DOCUMENT RESUME

ED 353 266 TM 018 967

AUTHOR Nevo, David; Friedman, Etel
TITLE Relevance and Methodological Adequacy: A Study of Evaluation Reports.
PUB DATE Apr 92
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Compensatory Education; Definitions; Disadvantaged Youth; Educational Assessment; *Educationally Disadvantaged; Elementary Education; Evaluation Criteria; *Evaluation Methods; Evaluators; Foreign Countries; Literature Reviews; *Program Evaluation; *Research Methodology; Research Reports; Special Needs Students; Theory Practice Relationship
IDENTIFIERS *Evaluation Reports; *Israel; Relevance (Evaluation)

ABSTRACT

The relationship among various characteristics of evaluation studies and their overall quality was studied using all of the evaluation studies (N=366) conducted in Israel over 30 years on special projects for disadvantaged students at the elementary school level. The following variables were defined to describe each study: (1) evaluation definition; (2) role of the evaluation; (3) object of the evaluation; (4) variables of the evaluation; (5) evaluation criteria; (6) clients and audiences; (7) evaluation process; (8) research methods; (9) types of evaluators; and (10) evaluation standards. The analysis showed that judgmental definitions of evaluation were more common among academic evaluators and in summative evaluations. Descriptive definitions were more often used by practitioners. Surveys and quasi-experimental designs were the most common research methods, and most evaluations were conducted by external evaluators. There was a positive relationship between the methodological adequacy of the evaluation reports and their level of relevance. Results suggest that building on the academic educational research community could be the best approach for program evaluation in Israel. There is a list of 15 references. (SLD)

RELEVANCE AND METHODOLOGICAL ADEQUACY:

A STUDY OF EVALUATION REPORTS

David Nevo and Etel Friedman

Tel Aviv University

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 20-24,1992.

BACKGROUND

The purpose of this study was to investigate the relationship among various characteristics of evaluation studies and their overall quality. The study was conducted within the framework of a national educational system of a small country and included all the evaluation studies that have been conducted during a period of over 30 years on special projects for disadvantaged students at the elementary school level.

Specifically, the study tried to address the following questions: (a) to what extent can actual evaluation activities be fully described by theoretical variables suggested by the evaluation literature?,(b) what are the inter-relationships among the various characteristics of the evaluation activities?, and (c)what is the relationship between the characteristics of an evaluation activity and its overall quality?

A comprehensive review of the literature on evaluation in education (Nevo,1983;Glasman & Nevo,1988) revealed ten dimensions representing conceptual issues addressed by major evaluation approaches in an attempt to clarify the meaning of educational evaluation and describe its major characteristics.These ten dimensions were used here as a basis for defining the major variables by which the various evaluations have been

described.Thus, the following variables were defined to describe each evaluation study:(a)evaluation definition, (b)role of the evaluation, (c)object of the evaluation, (d)evaluation variables, (e)evaluation criteria, (f)clients and audiences, (g)evaluation process, (h)research methods, (i)types of evaluators, and (j)evaluation standards. The standards by which we judged the overall quality of an evaluation study were based partially on the general scope of the Joint Committee on Standards for Educational Evaluation (Joint Committee,1981),relating to accuracy,or methodological adequacy, utility,or relevance,and readability, or clarity of the evaluation report.

Of special interest might be the fact that this study tried to apply evaluation concepts, that were developed and very much grounded in the American context, to analyze actual evaluation activities conducted in a different educational and social context of another country. We would like to suggest that important lessons can be learned from this study regarding the conduct of evaluation in both societies, in addition to its potential contribution to our overall understanding of the concept of evaluation.

METHODOLOGY

The data on which this study is based has been obtained from a systematic analysis of 366 evaluation reports that were produced during a period of over 30 years in the State of Israel,in relation

to special projects and programs for disadvantaged students in elementary schools. To avoid a priori elimination of low quality reports, we defined an "evaluation report" as "a written document presenting some data regarding the quality of a specific evaluation project or program." Some of the documents were only a few pages long, but most of them were full-blown reports. Limited resources confined our study to projects and programs that were originally developed for culturally disadvantaged students at the elementary school level.

Comprehensive search procedures were used to assure a complete collection of all existing evaluation reports. This procedure included computerized and manual searches in data bases and library catalogs of major research institutes and universities,reviews of major periodicals in education and related fields,and direct interaction with researchers and practitioner, who have been known for their involvement with relevant projects or programs.The bibliographies of all evaluation reports,that have been obtained,were also searched for additional relevant reports. The considerable number of evaluation reports (N=366),that we succeeded to collect, was a surprise not only to us but also to the people in the Ministry of Education,who couldn't believe that there were so many evaluation reports that they have ignored through out the years...

The research variables,delineated from the above mentioned

conceptual framework, were defined operationally and developed into a structured scoring guide which was then used to score all 366 evaluation reports. The scoring was done by trained scorers who reached a 76 percent inter-scorer agreement.

Data were analyzed using descriptive statistics,chi square and multiple regression with dummy variables.

FINDINGS

Following are some findings related to our research questions. A complete presentation of findings can be found in the technical report of the study (Nevo & Friedman, 1991).

1. Characteristics of the evaluation activities

1.1 Definition

Usually, evaluation reports did not provide an explicit definition of evaluation, but in most cases it was possible to infer how the author(s) perceived the concept of evaluation. We classified the various perceptions of evaluation into four major groups following some "classical" definitions suggested in the literature: (a) the Tylerian definition (Tyler, 1950) perceiving evaluation as an activity intended to determine the extent that goals have been achieved, (b) the "decision making definition"

(Stufflebeam, et al., 1971; Alkin, 1969), defining evaluation as providing descriptive information to serve decision making, (c) the judgmental definition (Scriven, 1967; Joint Committee, 1981) defining evaluation as the assessment of merit or worth,and (d) the definition which perceives evaluation as a combination of both,description and judgment (Stake, 1967; Guba and Lincoln, 1980).

The Tylerian definition was found to be the least common definition; only 8.4 percent of the reports focused on an attempt to determine the extent a project or a program achieved its goals.The most common definitions were the judgmental definition (34.0 %) and the decision making definition (31.7%), followed by the last definition, combining description and judgment (25.8 %). Thus,description and judgment seem to be two major components of evaluation studies with very little direct interest in goal achievement. This might be of special interest since other studies have shown that the Tylerian definition of evaluation seems to be very popular among Israeli teachers and school administrators (Nevo & Goldblat, 1988).

As we will see later on,the way evaluation is defined or perceived in an evaluation report is related to many other characteristics of that report.

1.2 Evaluation role

Like in the case of the definition,only a small number of reports had an explicit statement of the function that the evaluation was intended to serve,but some sense of purpose could be inferred from reading most (94.6 %) of the report. And since our study was based on reviewing written reports rather than interacting directly with the evaluators, we were unable to relate to the actual role of the evaluation,nor could we be aware of any latent functions that have been served by the evaluation,such as motivational and/or socio-political functions. Thus,our analysis was limited to two roles that the various evaluations intended to serve: the formative role,to provide feedback for improvement,and the summative role,for accountability,selection or accreditation.

Our analysis revealed that 42.9 percent of the evaluations were intended to serve summative functions,18.6 percent - formative functions, and the rest (38.6 %) tried to serve both, formative and summative functions. Obviously, there is a possibility that more formative evaluations were conducted over the years,but they were never published in some kind of written report that we could be aware of. On the other side, it is also possible that part of the reports that seemed to be intended to serve both functions, were actually summative evaluations,which paid some lip service to formative evaluation to obtain support and cooperation.

1.3 Evaluation variables

We tried to identify the variables that were investigated regarding each evaluation object. A distinction was made between goals,designs,processes,and outcomes. We found that the major focus was on outcomes.

Outcomes were investigated extensively in 71.7 percent of the reports; processes of implementation - in 44.7 percent of the reports; designs and strategies - in 15.7 percent; and goals were assessed only in 7.8 percent of the reported studies. Thus,in spite of what has been suggested in the evaluation literature in the last two or three decades,the major focus is still on outcomes with some concern for process evaluation, but very little in assessing the quality goals that were set and strategies that were chosen to achieve those goals. We have also found that the focus on outcomes was significantly stronger in evaluations of instructional materials than in other evaluations,such as evaluations of extra-curricular activities,teacher in-service training programs,or community projects.In evaluations of instructional materials, outcome evaluation meant usually assessment of student achievements.

1.4 Clients and audiences

Following client oriented approaches in evaluation,we tried to

identify the extent to which the evaluation reports were intended to serve specific clients and audiences. We found that 54 percent of the reports were addressed to specific clients,usually departments in the Ministry of Education or designated groups of educators.Most of the reports (60 %) that did not address any specific audience were written by university faculty members and graduate students,published as research articles,Master theses,or doctoral dissertations.On the other hand,73 percent of the reports that were written by professional evaluators (see 1.6) were directed to specific clients.

Several differences in structure and content were found between evaluation reports that were addressed to specific client or audiences and those that were not.

Regarding the structure of the report we made a distinction between (a)technical research type reports, (b)reports for policy makers and administrators, and (c)reports for the general public.Most of the reports (81 %) were of the first type,most appropriate for researchers,and members of the so called "scientific community",rather than policy makers,administrators,and the general public. Only 6.6 percent of the reports were written in a structure adapted to the needs of decision makers, and the rest were written in a narrative style appropriate for the general public. But a significant difference was found in the percentage of decision oriented reports among the reports that were intended

to serve specific clients (10 %), and those that did not address any specific audience (2 %).

Significant differences between the client-oriented and the non-client oriented reports were also found regarding the use of oral and interim reports throughout the evaluation process, and the inclusion of recommendations in the evaluation report. Only in 8.8 percent of the reports that were analyzed, some kind of interim report or oral presentation of findings was mentioned but in the client oriented reports this percentage was twice as much (11 %) as in the non-client oriented reports. It was also found that 28 Percent of the client oriented evaluation reports included specific recommendations for improvement, while only 13 percent of the other reports included such recommendations.

1.5 Research methods

A variety of methodological tools were used to conduct the various evaluations represented in the 366 evaluation reports that we have analyzed. We classified 58 percent of them as quantitative studies, 13 percent - as qualitative, and 29 percent as a combination of both. Over forty percent of the studies (42.4 %) were surveys, 43.8 percent used an experimental or quasi-experimental design, and 13.8 percent were case studies.

A variety of measurement instruments and data collection

procedures were used,the most popular being questionnaires (48.5%),achievement tests (42.9 %), and observations (33.9 %). Less popular were interviews (26.2 %), and psychological tests (23.3 %), and the least frequently used were content analysis (14.8 %) and expert opinion (4.1 %).

Interesting data was also obtained in relation to the psychometric qualities of the measurement instruments that were used in the various studies. Surprisingly enough,most evaluations did not report information regarding the reliability and validity of their measurement instruments.The best in this regard seemed to be those evaluations which used achievement tests; 36.7 percent of them reported some kind of information regarding the validity of their tests, and only 22.4 percent - regarding reliability. Those that have used psychological tests referred to the validity of those tests in 26.7 percent of the cases and to reliability only in 22.3 percent of the cases. The situation with other instruments was even worse. We found some differences between university based evaluators and other evaluators,but even at the university, where the concepts of reliability and validity are being widely taught and learned,not too many evaluators seem to be applying those concepts in their studies.

We have also looked into the sampling procedures that were used in the various evaluation studies.Only about one quarter to one third of the studies used some kind of defined sampling procedure,

about the same proportion of the reports did not include any information on their sampling procedures, and the rest used accidental samples, not applying any defined sampling procedure.

As to data analysis procedures we followed the classification suggested by Goodwin and Goodwin (1985), and found that 55.8 percent of the studies used elementary statistics,25.5 percent used intermediate or advanced statistical procedures,11 percent of the reports used qualitative analysis, and in the rest (5.7 %) it was not clear what procedures were used.

1.6 Types of Evaluators

Regarding the types of evaluators that conducted the various evaluations, we made a distinction between internal and external evaluators, and between professional and non-professional ("amateur") evaluators.We also tried to identify their level of formal training and their institutional affiliation.

We found that external evaluations were more prevalent (55.5 %) than internal evaluations (31.7 %). A combination of internal and external evaluation was found only in 1.9 percent of the evaluation reports, and in the remaining 10.9 percent of the reports we could not determine by what kind of evaluator they were conducted.

We considered evaluators to be "professional evaluators" by the extent that they had formal training in evaluation and evaluation

was a major component in their professional work. According to this definition,university professors and graduate students, who are not specializing in evaluation, were not considered as professional evaluators. This way only 22.7 percent of the evaluations were classified as been conducted by professional evaluators,18.0 percent by graduate students,and 37.7 percent by university professors (not in the field of evaluation), 7.4 percent by non-professional evaluators in the Ministry of Education,1.1 percent of the reports had a combination of both types,and in 13.1 percent of the reports we could not determine which type of evaluator conducted the evaluation.

some interesting relationships were found between the above mentioned two distinctions. About 68 percent of the professional evaluators and about 72 percent of the university professors not specializing in evaluation were external evaluators.On the other hand,two thirds of the non-professional evaluators, who were not university professors,were internal evaluators. About two thirds of the professional evaluators had a masters degree or a doctorate,while only 13 percent of the non-professional evaluators held similar degrees.

Most of the evaluations done by professional evaluators (76 %) or by university professors were conducted within the framework of universities or research institutes. And overall 83 percent of the evaluations were conducted by universities or research institutes.

The over all picture is that most evaluations are conducted externally,mainly by university professors,who do not specialize in evaluation, but also by professional evaluators who do specialize in evaluation. Most of the evaluations are conducted within universities and research institutes outside of the educational system.

2.Inter-relationships among the evaluation characteristics

The relationships among the various evaluation characteristics are still being analyzed, but some of them seem to be evident at the present time. From the perspective of the type of evaluators, who have conducted the various evaluation studies reflected in our reports, the following inter- relationships seem to be of special interest.

2.1 **Academic non-professional evaluators** (university professors and graduate students not specializing in evaluation) were inclined to use **judgmental definitions** of evaluation, they were involved mainly in **summative evaluation**, focused on **product evaluation**, used **quantitative research methods**,mainly **experimental** and **quasi-experimental designs**,and used **intermediate** and **advanced statistics** more than any other group. Most of their **reports were not addressed to specific clients or audiences.**

2.2. Non-academic amateur evaluators, working in the Ministry of Education or other non-academic institutions,were inclined to use descriptive (non-judgmental) definitions of evaluation,were involved mainly in formative evaluation, focused on process evaluation and on evaluation of strategies and designs, combined both qualitative and quantitative methods, use case studies more than any other group,and their statistical procedures were less sophisticated than those of any other group. Most of their evaluation reports addressed specific clients.

2.3 Professional evaluators (practitioners and academics) were inclined to use non-judgmental definitions of evaluation,did more process evaluation and did not focus their evaluations only on outcomes, but their research methods were similar to those used by academics (university professors and graduate students); they used more quantitative methods,more experimental and quasi-experimental designs, and more intermediate and advanced statistical procedures. Most of their evaluation reports were addressed to specific clients or audiences.

Thus, professional evaluators seemed to be providing, more than others, a combination of having a wide perspective on evaluation and using more advanced research methods,but they conducted only a small portion of the evaluations.

3. Relationships between the characteristics of the evaluation and their overall quality

As mentioned earlier, we have also rated the overall quality of each evaluation report on three dimensions: (a) relevance, (b)methodological adequacy, and (c) clarity. We examined the relationships among those dimensions, and the relationship between the overall quality of the reports their characteristics using chi square, correlations, and multiple regression with dummy variables.

3.1 Relevance

The relevance of a report was determined by the extent to which it addressed major issues pertinent to the evaluation object or its context. Using a three point scale, more than half (58%) of the reports were found to be at the higher level of relevance, and less than ten percent at the lowest level.

Significant relationships were found between the overall relevance of an evaluation and the types of variables it addressed, the research methods that it used,the format of the report, the type of the evaluator,and the overall methodological adequacy of the evaluation The more relevant evaluation reports addressed a wider range of evaluation variables (not only outcomes),used more adequate research methods,used technical research type reports and were typically conducted by external evaluators from universities or research institutes.

The multiple regression analysis also showed that the major variables that explained the variance of report relevance (20.5% of the variance) were: use of technical type reports,addressing more than one evaluation variable,evaluator from university or research institute,and the use of a survey research design.

3.2 Methodological adequacy

The overall methodological quality of an evaluation was judged holistically beyond specific reference to the adequacy of the research design,the reliability and validity of the measurement instruments and the sophistication of data analysis procedures. Less than half (45.3%) of the reports were classified at the high level of methodological quality, 38.1 percent at the intermediate level, and 16.7 percent at the lower level.

As mentioned earlier, significant relationships were found between the overall methodological adequacy of the reports and their relevance. The methodological quality of the reports was also related to other characteristics of the evaluation. The multiple regression analysis showed that the variables that had a significant contribution to the explained variance (explairing 36 % of the variance) were: technical research style report,evaluation conducted in a university framework,focus on outcomes and on summative evaluation, professional evaluator,using quantitative

research methods,and a judgmental definition of evaluation.

3.3 Report clarity

The third holistic rating of the reports was related to their clarity. Most of the reports (85.7 %) were rated by our scorers as clear and readable, few of them (11.3 %) as partially clear, and less than a dozen reports (3 %) were found to be unclear and confusing.

In spite of the lower variability on this variable some relationships were found with other variables. In the multiple regression analysis the variables that had a significant contribution to the explained variance (20.4 %) were: university non-professional evaluator, a judgmental definition of evaluation, evaluation conducted within a university or a research institute, and using a combination of qualitative and quantitative methods. However,the general notion that "university type" reports seemed to be clearer and more readable than "field originated" reports has to be qualified by the fact that our scorers were graduate students, familiar with academic writing and its jargon. The clarity level of the reports to other audiences has still to be determined.

SUMMARY AND DISCUSSION

Various concepts that have been suggested by American

evaluators, since the late sixties and early seventies, were found to be useful in analyzing evaluation activities conducted in the Israeli educational system. The analysis has shown that judgmental definitions of evaluation were more common among academic evaluators and in summative evaluations.Descriptive definitions of evaluation were used more by practitioners and in formative evaluations. Most evaluations focused on the assessment of outcomes and impacts. Some of them also assessed processes of implementation, but only a few assessed the designs and goals of the projects that were evaluated.

Surveys and quasi-experimental designs were the most common research methods used in the various evaluations, but samples were not representative in most cases. A small percentage of evaluators, even among university researchers,reported data regarding reliability and validity of measurement instruments. Evaluations that addressed specific audiences included more oral reports, and their written reports included more recommendations.

Most evaluations were conducted by external evaluators, usually graduate students and university professors whose area of specialization was not evaluation. The "professional evaluator", quite popular in the American context (e.g.,members of Division H at AERA), is not very common in the Israeli educational system.

A major finding of this study was the positive relationship between the methodological adequacy of the evaluation reports and their level of relevance. No support was found for the myth that "field based" evaluations tend to be more relevant than "academic"

evaluations. In our study, evaluations conducted by university evaluators appeared to be more accurate but also more relevant than those conducted by practitioners. Methodological adequacy seems to be a necessary condition for an evaluation to be relevant.

The distinction between methodological adequacy and relevance has been widely discussed in the evaluation literature as two major standards that all evaluation should meet. Although it was not suggested that methodological rigor is less important than relevance, the evaluation profession seemed to be especially proud of its relevancy and utility compared to "academic" (basic or applied) research. Thus, the Joint Committee on Standards for Educational Evaluation (1981) suggested four major evaluation standards, the first being "utility" and the last being "accuracy". The Joint Committee has never formally admitted that its standards were presented in order of importance. However, the order of the standards has been changed during the process of their development, and in the 1988 publication of the Joint Committee Personnel Evaluation Standards the "propriety" standards were moved up to the first place but the "accuracy" standards still remained last. Even if it would be only a symbolic act,our study suggests that the "accuracy" standards be upgraded to reflect a level of importance equal to "utility", and that should be the way accuracy should be treated in evaluation.

The association of evaluation with decision making,once strongly

advocated as a means to increase the relevance of evaluation (Cronbach, 1963; Alkin, 1969; Stufflebeam, et al., 1971),has been also criticized over the years for its oversimplification and its bureaucratization effect on evaluation (Cronbach,et al., 1980; Shadish,Cook & Leviton, 1991).On the basis of our findings regarding the differences between academic evaluators and practitioners, we are inclined to support such criticism and suggest that the American educational system might have gone too far in developing large evaluation bureaucracies at the state and school district level. Such bureaucracies, being released from high standards of academic research, on one side, and being detached from the school and classroom reality, on the other side, might end up producing evaluations that are not accurate nor useful. This proposition is not presented here as finding but rather as a recommendation for future research, which could benefit from our methodological experience.

As for the Israeli arena, this study provides some perspective for the typical inclination of the educational system to adopt procedures and approaches from the USA without due consideration. Our educational system could benefit very much from the American experience in conceptualizing the meaning of evaluation, developing its methods of inquiry, and getting a better understanding of its educational and social significance. But when mistakes are concerned, there is no need to repeat those of the American educational system; we could at least do our own mistakes,if we

can't avoid them. In a centralized system like the Israeli educational system it is very important to avoid as much as possible any further bureaucratization of evaluation activities. Our study suggests that building on the academic educational research community, while increasing its involvement in the educational system, and improving its ways of operation could be the way to go.

REFERENCES

Alkin, M.C. (1969). Evaluation theory development. Evaluation Comment,2,2-7.

Cronbach, L.J.(1963). Course improvement through evaluation. Teachers College Record,64, 672-683.

Cronbach, L. J., & Associates (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Glasman, N.S., & Nevo, D. (1988). Evaluation in Decision Making: The Case of School Administration. Boston: Kluwer.

Goodwin, L. D. & Goodwin, W.L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2),

Guba, E. G., & Lincoln, Y. S. (1989). Fourth generation evaluation. Newbury Park, CA: SAGE.

Joint Committee on Standards for Educational Evaluation (1981). Standards for evaluations of educational programs, projects, and materials. New York: McGraw-Hill.

Nevo, D.(1983). The conceptualization of educational evaluation: An analytical review of the literature. Review of Educational Research, 53(1),117-128.

Nevo, D., & Goldblat, E. (1988). Evaluation perceptions and school assessment in schools of Active Learning: A research report. Tel Aviv: School of Education,Tel-Aviv University (in Hebrew).

Nevo, D., & Friedman, E. (1991). A study of evaluation studies in Israel: An analysis of evaluation reports.Tel Aviv: School of Education, Tel Aviv University (in Hebrew).

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), Curriculum evaluation. AERA monograph series on curriculum evaluation 1. Chicago: Rand McNally.

Shadish, W. R., Cook, T.D., & Leviton, L. C. (1991). Foundations of program evaluation: Theories of practice. Newbury Park, CA: Sage.

Stake, R. E. (1967). The countenance of educational evaluation. Teachers College Record, 68, 523-540.

Stufflebeam, D. L., et al. (1971). Educational evaluation and decision-making. Ithaca, IL: Peacock Publishers.

Tyler, R. W. (1950). Basic principles of curriculum and instruction. Chicago: University of Chicago Press.