common arrangements. The first, which can be termed "in-house" or "inside" evaluation, is the type of project in which the researchers themselves are staff members in the organization whose programs are evaluated. The second approach has been labeled "outside" (or "out-house") evaluation. In this kind of evaluation, the evaluators are consultants from an organization other than the one whose program is being evaluated. There are advantages and disadvantages to both approaches.

On the one hand, as Alpern (1) notes emphatically with regard to the evaluation of head-start programs, in order to avoid experimenter bias, experimenters should not be connected with the program in any way. He feels that only then can an evaluation be objective. Alternatively, the federal government is now suggesting that in as many programs as possible, and particularly in community action programs, citizens or participants in the programs should have an opportunity to participate in all aspects of the programs. Consequently, Brocks (2) suggests that projects be "developed, conducted, and administered with the maximum feasible participation of residents of the areas and members of the group served." This theory is in keeping with the idea that if many perspectives are considered, the effort will be more successful. However, the idea can also be carried a bit too far. For instance, Guba (32) suggests that "evaluation practitioners on local, state, and national levels, evaluation consultants, evaluation research and development personnel, users of evaluation reports, consumers of evaluation reports, related professional groups and funding agencies might contribute" their expertise to the evaluation of a program.

A summary of the two positions is provided by Cameron, Kidd, and Price (5). They suggest that some advantages of outside evaluation are that the evaluator is:

1. Likely to be objective

2. Unlikely to be distracted by operational problems

3. Able to concentrate full efforts on assessments.

Some disadvantages, however, are that the outside evaluator:

1. Is incapable of intimately understanding the program

2. Ties up time of operational staff in learning about the program

3. Is likely to interfere with operations by imposing and by perturbing people with measurement activities

4. Imposes an external value structure on project purposes

5. Uses funds that would be better spent on refining optional aspects of the program

6. May cause operational staff to feel threatened and resentful.

On the other hand, the internal evaluator is:

1. Fully cognizant of all aspects of programs

2. Not a disruptive influence

3. Inexpensive.

But some of the disadvantages of internal evaluation are:

1. The evaluator lacks objectivity and perspective

2. Ego-involvement produces biases

3. Operational involvement may lead to an expensive evaluation (pp. 36-37).

An interesting compromise is suggested by a number of writers. Morris (44), Case (14), and Valentine and Larsen (70) argue that research should be viewed as a separate function but still placed within the local institution or the home program. With this arrangement, the evaluators are not project members themselves but are knowledgeable and their services might be less costly. The authors feel that the most efficient and valid outcomes can be anticipated through an institutional research system because this compromise incorporates many of the advantages of external and internal evaluation and negates many of the disadvantages.

Developing Contacts

The way in which contacts are developed with staff personnel and persons in positions of authority depends primarily on whether the evaluation staff is to be an inside or outside research team. Since the latter is more often the type used,our discussion will be oriented towards the external evaluation perspective.

The importance of establishing firm commitments from and excellent rapport with both the administrators of programs and the staff or operational personnel of programs cannot be overstated. More program evaluation efforts than this writer could list have been scuttled because of friction that arises between evaluators and the group they are working with. As Flint (24) suggests, "it is not known how many good research designs have failed (to fulfill program goals) because lack of tact and understanding lead to premature program termination." It is not hard to understand how such friction is created. Underlying this friction is what Rossi (50) terms the power of wishful thinking. "The will to believe that their programs are effective is understandably strong among administrators. As long as the results are positive (or at least not negative) relations between practitioners and researchers are cordial and even effusive. But what happens if results are negative?" At any rate, it should be clear that every attempt should be made to establish commitments on both sides towards maintaining excellent rapport.

One development that has the potential of either decreasing or increasing this problem is the emergence of a number of public laws requiring programs and projects to undertake program evaluation. The passage of such laws creates the need for systematic attempts to determine appropriate evaluative strategies. Seashore (55) and Dean, et al. (16) offer reports of such attempts. When public law requires that programs be evaluated, there may be less conflict between administrators and operational staff, and the evaluators.

A different topic that should be raised at this time has to do with programs that don't actually require evaluation, but are deemed likely prospects by evaluators. Under these conditions, it is incumbant upon the evaluator to "sell" administrators and project personnel on the need for evaluation. Further, it is often the case that evaluators with social science backgrounds wish to promote an alternative treatment program, new curriculum, or some such change. When this is the case, evaluators and researchers try to present the best case they possibly can for their position to convince the project personnel that such experimentation and evaluation are needed.

The problem comes when the idea is "oversold." Often, practitioners don't really understand empiricism, and when they are convinced in the initial argumentation that their approach should be tried, they are at the same time convinced their approach is the right answer. Researchers, on the other hand, even though they are the ones who have "sold" the program, can still sit back and wait for the evidence to tip the scale one way or the other. Meanwhile, project personnel are not so empirical, and sometimes begin to operate as if they have the "right" approach. Researchers are placed in a "must win" position;

when staff are oversold in such a fashion, they can react extremely negatively when the empirical evidence is presented, especially if the new implementation or innovation is not effective. In this case, the evaluator or researcher loses credibility and further efforts meet even more resistance. Clearly, evaluators should be careful not to oversell.

Implementation of Experimental Procedures

In discussing the implementation of program evaluation, the experimental approach will be considered an ideal. We should like, as Campbell did in his classic article "Reforms As Experiments" (7), to challenge the reader with the suggestion that social action programs can be conceived of in experimental paradigms. A number of topics are relevant to this implementation section: (1) definition of criteria (dependent variables); (2) definition of the independent variables (treatments); (3) development of instrumentation for measurement; (4) implementation; and (5) appraisal of the evaluation.

Definition of Criteria: Consider the Supreme Court decision discussed earlier. What could the possible benefits of such a law have been? Or, to state it another way, just exactly what did the jurists think they would be changing by means of that law? First, it seems clear that they believed the short-range impact would be to produce behavioral change in black children who attended predominantly white schools and white children who attended predominantly black schools. In other words, one of the legitimate areas in which to look for change would be in behavior. A second effect that the jurists probably anticipated was that the black children might pick up more content as a result of receiving "better" educations. Similarly, there were many who felt that white children would have an opportunity to learn more about black people and their culture by interacting daily with black children. Therefore, a second type of result that can be studied without regard for implementation of a program is the potential for change in knowledge for program participants.

In addition, many had proposed that there would be a lessening in prejudicial attitudes with forced integration. They believed that old stereotypes would be erased and black and white children would come to feel better about one another and interact with one another. Consequently, the third possibility is that programmatic intervention can bring about affective (attitudinal) change. In an example from another sphere, the State of Florida has recently funded a project to conduct in-service training for project and staff directors in the Division of Aging. Because it is possible to study behavioral, knowledge, and affective changes stemming from such educational programs, we attempted to assess all three by administering measures before and after implementation of the Florida program. We measured the following: (1) participants' knowledge in relation to the curriculum; (2) their attitudes towards the content of the curriculum; and (3) participants' behavior in relation to the content of the curriculum.

Existing literature offers information concerning studies of minority treatment using each of the above approaches. In two different studies, which were of the time series design (as discussed earlier in the quasi-experimental design section of this paper), Shaw (57) and Silverman & Shaw (58) utilized what might be termed affective dependent variables. In the first study, Shaw administered sociometric questionnaires at three different times to pupils in the fourth, fifth, and sixth grades of an elementary school. He administered them to both blacks and whites and hoped to note changes in the sociometric choices

across time. In the second study, both blacks and whites were questioned about interracial attitudes. This study also tapped behavior and specifically measured the frequency of interracial interaction across time as a result of sudden mass school desegregation.

A very important issue to be considered under the heading of dependent variable identification has been mentioned previously; namely, it is becoming more and more important to identify the goals of a particular program and to characterize the dependent variables in light of the goals of the program. In discussing the implications for the evaluation of programs for the disadvantaged child, Hodges (37) suggests that goals and objectives are generated directly from the assumed needs of the children. In turn, it should be obvious that if a program is funded to achieve a particular goal, one will want to demonstrate that the program has been effective so that future funding for similar ventures can be obtained. Because funding agencies are very concerned about accountability, it is quite understandable that dependent variables will more closely than ever before reflect the goals of the program.

Interestingly enough, there is an approach coming out of organizational behavior, industrial psychology, accounting, administration, and a variety of other disciplines that lends itself directly to the phenomenon that we have been discussing. The approach, known as management-by-objectives (MBO), involves three basic factors that will affect success: (1) goals and goal setting; (2) participation and involvement of subordinates; and (3) feedback and performance evaluation (68, 69, 77). The point is that many funding agencies are now requesting that grants be written in terms of behavioral goals and behavioral objectives. Further, for management purposes, many are suggesting that management also be run by objectives. The objectives of management and administration are to set behavioral goals, enlist staff support of these goals, and evaluate performance towards the attainment of these behavioral objectives. Funding agencies want precisely this approach for accountability studies. It should be clear, then, that for program evaluation the behavioral goals are set and the behavioral objectives should be those dependent variables or criterion variables that the program hopes to alter or attain (4).

Definition of Independent Variables

As Charters and Jones (15) point out in their article "On the Risk of Appraising Non-Events in Program Evaluation," we must be very careful to document the fact that a treatment is being undertaken when programs are implemented. As much time should be spent conceptualizing and delineating the independent variables as the dependent variables in any study. An example from the area of desegregation research is provided in the comprehensive work by Koslin, Josephson, and Pargament (40). The authors appropriately point out that the term desegregation can imply quite a few different things. Consequently, they attempt to delineate the characteristics of the term desegregation as specifically as possible for a particular school district. This delineation is focused on two concerns. First, what is the nature of the desired change in the student body composition? The answer to this question involves addressing

the issues of racial balance, racial heterogeneity, and social-class balance and social-class heterogeneity. Second, how extensive will the desegregation be? The answer depends on how many grades are to be involved, what proportion of the students in these grades are to be involved, and how many schools are to be involved. Consequently, it should be apparent that much thought should be given to the definition and conceptualization of independent variables (treatments). Secondly, as Charters and Jones suggest, we make sure by means of appropriate measurement that there is an actual treatment undertaken in treatment programs. Measurement of independent as well as dependent variables is of paramount importance.

Development of Instrumentation

As was indicated in both of the above two sections, it is clear that instrumentation has to be developed to tap both independent variables and dependent variables--independent variable taps to insure that treatments are actually in effect, and dependent variable taps to discover whether or not the program or treatment has made any impact. The problems connected with this kind of measurement are not small and they seem to become greater as the size of the program increases (26).

One of the dilemmas facing most researchers is the question of whether or not to use well-standardized, well-validated traditional measures or measures that are tailor-made for their particular needs. Clearly, this is not a problem for behavioral indices, but rather for content and affective indices. For each article that suggests that instrumentation be standardized and traditional (for example, 3 and 28) one finds another suggesting that new instruments be tailored to hope for responses (for example, 76). However, it seems that a compromise between the two positions would bear the greatest potential for most research purposes. Such a compromise would mainly involve incorporating several standardized and traditional measures to allow as much comparability as possible, while at the same time including new instrumentation to reflect as precisely as possible the nuances and differences in each unique program. It shall not be the purpose of this paper to go into the construction of such instrumentation, but reference can be made to an earlier work by Severy (56).

Implementation

One of the more thought-provoking approaches to implementation of experimental program evaluation is provided by Fairweather (21, 22). Taking the perspective of the nonviolent approach to social change, Fairweather challenges us to survive. He suggests that in order to do so, we must engage in a process very similar to that being suggested in his module; namely, we must define socially innovative experiments in which evaluation teams (1) define a significant social problem; (2) carry out naturalistic observations; (3) innovate a new social sub-system; (4) design an experiment to compare it with the traditional sub-system; (5) implant the two sub-systems in the appropriate social context; (6) evaluate the sub-system longitudinally; and (7) take responsibility for the welfare of the participating members. In addition, teams should be multidisciplinary in their make-up. Fairweather's provocative work continues with the suggestion that

experimental social innovation is really a marriage of the experimental approach and service. In fact, when he reviews the common research methods (descriptive theoretical approach, survey approach, laboratory approach, participant-observer approach service, and experimental approach) Fairweather claims that the two having the most common characteristics with regard to the eight aspects of social innovation identified above are service and the experimental method. This does not seem unreasonable, since we have been suggesting all along that the focus of program evaluation is to experiment through service.

Campbell appropriately points out that "from the point of view of natural laboratories, it is always going to be easier for us to evaluate the differential affect of a program innovation than it is to evaluate the whole package already existing" (52). When program innovations can be conceived ahead of time and interwoven into the experimental design, as a part of classic treatment or manipulation, evaluation will be much easier.

It is, of course, during the process of implementation that one chooses the experimental design and utilizes it. Consequently, it is at this point that the experimental approach, as described in the earlier section, should be approximated as closely as possible. For a comprehensive review of the advantages and disadvantages of five different program designs that have been utilized in desegregation programs, one should refer to the work of Koslin, Josephson, and Pargament (40).

Appraisal of the Evaluation

It is important to ask if you have done a good job of evaluation, just as it is important to ask if you have done a good job with your program. Consequently, Tallmadge and Horst (66) have developed a procedural guide that involves 23 steps towards validating the effectiveness of educational programs using existing evaluation data. The guide is constructed to allow "branching" and a particular answer to any one question leads you to another question in the guide. Although there is not enough space here to describe the complete procedure regarding the nature of the answers to the questions, perusal of the list of questions should indicate what would be classified as a good evaluation. The steps in the procedure involve the following questions:

1. Are the test instruments adequately reliable and valid for the population being considered?
2. Are pre-or post-test score distributions of any groups curtailed by ceiling and floor effects?
3. Is there reason to believe that the pre-testing experience may have been at least partially responsible for the observed experimental outcome?
4. Is there reason to believe that knowledge of group membership may have been at least partially responsible for the observed experimental outcomes?
5. Is there reason to believe that student turnover may have been partially responsible for the observed experimental outcome?
6. Does the evaluation employ a control group?

7. Were pre-test scores used to select the treatment group?

8. Are normative data available for testing dates which can be meaningfully related to the pre- and post-testing of the program pupils?

9. Do the norms provide a valid baseline against which to assess the programs of the treatment group?

10. Is the comparison between the treatment group and the norm group based on pre- and post-test scores or on gain scores?

11. Have appropriate statistical tests been employed to assess the significance of the gain in treatment group performance relative to the norm group?

12. Are pre- and/or post-test scores available?

13. Can appropriate statistical tests be employed to assess the significance of gain in treatment group performance relative to the norm group?

14. Were the children, either matched or unmatched, randomly asigned to the experimental and control groups?

15. Is there evidence that members of the experimental and control groups both belong to the same population, or to populations that are similar on all educationally relevant variables, including pre-test scores?

16. Are post-treatment comparisons made in terms of post-tests or gain scores?

17. Can data be obtained which would enable application of analyses of covariance techniques; would such analyses be appropriate; and is there reasonable expectation that they would produce significant results?

18. Is the control group superior to the experimental group on the balance of educationally relevant variables?

19. Have covariance analysis techniques been employed to adjust for initial differences between groups?

20. Have appropriate statistical tests been employed to compare post-test and gain scores?

21. Can data be obtained which would enable appropriate tests to be made?

22. Do analysis results favor the treatment group at the preselective level of statistical significance?

Feedback

Earlier we discussed the fact that there are different types of evaluation; namely, formative and summative. Recall that formative (ongoing) evaluation should be undertaken at appropriate intervals--as often as is reasonable and only as rapidly as is reasonable. One does not want to flood project personnel with meaningless trivia. It is important for conclusions to be made that can be brought to the attention of administrators and staff.

Writing an evaluation of programs designed for the disadvantaged is not a particularly easy task nor is it an enjoyable one. As will be discussed later, it is often difficult to leave ones own values and value judgments out of such

analysis. Further, it is encumbent upon the evaluator to be fully cognizant and apprised of each program's unique characteristics, goals, extenuating circumstances, and, most important. any special requirements or stipulations that have been imposed on the programs. For these reasons, many now suggest specific approaches that would be suitable for different problem areas. For example, Thonis (67) discusses concerns and principles connected with evaluating the effectiveness of programs tailored for Mexican-American students.

It is important that feedback to project staff and administration be reasonable, and it should not be delayed. The aim should be to put any program in a position to redirect itself or sell itself as much as is possible. The whole purpose of program evalution is to offer meaningful data to program personnel as quickly as possible. On the other hand, there is no point in over burdening the staff with meaningless data.

Dissemination and Follow-Through

Individuals and agencies in the professional community other than the project administration and staff like to know what's happening. They don't want to make mistakes that have been made before, and at the same time, they want to implement whatever has been successful. As has been pointed out by Larsen and Nichols (41), "if nobody knows you've done it, have you...?" As Welch (75) put it, "perhaps most unique to the evaluator...is the attention that must be paid to effective reporting of evaluation findings. Because the primary purpose is to provide information to decision-makers, the need for effective communication is paramount." However, disseminating information is not as easy as it sounds. "Information may be derived from the evaluation, but may lend itself to reports for several audiences. This may include reports to program administrators, press releases, and papers and articles for scholarly research circles. The evaluator, as a scientist, has a commitment to obtain and share knowledge" (30). Often an evaluator must write up a variety of evaluative reports--each one for a different audience. For example, one for practitioners and one for scholarly purposes might be useful.

In the dissemination of information, there seems to be a proclivity for publishing positive findings. As Alpern and Levitt (1) suggest, some people have misinterpreted the effectiveness of head-start programs because only positive reports have been published. Whenever programs are to be compared, negative information is as important as positive findings. Evaluation and "pure" research have different intents, and it is especially important to publish both the negative and positive findings of program evaluations. The question of follow-through is another that distinguishes program evaluation from traditional research. Fairweather as was already indicated, made the plea that program evaluators be intricately concerned with the welfare of the participants of the programs and follow through as long as is necessary to make sure that implementation of the appropriate programs takes place. Others, such as Larsen and Nichols (41) are vitally concerned with building better utilization models. It is their feeling that researchers simply don't go far enough in aiding practitioners with their problems. It could be that researchers are so used to conceiving of alternative programs or alternative approaches that they never make the committment to claim confidence in any one approach.

At any rate, often the knowledge that can be imparted to practitioners is not forwarded to them. This is really a double attack on the participants or potential participants of a program. Not only is the researcher not personally involved in helping the participant, but he's also not letting whatever pieces of information he does have filter through to the participant.

Problems and Alternatives

Topics to be considered in this final section include problems with the experimental method, alternatives to it, and concluding comments.

There are two classes of problems that should be addressed: first, problems connected with program evaluation in general; and second, problems unique to the experimental method.

General Problems in Program Evaluation

Throughout this paper potential stumbling blocks in program evaluation have been pointed out. However, there are some more general problems that should be discussed. First, there is the problem of personal values and personal judgment whenever a human being evaluates a social program. The problem is not always circumvented by simply adding more people to the evaluative team. In fact, that approach might compound the problem because two or three or more people, all with different values, might end up in a conflict and really bog down the evaluation. As Neufeldt (46) points out, there is no single way to perform evaluation. There is no logical structure that assures the right method. Evaluation eventually becomes judgment as long as there is no ultimate ordering or priorities, and the critical question in evaluation is: Who has the right to decide? Consequently, as objective as we might try to make an evaluation, there is a personal value structure and a personal value judgment that will be made. Similarly, Stake (60) feels that "it is likely that judgments will become an increasing part of the evaluation report."

As we will discuss shortly, there are now suggestions that more subjective approaches be included in program evaluation, but there are problems as can be seen when Stake (60) states that "evaluators will seek out and record the opinions of persons of special qualification. These opinions, though subjective, can be very useful and can be gathered objectively, independent of the solicitors opinion. A responsibility for processing judgments is much more acceptable to the evaluation specialist than one for rendering judgments himself." I would question how much better the judgment becomes when the evaluator asks those involved in a program rather than depending on his or her own judgment, especially in an outside evaluation. Instead of circumventing responsibility for judgment, Stake is suggesting that we diffuse responsibility for those judgments.

There may be a very good reason to diffuse the responsibility for such judgments, for as Wortman (78) points out, one cannot escape politics when one talks about program evaluation. "Action programs for ameliorative innovation

are not free of the political controversies surrounding their implementation... social experimentation is a political act...this implies that such projects and the persons involved are thereby subject to the same political pressures already enveloping the issue."

Another problem that is independent of the particular method of evaluation is that of the uncertain trends of funding and the drifting of the goals and priorities of funding agencies and governmental policy and decision makers. This uncertainty, of course, works against the longitudinal approach that most recognize as ideal, and instead works in favor of short-term designs (31).

These particular problems will always confront those who decide to embark upon program evaluations of any kind. However, there are some problems that are specifically related to the experimental method.

Problems Connected with the Experimental Method: Perhaps the best review of the problems of social experimentation is provided by Rivlin (49). She suggests that there are at least six problems, but her perspective is one of optimism, and she looks more to the promise than to the problem. At any rate, she suggests that:

1. There are design dilemmas, some of which arise from the conflict between the desire to obtain valid, reliable results, and the equally urgent desire to obtain results quickly and at low cost.

2. There are implementation dilemmas.

3. There are dilemmas attached to the evaluation itself.

4. There are timing dilemmas, for if the results of the social experiments are to effect decisions, they have to be available when the decisions are being made.

5. There are difficult moral questions associated with experimenting with people.

6. There are a series of dilemmas having to do with the openness of experiments such as maintaining the privacy of participants, and how much the experimenter should divulge to different participants.

These problems and others have led a number of writers such as Stufflebeam (62), English, Frase, and Melton (19) to seriously question the experimental design approach. However, the most adamant critics, logically, are those individuals who have proposed alternatives to the experimental approach. For the sake of order, let's first look at the criticisms, such as those leveled by Guttentag (33,34,35) and Weiss and Rein (74). Guttentag points out that problems with the experimental method can be viewed from both the program administrator's perspective and the program evaluator's perspective. She suggests that in the program administrator's view administrators often find that researchers assume that action programs are designed to achieve some specific ends, when, in fact, program people believe that this kind of approach is misleading, especially when action programs have broad aims and unstandardized forms.

Program people often cite illustrations of research disasters in the evaluation of broad social programs, which suggests that if a program is forced into an experimental paradigm, logical choices and decisions are often not made. She suggests that sometimes a programmer-administrator may feel that evaluation researchers studied "interesting questions (when they could understand what they were studying), but the work seemed to be basic research, and had little to offer to anyone who had to decide what to do in a real situation."

Guttentag also views problems from the researcher's point of view, and suggests that they are the inverse of those faced by the program administrator. She suggests that most of the assumptions of the experimental model cannot be fulfilled. Because he does not begin with his own hypotheses, and the researcher may have no control over what he is studying and usually cannot randomize his subjects or his treatments or control the flow of subjects into or out of programs. According to Guttentag, "when he does try to do so, conflicts with program administrators result. Even when a control group is established, true random assignment of subjects to experimental and control groups is rare." The author distains the logical positivist approach that she feels is perfectly reflected in the experimental approach, and further, attacks the use of T or F tests in such research. She feels that "virtually all of the assumptions underlying these tests are violated" and states, "clearly, the experimental model does not ask the right questions for evaluation research."

Weiss and Rein (74) also approach a variety of problems, but concentrate most heavily on what they feel are technical difficulties with the experimental design. They may be enumerated as follows:

1. There is a problem in developing the criteria.

2. There is a problem in that the situation is essentially uncontrolled.

3. Oftentimes treatment is not standardized.

4. Experimental designs discourage unanticipated information.

As a consequence of these criticisms, alternative approaches have been suggested for program evaluation.

Alternatives

Rather than specifically delineating a variety of alternatives, it shall be the purpose of this section to indicate current views of alternatives. A more complete delineation of a variety of methods can be found in Guttentag (33). In the discussion concerning the problem of judgment in program evaluation, we pointed out the potential of increasing the use of subjectivity in evaluation research. Whether it be described as increasing the number of judgments or the utilization of subjectivity, participant-observers, decision-theoretic evaluation, or process-oriented qualitative research, these program evaluators--Stake (60), Weiss and Rein (74), Glaser and Backer (29), and Guttentag (33,34,35)--are all arguing for what may be described as subjectivity in evaluation research. Sociologists talk about ethno-methodology and anthropologists talk about the anthropological perspective through participant observation. Whatever the

discipline is, it should be recognized that there is probably a place for such subjectivity. There are, however, arguments against such an approach.

Campbell (6), who has openly defended the potential of such subjective approaches, attempts to place the subjective approach in proper perspective. He suggests that Weiss and Rein's approach "contains mutually imcompatible elements, which are individually compatible with the experimental method, and requires the experimental method for reducing equivocalities of inference." He further suggests that the four problems pointed out by Weiss and Rein are, in fact, often weaknesses in would-be experimental program evaluations, but that they can be avoided by better experimental methods and better experimenters. He also suggests that if some of the subjective approaches are to be thought of as "making an argument in favor of common-sense knowing,....I agree with them. But if they are arguing for an alternative that is as good as or superior to an experimental design, then I've got to wait for an example to see whether or not I feel it would have been strengthened with more attention to experimental design" (52).

Conclusions

Mushkin (45) concludes that "after looking very closely at the findings emerging from evaluations of social programs...the methodology of evaluation is still inadequate to serve as an overall policy guide. Further,....the process of evaluation is not understood well enough for public debate on policy options." The point is well taken. It's very probable that the experimental method has not been examined by enough people; the process is not very well understood and therefore is often criticized. In fact, Campbell states that "you will find a general concensus that 99 percent of our ameliorative programs have not been evaluated in an interpretable way" (52). The author further states that "we have papers...saying that his model is inappropriate, out of date, doesn't do enough, or whatever. Yet it's almost never been tried. There are very few program evaluators that have used the experimental method. The orthodoxy that people are rebelling against has only been an orthodoxy of practice." Consequently, the challenge for new program evaluators is: Does it work? The question can only be answered in relation to individual programs. After a series of experiences in program evaluation, Evans (20) offers a few thoughts about the future.

1. My experience leads me to disagree with the cynic's view that evaluation is generally a waste of time—that partisan-political considerations are the overwhelming factor in determining what happens to government social action programs, and that empirical evaluation, and rational analysis, can never hope to be more than an insignificant input to the decision making process.

2. An important lesson we must all learn is that our task in evaluating social action programs in the real world is not to produce methodologically perfect studies, but rather to improve decisions by doing the best that can be done in a timely and relevant way.

3. Continuing to belabor the point of the "poor state of the art" as many of us do, is a poor excuse for not getting at the task at hand, and it serves only to delay, not to accelerate the contributions that social scientists make to program assessment and policy determination (pp. 577-578).

Even in light of the criticism leveled against the experimental approach, Campbell's challenge and Evans' optimism serve to support the view that the experimental approach to program evaluation is the ideal.

Resources in Experimental Program Evaluation

This paper cannot cover all of the complexities, principles, procedures, and problems connected with program evaluation. However, the following section can serve as an introduction to these areas, and it is hoped that it will stimulate the interest of the reader. It is with this thought in mind that brief resumes of a number of rather comprehensive volumes have been prepared. Although each of the more than 75 references included in this paper provide specific information that may be helpful to the reader, the volumes presented below (in alphabetical order and briefly annotated) represent some of the more comprehensive resources available in the area of program evaluation.

1. Readings in Evaluation Research (12). This volume contains a comprehensive set of readings in the social and behavioral sciences and extensive information regarding evaluation research. It contains views on: the scope of the field; methods; how to conduct research in an evaluative way. Examples of actual studies are included. The volume was supported by the Russell-Sage Foundation in the interest of improving and developing the field of evaluation research. The readings themselves are presented in four sections that follow an interesting discussion of evaluation research by Caro. The four sections are entitled "Basic Issues: Program Development and Scientific Inquiry"; "The Organizational Context: Establishing and Maintaining the Evaluative Research Role"; "Methodological Issues: Measurement and Design"; and "Case Materials." The 31 different papers that follow Caro's discussion are well chosen and represent a broad spectrum of ideas in program evaluation.

2. Methods for Experimental Social Innovation, (21). Fairweather, the author of this book, is interested in answering two questions. First, "how can society affect needed changes in ongoing social processes with a minimum of disruption? I propose that the answer to this question is to create a new social subsystem whose methods include innovating models as alternative solutions to social problems, experimentally evaluating them, and disseminating the information to those who can make the appropriate changes." Second, "how can this be done?" The author sets up a model for the first question, and then, in the main body of the volume answers the second question. In effect, this is a how-to-do-it book, and a very large amount of information regarding a particular procedure for social program evaluation is covered in this rather short volume.

3. Social Experimentation: A Method for Planning and Evaluating Social Intervention (48). "This book is the product of the committee appointed by the Social Science Research Council in 1971 to summarize the available knowledge about how randomized experiments might be used in planning and evaluating ameliorative social programs...the result is a comprehensive statement of the promise and the problem of social experimentation." In this volume, a comprehensive statement of social experimentation is provided by a committee of competent social researchers and, although it is an edited volume, it is not a set of readings; rather, the committee combined their efforts in a format that reads like a general text. The major sections of the volume include: (1) Experimentation as a method of program planning and evaluation; (2) Why and when to experiment; (3) Experimental design and analysis; (4) Quasi-experimental designs; (5) Measurement in experiments; (6) Execution and management; (7) Institutional and political factors in social experimentation; and (8) Human values and social experimentation. Following these sections, one finds an interesting appendix entitled "Illustrative controlled experiments for planning and evaluating social programs." A large number of examples are abstracted and are classified as to programs in delinquency and criminal reform, law-related programs, rehabilitative programs, mental health, special education programs, sociomedical and fertility control, communication methods, and so forth.

4. Evaluating Social Programs (51). This is another book of readings, which are grouped into four different sections: an overview; a look at the theory in evaluation research; a look at the practice of evaluation research; and, organizing for large-scale evaluation research. The editors of this volume address three related questions. "First; why have the quantity and quality of evaluative activity to date been so slow? Second; what are the problems and risks associated with developing more evaluation research and using the results in the social policy process? Third; what steps should be taken by the government and the social science research community to increase significantly the level of soundly conceived and executed evaluative studies and to reduce the dangers intendent in the use of the results?" In response to these questions, the editors focus their essays on five different kinds of problems: conceptual, methodological, bureaucratic, political, and organizational.

5. Evaluative Research (64). This brief text presents a comprehensive discussion of what evaluation research involves and views the social scientist in the world of action. This volume was supported by a Russell-Sage Foundation program to improve and develop the field of evaluation research. The Caro volume of readings described earlier was developed as the companion volume to this introduction to the field. The author, Suchman, has included chapters on topics such as experimental design, ties with program administrators, types of evaluation, and the administration of evaluation studies.

6. Evaluating Action Programs: Readings in Social Action and Education (72). Following an overview entitled "Evaluating educational and social action programs: a treeful of owls", Weiss presents a series of 20 different readings, including some of the classic early articles in the field of program evaluation. According to the author, "the book aims to help the reader conceptualize and understand the purposes of evaluation and the methods by which it obtains information and generates conclusions. It assumes an elementary knowledge of social science research methods; even a passing acquaintance would get the reader through the book. Rather than giving a set of pre-fabricated rules and instructions, it points out the constraints within which evaluation operates and suggests alternative strategies of design, measurements, structure, relationship, and communication in order to accomodate two existing constraints and to serve the informational needs of programs."

7. Evaluation Research: Methods of Assessing Program Effectiveness (71). This volume is a companion to the preceding book written by Weiss. It is a short work that deals with "the application of research methods to the evaluation of social programs--programs in education social work, corrections, health, mental health, job training, technical assistance, community action, law, and so on...the basic theme of the book is that evaluation uses the methods and tools of social research, but applies them in an action context that is intrinsically inhospitable to them...a principle aim of the book is to acquaint the reader with the realities of evaluation life." The book has short sections on the purposes of evaluation, formulating the question and measuring the answer, design of evaluation, the turbulent setting of the action program, and the utilization of evaluation results.

The above resources are probably the best available at this time. Four general text-like volumes and three sets of readings are included, and together they represent a rather comprehensive statement of the concerns, principles, procedures, and problems of program evaluation.

REFERENCES

1. Alpern, G. D., & Levitt, E. E. Methodological considerations in devising head start program evaluations. 1967.

2. Brooks, Michael. The community action program as a setting for applied research. Journal of Social Issues, 1965, 21, 29-40.

3. Brown, R. D. Student development in an experimental college: some evaluation strategies and outcomes. 1971. ED 049 291

4. Cain, G. G., & Hollister, R. G. The methodology of evaluating social action programs. Public-Private Manpower Policies, November, 1969, 5-37.

5. Cameron, B. J., Kidd, J. S., & Price, H. E. Operational evaluation from the standpoint of the program manager. Falls Church, Va.: Bio Technology, Inc., 1971. ED 069 747

6. Campbell, D. T. Considering the case against experimental evaluations of social innovations. Administrative Science Quarterly, 1970, 15, 110-113.

7. Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

8. Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make conpensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education: national debate, Vol. 3, disadvantaged child. New York: Brunner/Mazel, 1970.

9. Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental design for research. Chicago: Rand McNally, 1966.

10. Caporaso, J. A., & Roos, L. L., Jr. (Eds). Quasi-experimental approaches. Evanston, Ill.: Northwestern Univer. Press, 1973.

11. Carithers, M. W. School desegregation and racial cleavage, 1954-1970: a review of the literature. Journal of Social Issues, 1970, 26, 25-47.

12. Caro, Francis G. (Ed.) Readings in evaluation research. New York: Russell Sage Foundation, 1971. ED 058 327

13. Caro, F. G. Approaches to evaluation research: a review. Human Organization, 1969, 28, 87-99.

14. Case, C. M. The application of PERT to large-scale educational research and evaluation studies. 1969. ED 030 967

15. Charters, W. W., Jr., & Jones, John E. On the risk of appraising non-events in program evaluation. Educational Researcher, 1973, 2, 5-7.

16. Dean, G. S., Robinson, S. A., Strem, B. E., & Thompson, J. E. Regional medical programs: guidelines for evaluation. Los Angeles: Univer. of Southern California School of Medicine, 1968. ED 042 104

17. Doyle, W. J., & Schwarty, H. S. Methodology: a critical issue for research and evaluation in experimental programs. Chicago: Graduate School of Education, Univer. of Chicago. ED 074 045

18. Eash, M. J., Talmage, H., & Walberg, H. J. Evaluation designs for practitioners. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, & Evaluation, 1974. ED 099 430

19. English, F. W., Frase, L. E., & Melton, R. G. Evaluating the effects of implementing a differentiated teaching staff: problems and issues. A tentative position paper for use in project evaluation. Mesa, Ariz.: Mesa Public Schools, 1971. ED 056 993

20. Evans, J. W. Evaluating social action programs. Social Science Quarterly, 1969, 50, 568-581.

21. Fairweather, G. W. Methods for experimental social innovation. New York: Wiley, 1967.

22. Fairweather, G. W. Social Change: the challenge to survival. Morristown, N.J.: General Learning Press, 1972.

23. Farmer, J. A., Jr., Sylvester, R. K., & Weagraff, P. J. Western region AMIDS evaluation: a description of evaluative research design and methodology. Los Angeles: Univer. of California at Los Angeles, Division of Vocational Education, 1970. ED 044 576

24. Flint, R. T. Evaluation overview--evaluation: the key to better police service? Evaluation, 1972, 1, 6-8.

25. Fox, D. J. Issues in evaluating programs for disadvantaged children. The Urban Review, 1967, 2, 6-9.

26. Freeman, H. E., & Sherwood, C. C. Research in large scale intervention programs. Journal of Social Issues, 1965, 21, 11-28.

27. Gage, N. L. (Ed.) Handbook of research on teacning. New York: Rand McNally, 1963.

28. Gladkowski, G. Evaluation considerations of compensatory education. University Park, Pa.: Pennsylvania State Univer., 1973. ED 090 323

29. Glaser, E. M., & Becker, T. E. A look at participant observation. Evaluation, 1973, 1, 46-49.

30. Griessman, B. Eugene. An approach to evaluating comprehensive social projects. Educational Technology, 1969, 9, 16-19.

31. Grotberg, E., & Searcy, E. A statement and working paper on longitudinal/ intervention research. Washington, D.C.: George Washington Univer. Social Research Group, 1972. ED 091 056

32. Guba, E. G. A look to the future. 1971. ED 052 223

33. Guttentag, M. Models and methods in evaluation research. Journal for the Theory of Social Behavior, 1971, 1, 75-95.

34. Guttentag, M. Evaluation of social intervention programs. In Annals of the New York Academy of Science, 1973, 218, 3-13.

35. Guttentag, M. Subjectivity and its use in evaluation research. Evaluation, 1973, 1, 60-65.

36. Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.

37. Hodges, Walter L. The implications of design and model selection for the evaluation of programs for the disadvantaged child. Merrill Palmer Quarterly, 1973, 19, 275-288.

38. Hyman, H. H., & Wright, C. R. Evaluating social action programs. In Lazarsfeld, P. F., Sewell, W. H., & Wilensky, H. L. (Eds.), The uses of sociology. New York: Basic Books, 1967, 741-783.

39. Kooi, B. Y. Planning and measurement in school-to-work transition. New York: National Institute of Education, 1972. ED 087 856

40. Koslin, S. C., Josephson, B., & Pargament, R. Guidelines for the evaluation of desegregation programs in school districts. New York: Riverside Research Institute, 1972. ED 094 042

41. Larsen, J. K., & Nichols, D. G. If nobody knows you've done it, have you...? Evaluation, 1972, 1, 39-44.

42. McCall, W. A. How to experiment in education. New York: Macmillan, 1923.

43. Mondale, W. F. Social accounting, evaluation, and the future of human services. Evaluation, 1972, 1, 29-34.

44. Morris, R. Task force report on assessment and evaluation. Boston: Action for Boston Community Development, Inc., 1961. ED 001 523

45. Mushkin, S. J. Evaluations: use with caution. Evaluation, 1973, 1, 30-35.

46. Neufeldt, A. H. Considerations in the implementation of program evaluation. 1973. ED 079 642

47. Pettigrew, T. F. Social psychology and desegregation research. American Psychologist, 1961, 16, 105-112.

48. Riecken, H. W., & Boruch, R. F. (Eds.) Social experimentation: a method for planning and evaluating social intervention. New York: Academic Press, 1974.

49. Rivlin, A. M. Social experiments: the promise and the problems. Evaluation, 1973, 1, 77-78.

50. Rossi, P. Evaluating social action programs. Transaction, 1967, 4, 51-52.

51. Rossi, Peter H., & Williams, Walter (Eds.) Evaluating social programs: theory, practice, and politics. New York: Seminar Press, 1972.

52. Salisin, S. Experimentation revisited: a conversation with Donald T. Campbell. Evaluation, 1973, 1, 7-13.

53. Scheier, E., Senten, D. R., & Dionne, J. L. Learning 100 evaluation manual. Hunting, New York: Educational Development Laboratories, Inc., 1969. ED 047 192

54. Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967, 39-83.

55. Seashore, C. M. Regional meetings in evaluation research, final report. Washington, D.C.: National Training Laboratories, National Education Association, 1966. ED 010 229

56. Severy, L. J. Procedures and issues in the measurement of attitudes. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1974. ED 099 426

57. Shaw, M. E. Changes in sociometric choices following forced integration of an elementary school. Journal of Social Issues, 1973, 29, 143-157.

58. Silverman, I., & Shaw, M. E. Effects of sudden mass school desegregation of interracial interaction and attitudes in one Southern city. Journal of Social Issues, 1973, 29, 133-142.

59. Staats, E. B. The challenge of evaluating federal social programs. Evaluation, 1973, 1, 50-54.

60. Stake, R. E. The contenance of educational evaluation. Teacher's College Record, 1967, 68, 523-540.

61. Strevell, W. H. (Ed.) Rationale of education evaluation. Pearland, Tex.: Gulf Schools Supplementary Education Center, 1967. ED 034 292

62. Stufflebeam, D. L. The use of experimental design in educational evaluation. AERA 1970. ED 045 706

63. Suchman, E. A model for research and evaluation on rehabilitation. In M. Sussman (Ed.), Sociology and Rehabilitation. Washington, D.C.: Vocational Rehabilitation Administration, 1966. Pp. 52-70.

64. Suchman, E. Evaluative research. New York: Russell Sage Foundation, 1967.

65. Suchman, E. Evaluating educational programs. The Urban Review, 1969, 3, 15-17.

66. Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Los Altos, Calif.: RMC Research Corp., 1974. ED 096 344

67. Thonis, E. W. Evaluating the effectiveness of programs designed to improve the education of Mexican-Americans. Sacramento, Calif.: California State Department of Education, 1971. ED 062 047

68. Tosi, H. L., & Carroll, S. J. Management by objectives. Personnel Administration, July-August, 1970, 44-48.

69. Tosi, H. L., Rizzo, J. R., & Carroll, S. J. Setting goals in management by objectives. California Management Review, 1970, 12, No. 4, 70-78.

70. Valentine, I. E., & Larsen, M. E. A systems approach for implementing an institutional research program. Journal of Industrial Teacher Education, 1974, 12, 38-48.

71. Weiss, Carol H. Evaluation research: methods for assessing program effectiveness. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

72. Weiss, Carol H. Evaluating action programs: readings in social action and organization. Boston: Allyn & Bacon, 1972.

73. Weiss, C. H. Where politics and evaluation research meet. Evaluation, 1973, 1, 37-45.

74. Weiss, Robert S., & Rein, Martin. The evaluation of broad-aim programs: experimental design, its difficulties, and an alternative. Administrative Science Quarterly, 1970, 15, 97-109.

75. Welch, Wayne N. The process of evaluation. Journal of Research in Science Teaching, 1974, 11, 175-184.

76. Wellisch, J. Problems and approaches in ESAA data collection. AERA 1974. ED 095 199

77. Wexley, K. N., & Yukl, G. A. Performance appraisal and management by objectives. Chapter 5. Organizational Behavior and Industrial Psychology. New York: Oxford Univer. Press, 1975.

78. Wortman, P. M. Evaluation research: a psychological perspective. American Psychologist, 1975, 30. (In Press.)

Items followed by an ED number (for example, ED 049 291) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of Resources in Education for the address and ordering information.