DOCUMENT RESUME

ED 164 565 TM 007 734

AUTHOR Eichelberger, R. Tony
TITLE Multiple Stakeholders and Evaluation.
PUB DATE 29 Mar 78
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Accountability; Control Groups; Educational Programs; *Evaluation Criteria; *Evaluation Methods; Evaluative Thinking; *Information Utilization; Program Evaluation; *Research Design; *Research Problems; *Values

ABSTRACT

Evaluations occur within a political decision-making mileau, where multiple stakeholders are contending for limited funds. Given the subjective basis of empirical information, different conclusions or recommendations about a program may result from different ideological, theoretical, and disciplinary perspectives. The logic behind the interpretation of results, and the assumptions necessary for such interpretations, must be specified and explained to facilitate the most appropriate use of an evaluation. Because of the complexity of many statistical techniques presently used, much work is needed to identify what assumptions must be met for meaningful and useful interpretations of results in a specific decision-making situation. The rationales for both the inclusion and the exclusion of the variables to be considered in an evaluation should be made explicit. The problem of obtaining a matched control group is often nearly impossible. The relationship between the statistical analysis and the evaluation question is often based on tenuous assumptions. The evaluation of Project Follow Through is used to exemplify these problems. (Author/CTM)

Multiple Stakeholders and Evaluation

by

R. Tony Eichelberger

Learning Research and Development Center

University of Pittsburgh

Session #18.18

Presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 29, 1978.

This paper includes material worked out jointly with others, especially Nancy Cole and James L. DiCostanzo. The ideas in the paper have also benefited greatly from discussions with William W. Cooley, Gaea Leinhardt, members of the LRDC Follow Through program, and members of other Follow Through Sponsors' staffs. The editing of this paper and earlier drafts by Connie Faddis was extremely useful.

Multiple Stakeholders and Evaluation

R. Tony Eichelberger
Learning Research and Development Center
University of Pittsburgh

The primary distinction between evaluation and other empirically-based activities identified as research is that evaluation serves decision making, or is decision-oriented, while other research is conclusion-oriented (Cronbach & Suppes, 1969). In "models" of educational evaluation, decision making is usually viewed as the work of a single decision maker or decision-making body. When evaluators actually become involved with educational institutions or agencies, they often discover that there are multiple stakeholders in the decision situation, rather than just the specified decision maker. For example, in every school system, the administrators and school board are ultimately responsible to the community, and must be responsive to it. Results of an evaluation study must be credible to the various stakeholder groups in the community, or the results will not be useful.

Stakeholder groups usually have different ideological, theoretical, and practical perspectives. A simplistic example is the "humanist" groups versus the "back-to-basics" groups in many of our communities. Each has a different view of the primary functions of the schools and each group assigns different weights to educational outcomes. For an evaluation to address the most important issues in a specific setting and remain credible to the various stakeholder groups is a difficult task.

This very real problem has been identified by many evaluators, and several authors have recommended procedures for systematically dealing with the discrepant views and values of different stakeholders (Edwards & Guttentag, 1975; Stake, 1975; Stenner, 1976). Edwards and Guttentag suggest a "multi-attribute utility analysis" that takes the differing values into account when setting up the evaluation plan and analysis procedures. Stake has described an approach that he calls "Responsive Evaluation." According to Stake,

> An educational evaluation is responsive evaluation (1) if it orients more directly to program activities than to program intents, (2) if it responds to audience requirements for information, and (3) if the different value perspectives of the people at hand are referred to in reporting the success and failure of the program. (1975, p. 10)

Stenner uses "Policy Implications Analysis," which asks members of the various stakeholder groups to identify the types of evaluative information and appropriate reporting formats that they would like from the evaluation at the end of the program (or other future date specified).

Coleman (1972) discusses a related problem of the limitations of any one evaluation study, especially if it is carried out from only one theoretical or disciplinary perspective. Every decision situation can be viewed from various perspectives, each of which may lead to very different decisions. His suggestion that a number of concurrent evaluations take place, using different theoretical and disciplinary bases, is especially pertinent for dealing with multiple stakeholders.

In decision situations, various groups are contending for limited resources. Each group will use whatever information is available to support

its own position, regardless of the quality of the data. If the evaluator wants to contribute information useful to the decision-making process, (s)he must attempt to represent the major differing perspectives and report the information as comprehensively and accurately as possible. My own experiences as an evaluator, combined with the experiences recorded in other major evaluation reports, have revealed that this task is an extremely difficult one.

A major problem in meeting the needs of the various stakeholders is that methodology is often used without recognizing the assumptions that are required for meaningful interpretation of the results for the specific situation. This relates to the role of "methodologist" as discussed by Lazarsfeld and Rosenberg (1955):

> The term methodology . . . implies that concrete studies are being scrutinized as to the procedures they use, the underlying assumptions they make, the modes of explanation they consider as satisfactory. (p. 3)

In order to report data accurately and to make appropriate interpretations of results, a number of fundamental considerations about the use and interpretation of empirical information is needed.

Evaluation Data and Interpretation

In this paper several issues related to accurate specification of data and interpretation of results are discussed. The focus of these comments is to make an evaluation report more readily interpretable by readers who are not experts in research methodology or evaluation, which is often the case with many educators and "lay" groups. Several types of information are recommended for inclusion in reports, and examples are given

where possible. The issues addressed deal with (a) the subjective basis of all data and its interpretation, (b) rationales for including variables and measures in an educational evaluation, (c) assumptions required for meaningful interpretation of data in a specific setting, and (d) inadequacies of the experimental paradigm for evaluations.

Subjective Basis of Empirical Information[1]

In general, laws, theories, variables, and measures in the behavioral sciences are man-made conceptualizations. For example, what is "reading" from one perspective is "symbol processing" from another. The measures that would be used for assessing each would be very different, and different researchers could use quite different measures of the same variable.

Another example of the subjective basis of the interpretation of data is the varied uses of the Coloured Progressive Matrices Test. Raven (1962) developed it as a measure of nonverbal IQ, which was believed at that time to be a genetic characteristic. The test scores were used in the evaluation of the National Follow Through Program (FT) as a nonverbal problem-solving measure, skills assumed to be learned and affected differentially by the various instructional models.

In the design stage of an evaluation, the evaluator must decide which variables and measures to include. These decisions are based on the evaluator's perspective of the program, its context, and its purpose. This view is often discipline-based. For example, the evaluation of Follow Through, like most educational evaluations, utilized only academic achievement, self-concept, and individual responsibility measures. This represents

primarily a psychology-based view of education. In order to make sense out of an evaluation such as that of Follow Through, the reader must be aware of the evaluator's perspective of the program and the evaluation, and the evaluator's rationales for including whatever information is presented.

In reviewing modern developments in the philosophy of science,[2] Campbell (1974) indicated that:

> Non-laboratory social science is precariously scientific at best. But even for the strongest sciences, the theories believed to be true are radically underjustified and have, at most, the status of "better than" rather than the status of "proven." All common-sense and scientific knowledge is presumptive. In any setting in which we seem to gain new knowledge, we do so at the expense of many presumptions, untestable--to say nothing of unconfirmable--in that situation. While the appropriateness of some presumptions can be probed singly or in small sets, this can only be done by assuming the correctness of the great bulk of other presumptions. Single presumptions or small subsets can in turn be probed, but the total set of presumptions is not of demonstrable validity, is radically underjustified. (p. 2)

Conclusion-oriented researchers have the freedom, if not the responsibility, to carry out their studies within a well-defined theoretical perspective in order to test the theory and contribute to knowledge within that perspective, regardless of the extent to which it is justified by empirical evidence. Evaluators in real-life situations have the responsibility of providing information useful to <u>that</u> situation. In order to do this, the presumptions upon which the data and their interpretations are based must be specified, and the extent to which they are met in a particular situation must be estimated. The presumptions, or assumptions, and the extent to which they are met in an evaluation, are discussed in the next two sections of this paper.

Rationales for Variables and Measures in an Evaluation[3]

Whenever an evaluation is planned, a wide range of variables and measures are initially identified for possible inclusion. Some variables and measures are inevitably excluded during the selection process. The evaluation contractor is usually most knowledgeable about the compromises and deletions that are made at this time. Unfortunately, a discussion of the selection process is seldom, if ever, included in an evaluation report. Thus, the best thinking about this problem and the rationales for the decisions are lost to the field and to society. They are also not available to the stakeholders, who need that information so that they can more appropriately assess the relative value of the evaluation's conclusions as they relate to decision alternatives. Without such a discussion, only the most knowledgeable reader will be able to recognize the limited nature of the evaluation, and weight the possible alternatives appropriately.

A good example of such a discussion appeared in Design for the Individualized Instruction Study (Cooley & Leinhardt, 1975). The rationale for excluding noncognitive variables in the evaluation design was included as an apprendix to that study. It indicated the steps that were followed and the criteria that were used to arrive at the recommendations. Cooley and Leinhardt also presented their rationale for using a standardized achievement test to assess cognitive outcomes. The criteria utilized to compare possible tests were delineated. The actual test reviews were included as another appendix, in which the subtests of each achievement battery, the psychometric characteristics, the available norms, and other characteristics were described.[4]

There are many pressures on evaluation contractors to make an evaluation appear as comprehensive and competent as possible. Thus, the omission or weaknesses of the chosen set of variables and associated measures are seldom presented and discussed. When they are not presented, the reader may be left with the impression that all important variables were included in the evaluation, and the procedures used did measure them adequately (if not comprehensively). As a result, the particular groups that these measures favor will use the results to fight for a decision that supports their position and give them more of the resources.

A good example of this type of use by a stakeholder involved the use of Follow Through (FT) evaluation results (Stebbins, 1976) by the Oregon FT Program and SRA (the publisher of DISTAR, a central component of the Oregon model). These results were immediately put into a short paper indicating that Oregon was the one successful FT model, yet no plans were being made to provide additional funds for dissemination of this program. SRA disseminated these results broadly.

A closer reading of the FT evaluation results would indicate the limited sense in which the Oregon model was the "most successful." The lack of clear articulation of the sense in which the model was "best" gave Oregon and SRA the license to use the evaluation results and language to their best political advantage.[5]

In interpreting evaluative data, stakeholders may use it in ways that are inappropriate in the view of the evaluator (although I am not saying that was the case with the Oregon and SRA uses and interpretations). However, in any situation where the rationales and caveats do not appear

appropriately in the report, the evaluator must take some responsibility for any misuse.

Assumptions Required for Meaningful Interpretations of Evaluative Data

As indicated previously, all quantitative data are based on presumptions about the data. Some of these are often the assumptions of the particular statistical technique used to analyze the data. For example, the usual parametric assumptions about data for analysis of variance (ANOVA) include:

1. Independent observations
2. Populations are normally distributed
3. Populations have equal variances
4. Variables are measured on interval or ratio scales.

If these assumptions are adequately met, then ANOVA results can be meaningfully interpreted. Much is known about the effects on ANOVA when data do not precisely meet the assumptions, and that knowledge must be considered when deciding about the adequacy of the data in a specific situation. When more sophisticated techniques, such as multiple regression or ANCOVA are used, the assumptions are more numerous and the effects of failing to meet them precisely are usually not accurately specified. (See DiCostanzo & Eichelberger, 1977, for a discussion of information needed to assess assumptions required by ANCOVA.)

In most settings, numerous other assumptions must also be met if interpretations that are meaningful for decision making are to be made. In general, these involve threats to internal and external validity, as described by Campbell and Stanley (1963) and Bracht and Glass (1968). These threats

are seldom controlled adequately in any natural setting--especially one as complex as the educational setting, where buildings, teachers, administrators, and the social contexts of the schools vary so greatly. The particular strengths and weaknesses of the analytic techniques used and the confidence that one can have in the results and interpretations must be specified.

In addition to these logical concerns, the relationship between the analyses being carried out and the evaluation question being addressed is based on assumptions about the data and the education setting that are often tenuous. For example, in the FT evaluation the two major concerns to be addressed originally were:

1. Assessing program impact on pupils, parents, schools, and community (Emrick, Sorensen, & Stearns, 1973, p. 72).

2. Assessing relative effectiveness of different programs and program approaches (Sorensen & Madow, 1969, p. 4).

The evaluation design on which the FT final report was based essentially involved measuring pupil outcomes (academic achievement, self-concept, and individual responsibility for learning) at the end of third grade for pupils in the FT classrooms, and comparing the results to the outcomes for "similar"[6] students not in FT classrooms. The differences in outcomes, after numerous covariates were used to adjust results statistically, were identified as "program effects." In order to interpret the results as program effects, a number of assumptions had to be met. The simplest two, for illustrative purposes, were that: (a) the groups were initially similar, and (b) whatever differences obtained were due to different educational experiences of students.

If the first assumption of the evaluation design, that the two groups were similar, was met, then the general question of the impact of the total national FT program was addressed to some extent. (Keep in mind that the FT program included psychological, medical, dental, and nutritional support, as well as the use of classroom aides, etc., and not merely innovative educational programs.)

If the second assumption, that the children experienced different educational programs, was also met, then the second evaluation question was also addressed to some extent. The evaluators did not attempt to identify the differences in the educational experiences of the two groups (FT and non-FT) that were tested. All that is known about these students is that one group participated in classrooms identified as "Follow Through" and the other group did not. Other factors also existed that question the adequacy with which that second question was addressed. For example, the program effect was measured by the adjusted differences between a single FT site and its non-FT comparison group; thus, each value was on a different metric, and the relative effectiveness could not be addressed directly.

In the FT evaluation report (Stebbins, 1976) these assumptions were not specified, although some information was provided that described the similarity of the groups compared. The tenuousness of the inferences from the data to the interpretation of results, as they related to the evaluation questions, was not presented in the report, however. This left the impression that the evaluation questions were indeed adequately (if not comprehensively) addressed.

Inadequacies of the Experimental Paradigm

American society, and especially the academic community, have been oversold on the applicability of the experimental method to address almost any type of question in any type of setting. It is such a pervasive belief that if there is no control group nor tests of statistical significance in an evaluation study, the study is immediately suspect. This view is reflected in a quote from the evaluation of FT:

> It is an axiom of evaluation that in order to attribute observed outcomes conclusively to a program, children who participate in the program must be compared to similar children who do not. (Stebbins, 1976, p. A-45)

Numerous authors have discussed the inadequacies of the experimental paradigm for educational research and evaluation in natural settings (e.g., Guba, 1965, 1977, Edwards & Guttentag, 1975). A major problem with the experimental paradigm is its assumption that a program is static rather than dynamic, (i.e., the situation is such that an identifiable independent variable is operating). Guba (1965) questions the value of this assumption for educational programs (because programs must adapt to the educational requirements of different kinds of students); and, he also questions the likelihood that the assumption is usually met.

Edwards and Guttentag (1975) point out four kinds of dynamic changes that occur in educational programs:

1. The values of those served by the program and those who operate the program change.
2. The program evolves--changes shape and character.
3. The external circumstances to which the program is a response change.

4. Knowledge of program events and consequences change (p. 415).

Each of these four types of changes occurred within the FT program--as they will in any longitudinal program. Every FT sponsor's program changed over the years. For example, FT sites associated with the Learning Research and Development Center (LRDC) adapted the Center's instructional materials to meet their particular needs. In addition, some major changes were made in the content of the kindergarten and first grade curricula across all sites. These changes were partly based on knowledge of events and consequences at the sites, and were partly normal evolutionary changes. Also, in 1967-68, American society viewed the Head Start Program as a positive first step in compensatory education, which Follow Through was to continue. The original evaluation issue in FT was to develop and identify the "best" or the "successful" models. Later, in 1975-77, the value of all compensatory education was being questioned, and the desired outcomes of primary education tended to expand beyond the reading and math skills emphasized by FT and measured by the standardized tests used in the evaluation.

In addition to these dynamic problems, it is frequently impossible to obtain a group of truly comparable groups of children in stable circumstances that allow only the program or other treatment variable to operate. When Richard Anderson, Director of the FT evaluation for Abt Associates, was asked (at the 1977 AERA convention) whether he would use a control group if he were to do anything like the FT evaluation again, he indicated that he would not, because of the many problems that were experienced in obtaining comparable groups.

The appropriateness and adequacy of evaluative information provided for decision making must be assessed in each situation. The tenuousness of interpretations from complex experimental designs with sophisticated multivariate analyses must be recognized and reported accurately. In the words of John Tukey (1954), "Experimental statisticians should be honest and expository about the relation of precise assumptions and exactly optimum solutions to real situations" (p. 719). The same types of assessments of results from "responsive" and other types of evaluations are also needed. This is work for the methodologist as identified by Lazarsfeld and Rosenberg (1955).

Summary

The thrust of this paper has been to point out that evaluations occur within a political decision-making mileau, where multiple stakeholders are contending for limited funds. Given the subjective basis of empirical information, different conclusions or recommendations about a program may result from different ideological, theoretical, and disciplinary perspectives. The logic behind the interpretation of results, and the assumptions necessary for such interpretations, must be specified and explained to facilitate the most appropriate use of an evaluation.

Each of the issues raised in this paper need further study and explication if evaluators are to learn how to provide the most useful information for decision making. Because of the complexity of many statistical techniques presently used, much work is needed to identify what assumptions must be met for meaningful and useful interpretations of results in a specific decision-making situation. The persons best prepared to do this fundamental work are

probably research methodologists who are not practicing evaluators. Perhaps we can coax our colleagues in research methodology to join us in doing such needed work.

Footnotes

[1]Scriven (1972) discusses qualitative and quantitative sense in "Objectivity and Subjectivity in Educational Research."

[2]Campbell (1974) argues that the qualitative basis of quantitative data must be recognized and that both types of data are needed as cross-validating sources.

[3]Much of this discussion is taken from a paper written with James L. DiCostanzo (DiCostanzo & Eichelberger, 1977).

[4]One oversight, in my view, was the lack of some discussion of the inadequacies of the test battery that was selected by Cooley and Leinhardt for use in the evaluation.

[5]Coleman (1972) differentiates between the world of action and the world of the disciplines. It is my view that in the world of action, persons bright enough to recognize such an opportunity (as in the FT evaluation) would consider it foolish not to seize the opportunity. The Oregon and SRA usage is an example of stakeholders using whatever information is available to support their position.

[6]The degree of similarity has been a continuing problem for the FT evaluators. Such problems are identified for one program Sponsor by Eichelberger (1977).

References

Bracht, G. H., & Glass, G. V The external validity of experiments. American Educational Research Journal, 1968, 5(4), 437-474.

Campbell, D. T. Qualitative knowing in action research. Paper presented at a meeting of the American Psychological Association, New Orleans, September 1974.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally & Co., 1963.

Coleman, J. S. Methodological principles governing policy research in the social sciences. Address prepared for the 139th meeting of the American Association for the Advancement of Science, Washington, D. C., December 29, 1972.

Cooley, W. W., & Leinhardt, G. Design for the individualized instruction study: Final report. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1975. [NIE Contract No. 400-75-0071, 1975]

Cronbach, L. J., & Suppes, P. (Eds.) Research for tomorrow's schools: Disciplined inquiry for education. New York: Macmillan, 1969.

DiCostanzo, J. D., & Eichelberger, R. T. Design, analysis and reporting considerations when ANCOVA-type techniques are used in evaluation settings. Paper presented at a meeting of the American Educational Research Association, New York, April 1977.

Edwards, W., & Guttentag, M. Experiments and evaluations: A re-examination. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment. New York: Academic Press, Inc., 1975.

Eichelberger, R. T. Comments on the national evaluation of Follow Through: Some lessons learned. Paper presented at Follow Through Sponsors' conference at Sea Island, Georgia, July 11, 1977.

Emrick, J. A., Sorensen, P. H., & Stearns, M. S. Interim evaluation of the National Follow Through Program 1969-1971. A technical report. Menlo Park, Calif.: Stanford Research Institute, 1973. [USOE Contract No. OEC-0-8-522480-4633(100)]

Guba, E. G. Methodological strategies for educational change. Paper presented to the Conference on Strategies for Educational Change, Washington, D.C., November 8-10, 1965.

Guba, E. G. Educational evaluation: The state of the art. Invitational address presented to the Evaluation Network Conference, St. Louis, September 27, 1977.

Lazarsfeld, P. F., & Rosenberg, M. (Eds.) The language of social research: A reader in the methodology of social research. New York: The Free Press, 1955.

Raven, J. C. Coloured progressive matrices (sets A, Ab, B). London: E. T. Heron & Co., Ltd., 1962.

Ross, L., & Cronbach, L. J. Handbook of evaluation research. (Review of M. Guttentag & E. L. Struening, Eds., Handbook of evaluation research, 1975). Educational Researcher, 1976, 5(10), 9-19.

Scriven, M. Objectivity and subjectivity in educational research. In L. G. Thomas (Ed.), Philosophical redirection of educational research. In seventy-first Yearbook of the National Society for the Study of Education, Part I. Chicago: The University of Chicago Press, 1972.

Sorensen, P. H., & Madow, W. G. A proposal for research: Longitudinal evaluation of the National Follow Through Program, 1969-70. Menlo Park, Calif.: Stanford Research Institute, June 13, 1969.

Stake, R. E. Program evaluation, particularly responsive evaluation. Kalamazoo: Western Michigan University, Evaluation Center, 1975. (Occasional Paper Series No. 5).

Stebbins, L. B. (Ed.) Education as experimentation: A planned variation model (Vol. 3). Cambridge, Mass.: Abt Associates, Inc., 1976. [USOE contract No. 300-75-0134]

Stenner, J. Policy implications analysis: Design for evaluation of the state capacity building program in dissemination. Durham, N. C.: National Testing Service, Inc., 1976.

Tukey, J. Unsolved problems of experimental statistics. Journal of the American Statistical Association, 1954, 49, 706-731.