DOCUMENT RESUME

ED 238 917 TM 840 007

TITLE	On Evaluation Policy in the United States and Israel. Report No. A-112.
INSTITUTION	Northwestern Univ., Evanston, Ill. Dept. of Psychology.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Mar 81
CONTRACT	300-79-0467
GRANT	NIE-G-79-0128
NOTE	33p.; Paper presented at the joint Israel-American Seminar on Educational Evaluation (Jerusalem, Israel, June 1980). For related documents, see ED 192 466 and TM 840 004. This document is chapter 1 of TM 830 628.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Educational Policy; *Evaluation; Evaluation Methods; Evaluation Utilization; Evaluators; Federal Programs; *Foreign Countries; Guidelines; *International Educational Exchange; Standards; State Programs
IDENTIFIERS	*Israel

ABSTRACT

This paper has two aims, each bearing on recent developments in evaluation policy. The first is to summarize a report presented in 1980 to the United States Congress and Department of Education, concerning evaluation policy and practices at the national, state, and local levels of government. The second aim is to link some of the United States recommendations to ideas presented at the Israel-United States seminar on education evaluation, and in related papers by the seminar participants. The intent is (1) to outline similarities, differences, and analogs between the two perspectives in order to learn how the Israeli experience can be adapted by the United States, and vice versa; and (2) to examine how the problems evidenced in the arena of evaluation are not confined by national borders, ethnic origin, or history. (PN)

TITLE: On Evaluation Policy in the United States and Israel

AUTHOR: Robert F. Boruch

SUPPORTED BY: OED-300-79-0467 and NIE-G-79-0128

REPORT NO.: A-112

PRESENTED AT: Israel-United States Seminars on Evaluation, Jerusalem, Israel, June, 1980.

PUBLISHED IN: A. Lewy and S. Kugelmass (Eds.) Decision oriented evaluation in education: The case of Israel. Philadelphia and Rehovot: International Science Services Press, 1981, pp. 1-32.

NOTE: This is an expanded version of oral remarks made by the author at the Seminar.

DATE: March, 1981.

Northwestern University
Department of Psychology and the School of Education
Methodology and Evaluation Research
Evanston, Illinois 60201 USA

ON EVALUATION POLICY IN THE UNITED STATES AND ISRAEL[1]

Robert F. Boruch

1. INTRODUCTION

Attempting to understand which of several policies or programs has the greater benefit is not, of course, a novel human enterprise. Comparative tests to understand conditions under which childeren learn speech, for instance, were undertaken by the Arab conqueror Akbar the Great in 14th century India. Competing theories of human development fired rabbinic argument and theories of evidence during the same period in the Middle East. Nor are sophisticated logic and statistical theories underlying the fair comparison of programs especially new. They are represented in the 18th century scholars' attempts in Europe to understand numerical evidence and independent contributions during the same period to characterize the toxicity of metals, chemicals, and drugs. Finally, there are some distinctive early precedents for controlled field tests of social programs. They include experiments on the effects of sanitation instruction in Syria during 1931-33, and on the comparative benefits of raw and pasteurized milk in nutrition programs for English school children in 1930[2].

What is relatively novel about evaluation is the regularity of formal government interest in understanding the comparative effects of new social programs and increased government willingness to estimate effects in pilot tests of the programs. This interest in effectiveness is linked in principle to systematically establishing the need for programs and the quality of their implementation. The latter are no less important than estimating effects, but, until recently, had not been routinely required by law.

This paper has two aims, each bearing on recent developments in evaluation policy. The first is to summarize a report that we presented in 1980 to the U.S. Congress and Department of Education, concerning evaluation policy and practices at the national, state and local levels of government. The focus here is on recommendations and the treatment is very brief. The report itself (Boruch and Cordray, 1980) provides details, is readily accessible, and the literature review on which it is based has been published elsewhere (Boruch and Wortman, 1979).

The second aim is to link some of the U.S. recommendations to ideas presented at the Israel-U.S. seminar on education evaluations and in related papers by the seminar participants. This examination, too, is brief, rather too brief to do real justice to the ideas proposed. But the intention is simply to outline similarities, differences, and analogs between the two perspectives. One of my motives is pragmatic: to learn how the Israeli experience can be adapted by the United States, and vice versa. The second motive is based on the simple premise, suggested earlier, that problems of evidence in this arena are not confined by national borders, ethnic origin, or history. Understanding how durable problems and their solutions are is a task that, as a theoretician, I can ill afford to ignore.

2. THE HOLTZMAN REPORT TO THE CONGRESS

The Education Amendments of 1978 required that the Secretary of the U.S. Department of Education conduct a comprehensive review of federal evaluation practices and procedures. Introduced as a bill by Congresswoman Elizabeth Holtzman, the law directs attention to federally supported programs at the national, state, and local levels of government. In response, two projects were initiated by Department and Congressional staff. A group at Northwestern University was asked to undertake the first in September 1980. The National Academy of Sciences' Committee on Program Evaluation was asked to initiate parallel, independent work. The results of each are reported in Boruch and Cordray (1980) and Raizen and Rossi (1981) respectively[3].

The questions covered in Northwestern's report to the Congress and the Department of Education are fundamental. They were implied by the law and the conference reports preceding it:

- Why and how are evaluations carried out?
- What are the capabilities of those who carry out evaluations?
- How are the results of evaluation used?
- What recommendations can be made to improve procedure or practice?

The study was prospective in its orientation, designed to provide evidence and argument bearing on these questions and to provide recommendations which would help to ameliorate the problems that were identified. The findings and recommendations stemmed from two general sources of information: contemporary investigations by other researchers and agencies, and direct investigations by Project staff. The latter included site visits to local and state education agencies and telephone surveys of local units, both based on a stratified random sample. Round-table discussions were undertaken to

capitalize on special expertise in topics such as school board interest in evaluation. Interviews with some staffers of all major federal agencies with an interest in educational evaluation were carried out. This included the U.S. General Accounting Office and the Congressional Budget Office as well as the education agencies. The literature review covered both unpublished and published documents, including reports maintained by ERIC and, in the case of statute, by the LEXIS system. An earlier review served as a guide to sources on national studies published before 1979.

The project report made recommendations to the Congress and to the U.S. Department of Education. The two sets are condensed in the following treatment and coupled to a brief rationale for each. The links to Israeli work are discussed after each recommendation.

3. REQUESTING AND PLANNING EVALUATIONS

Three of the Report's recommendations concern the process of deciding what kind of evaluations can or should be done, and the way they should be done. They stress the necessity for regular meetings to establish information needs, the merit of specificity in evaluation law, and the reduction of constraints on exchange of information.

On clarifying needs, audiences and options

We recommended that the Congress direct the relevant staff of Congressional committees and support units such as the U.S. General Accounting Office and the Congressional Budget Office to meet with evaluation staff of the Department of Education regularly with instructions to: 1) identify specific committees and groups as audiences for evaluation results, 2) reach agreement about when particular evaluations are warranted and the extent to which each evaluation required by law is possible, 3) clarify Congressional information needs, the quality and type of evidence required, and a planning cycle for each major evaluation required by law, and 4) identify the changes in programs or understanding which could occur on the basis of alternative findings. Parallel suggestions were also made to the Department of Education.

The recommendation is, at its simplest, embarrassingly mundane: it asks that the principals meet. And indeed periodic efforts have been made by Congressional and Department staff to assure that the production of reports coincides with authorization cycles and that Congressional needs are understood. But the process has been less orderly, less regular, and less thorough

than it ought to be. This recommendation depends heavily on the fact that a legislative demand for "evaluation" is often ambiguous. The word can imply any activity from journalistic reporting to full-blown long-term field experiments dedicated to estimating the effects of innovation on children. The involvement of multiple interest groups is often necessary. But this complicates matter further since all are unlikely to agree on just what sort of evaluation is warranted. At worst, general legal demands to evaluate that are unaccompanied by serious discussion obscure the fact that the feasibility of particular kinds of evaluation varies enormously and that elaborate evaluation may be unwarranted.

Specification in law

We recommended that in constructing laws for evaluation the Congress: 1) specify exactly which questions ought to be addressed and the audiences to whom results should be addressed, when specification is possible; 2) provide for formal assessment of the evaluability of the relevant program where specification of questions is not possible; 3) provide for statistically valid field testing of proposed evaluation requirements where specification is not possible and in-house assessment insufficient.

Though statutes are frequently explicit about routine reporting requirements; references to evaluation are often ambiguous. The common requirement, for instance, to evaluate "whether the program meets the objectives of the statute" is common but vague. The published Hearings, covering public testimony submitted prior to enactment of a law, are not always informative.

Defining evaluation requirements in terms of the questions that should be addressed is sensible so long as the questions themselves are clear, answering them is feasible, and the answers are likely to be useful. The particular questions that often need to be addressed are: How many are served and how many need service? What are the services and their costs? What are the effects of programs on their primary or secondary clients? What are the costs and benefits of services? The early specification of audiences, especially particular committees or Congressional support agencies, should enhance the usefulness of reports.

We recognized that explicitness in law is often not feasible or desirable. Consequently, we suggested formal investigation of evaluability (Wholey, 1977) to clarify questions, audiences, and the ways in which results could be used, within a year after enactment of a demand for evaluation. We recommended field

tests of reporting requirements in the interest of assuring that costs and benefits of reports, users and uses of reports were well understood.

Authority for technical discussion

The third recommendation in this class urged that the Department authorize the technical staff of evaluation units to initiate discussion of evaluation plans with pertinent Congressional staff, at their discretion, and to refrain from directives which might impede direct discussion.

The impetus for the recommendation was simple: Competent evaluators can expect to do a good job only when they have the opportunity to frequently discuss Congress's information needs. Restrictions on the evaluation unit's initiating discussion with the Congressional staff of committees that demand evaluation prevent the job from being done better. Such restrictions were made formal by, among others, Joseph Califano during his tenure as Secretary of the Department of Health, Education and Welfare. The Report recognized that some restrictions on bureaucratic lobbying for programs are warranted, and that some administrative rules are necessary to keep the process of communication between agencies and the Congress orderly. Restrictions engender a lack of clear opportunity to identify which information Congress can use. This in turn decreases the likelihood that evaluations will be timely, relevant, and credible and the likelihood that the Congress will find the results useful. Relaxing restrictions will not, of course, guarantee usefulness.

Remarks

Three aspects of the Israeli papers are pertinent to these recommendations.

Mapping the Question. The first concerns Louis Guttman's mapping sentence which, as Lewy describes it, is a remarkably terse statement for helping one identify critical decision points in the evaluation: when information is warranted (at what stage of the program), what entity ought to be evaluated, why the information is needed, how it should be obtained. This literal map is implicit in our first two recommendations to the Congress. Moreover, it constitutes a neat working rule for the individuals responsible for planning evaluations at the national level. It takes no wit to see that it can be formally adopted as well in work at state and local levels of government.

Audiences for Results. The mapping sentence dedicates no explicit attention to the matter of whose needs must be served by the evaluator, but both the Holtzman Report and the papers in this volume do so. For instance, one

of our major suggestions was to identify audiences for results as soon as possible. At its worst, this is merely pious exhortation: the turnover of staff members of the Congress and at the executive level is high enough to threaten some short-run projects and most long-term research. But it is a practical suggestion to the extent that career bureaucrats and Congressional committee staff that are responsible for evaluation are a stabilizing influence and can serve as a vehicle for identifying both transient and durable users of information.

The idea of regular meetings among both evaluators and users of information is not different in principle from the tactic already used at the Israel Curriculum Center, judging from Lewy's paper. The ICC's use of a liaison person as a bridge to users and as a expeditor seems sensible for information exchange, building trust and a common vernacular. Note that our recommendations, though, concern only evaluators and users. Production or development agencies are ignored. In principle at least, the ICC liaison approach is adaptable to working relations between these two groups in the U.S. as well at the federal, state, and local levels.

Developing an Evaluation Portfolio. Lewy cites a 1970 article by Alkin, suggesting that the task of evaluators is in no small measure, "ascertaining the decision areas of concern". Lewy expands on this to argue that the evaluation unit adopt this as a fundamental operating principle, and moreover that the selection of the topic for evaluation "should be done on the basis of consent between or at least compromised among the two teams", the two teams being the evaluation unit and the program development team. The Kugelmass discussion of the Reform-Junior High School change instituted by the Israeli Knesset makes a similar point. In 1968, the law altered the then conventional 8 year primary school and 4 year high school program into a new 6-3-3 sequence in government schools. Kugelmass suggests that there was a great deal of difficulty in meetings among the administration of the Ministry of Education and researchers, and stresses that making a decision about what sort of evaluation to do, under differing pressures from diverse interest groups, is not simple or easy. Dan Davis too makes the point. But he emphasizes that once a preference is made explicit about the desirability-outcome evaluation at least, the evaluation must be under considerable control by the evaluator to assure a reasonably successful evaluation.

The problem here is not new, of course. It underlies any attempt to build a coherent research and development agenda, any effort to choose among products for manufacture or corporations for acquisition. Nick Smith's labelling of

the process of trying to choose among evaluation enterprises as portfolio development is also apt. (The label may assist in translating ideas about evaluation to a nontechnical audience, such as a legislature, that is sometimes more sensitive to business than to the evidential basis for government).

Nor is the difficulty of choosing what sort of evaluation to perform confined to Israeli borders. The U.S. encounter with the same problem is one of the reasons for making clear the choice and the basis for choice in the Holtzman Report's recommendations. The pertinent evidence comes from a variety of sources. Mary Kennedy (1980) for instance suggests that efforts to compare the relative effectiveness of two or more strategies of (say) instruction are not common at the school district level. Her message is that impact assessment is less frequent and less important than other evaluation questions in the local agencies. Charles Stalford (1980) quite properly warns against a "testing-only model of evaluation" (p. 6). The Holtzman Project and other work seems to support this contention: activities other than impact estimation are important and the importance varies with the level of government and with the agency within government.

This of course does not mean that comparisons are unimportant, merely that they can receive low priority. The reasons may include simple inability to create variations that are cheaper and more productive than the existing one and that are worth testing. For instance, in response to my suggestion that the Agency for International Development test variations, one AID staffer complained plaintively that they had had enough trouble creating one variation and that creating more just to be able to find the most effective one was too onerous to countenance.

The more general implication is that within a school district or at any other level of government a de facto portfolio of evaluation activities is created. The question this engenders is how such a portfolio can or should be built. Consider for instance the first factor that might be taken into account when developing a portfolio: the source of the inquiry or target audience for evaluation results. Is it sufficient to rely solely on evaluative questions from instructors, parents, or program managers to develop a portfolio? Probably not, since asking the right questions requires some skill and informed conceptions of evidence. Can one rely solely on the evaluator? Probably not, since this assumes too much knowledge of substantive problems. What do the stereotypical portfolios look like in this respect? The U.S. General Accounting Office initiates about two-thirds of its own inquiries and most of these are managerial studies. Should evaluation offices

at the state level build a portfolio in the same way? We know very little about this. Nor can we give much advice.

The second factor is time. We know that fast turnaround studies are essential to satisfy a public or a superior with a short attention span, if not to actually resolve durable problems. And so perhaps most evaluative studies need to be short in the interest of evaluator survival. But all evaluators, especially those in government and academe, do have some responsibility for finding long-term effects of programs, and for understanding long-term social problems as well. There is no technology for designing evaluations which produce short, interim, and long-term results.

The strategy has been to elicit suggestions for evaluations from directors of substantive programs. The ultimate choice is based partly on agreement between evaluators and these agencies in principle. But it may be superseded by agreements with the Secretary of Education or by other criteria used in making decisions. The other criteria include expiration dates for legislation bearing on programs, the period during which a legislative committee could be expected to use information or whether high priority programs have been evaluated. The choices are incorporated into three plans or evaluation.

Very little intellectual attention has been dedicated to a third factor, the administrative mechanism which yields the evaluation portfolio, and which can be used to terminate projects which are not turning out well. The system at the U.S. General Accounting Office appears to be hierarchical, involving screening committees to ultimately determine the choice of project, and a special committee for termination. Nomination of topics to be investigated come from staff groups with operating responsibility. Until recently, the Office of Education and Dissemination at the Office of Education had a similar system. But it is not clear how termination decisions were made.

The main point is that criteria for developing and assessing the value of an evaluation portfolio are not yet clear. Serious attention has been given the matter by bureaucrats, not by executives or academics, and their effort can be augmented profitably by others.

4. EVALUATOR CAPABILITIES

The Project staff had been asked to investigate the capabilities of those who do evaluations and to make recommendations based on our findings. We recommended that the Congress and the Department: 1) assess capabilities of local and state education staff before new statutory evaluation requirements

are directed at them in order to determine where resources are adequate to meet the demand, 2) expand training or technical assistance when the demands are notable and capabilities low, and 3) explore the feasibility and desirability of direct contract programs to capitalize on capabilities in strong local and state education agencies.

The first section of the recommendation stems partly from the fact that no real standard for assigning the title "evaluator" exists and that skills required of the evaluator depend heavily on the nature of the evaluation demand and on local and state interest in evaluation. The second part is based on the finding that most local and state agencies need assistance when the evaluation requirements are technical. The minority of these agencies that do have strong evaluation units are a major resource, and we believe that direct grant opportunities should be expanded to capitalize on them.

By determining capabilities here we mean understanding whether there can be a reasonable match between what the law demands of local and state evaluators and the skills of these individuals. A formal assessment of this sort is unlikely to be easy for three reasons. First, within a school district or state office, evaluation responsibility may be split up among several individuals, none of whom may have any pertinent training, and this responsibility can often change. Second, evaluation duties may have a considerable range depending on local interest in exploiting systematic information to improve programs. Just meeting minimum federal requirements requires far different resources than establishing a long-term research program. Finally, the methods one might exploit to perform capabilities assessments are not clear. They range from intensive task analyses during, say, pilot tests of new regulations that require a specific type of evaluation to telephone surveys that enumerate skills and tasks. Drs. Georgine Pion and David Cordray are developing plans now to accommodate such problems and to implement assessments for local education agencies and for community mental health centers.

A critical influence on the matter is whether an education agency decides to just accommodate federal requirements or goes beyond these to mount a stronger evaluation program. Even just meeting requirements demands some skill. The notion of temporal instability or reliability of tests is not obvious to many people despite training in a substantive education area and in the history of testing. As a consequence, we urged that the federal program sponsor make an effort to capitalize on evaluation unit expertise in training local or state staff in meeting demands, and that funds be allocated

for such training. Options such as cooperative arrangements among small education agencies for joint support of an evaluation unit and expansion of federally supported technical assistance centers need to be explored, and the Report suggests doing so.

For local and state agencies that are willing to go beyond federal requirements, we stressed direct grants from the federal government for two purposes. First, some agencies are capable of mounting research and evaluation programs that match federal efforts in quality and are more pertinent to local interests. They are in the minority, accounting for probably no more than 400 of the 15,000 school districts and less than half the state agencies, and they deserve to be given an opportunity to produce good work that can be applied to other areas. The second purpose is to foster closer ties between local agencies and university evaluation groups. Such arrangements are bound to be difficult, but it is hard to see how the state of the art in evaluation can be advanced without better ties between the two.

Remarks

There are several points of correspondence between the findings on which these recommendations are based and the opinions registered in the seminar papers. Consider, for example, Lewy's conclusion that "a realistic assessment of actual need in terms of manpower and other resources and their satisfactory provision constitute a prerequisite for the successful operation of the evaluation unit. It is not the absolute size of the budget which determines the successful operation... but rather the match between the resources available and appropriate definition of the evaluation tasks". This is remarkably similar to the conclusions tendered by Pion and others in the Holtzman Report that evaluation tasks vary widely among school districts, that the skills required to do those tasks vary as well, and that the tasks have to be understood before resources can be intelligently allocated to training and before laws demanding wholesale evaluation can be conscientiously constructed.

The second point of correspondence lies in Kugelmass' observation that as a result of the academic emphasis on basic research and theory, the manpower available for task-oriented research such as evaluation is spare. Tamir points out that even where university-trained researchers are available, there will be a notable tension between the research-oriented view of what should be done at what level of accuracy, and what the manager or practitioner believes

is warranted. That same problem has appeared in the United States and is being resolved in several ways. A tight market in university jobs seems to have resulted in better people going into government, into independent contracting research institutes, and into evaluation units at the local and state levels. The migration engenders problems but it is also reasonable to expect better understanding of local practical problems and wiser evaluators. The federal government has assisted by creating technical assistance centers to respond solely to local needs for advice. The centers are staffed by university trained people; not all local programs, however, are assisted by such centers.

A third point of correspondence is Lewy's conclusion that the evaluation unit serves as a "catalyzer for initiating evaluation activities, the limits of which exceed the working capacity of the unit itself". This suggests that developers get interested in evaluation to the extent that the working relationship with evaluators is close and that this interest can be used to expand the effective size of the evaluation unit's staff. There is an analog here to efforts at the local and state levels in the U.S. to augment the evaluation requirements set out by the federal government. In particular, though only a minority of the evaluation units and research units within school districts are strong, this minority used the minimal requirements as a vehicle for collecting additional information of more direct relevance to local interests. States such as California and Massachusetts also have this utilitarian perspective, building on federal investments and requirements to do a better job in meeting federal, state, and local demands.

One feature of the Israeli experience for which there is no routine analog is the use of a liaison person to link the program development group with an evaluation group. Informal arrangements of the sort do appear in the U.S., but the role seems much better articulated in the ICC. Lewy's description of the liaison person's training and skill is especially interesting. If I understand it correctly, the three types include those with evaluation training of a substantial sort, those with substantial substantive training, and the project director. Lewy suggests that the project director does not work out too well because he hasn't got the time, and that the substantive area expert is probably best because he is immersed in the project itself.

5. DESIGN AND EXECUTION OF OUTCOME EVALUATIONS

Once said, it is obvious that quality in design of an outcome evaluation affects quality of the data and of conclusions. The evidence that bad design

can make programs look worse than they are, or better than they are, or yield ambiguous evidence is substantial. The theory that organizes understanding of biases in estimating program effects is, however, reasonably well articulated.

The idea that quality in design ought to be recognized as a formal part of evaluation policy is explicit in federal education agency attempts to yoke the introduction of new programs with design, as in evaluation of the programs supported under the Emergency School Assistance Act, in attempts to review designs with more vigor within agencies, and in efforts to provide technical assistance programs in the interest of better local design. It appears also in the U.S. General Accounting Office's attention to competing explanations characteristic of poor designs, to the elements of reasonable design, and to the need for designing evaluations before a new program is put into the field. It has been recognized by the courts in cases outside education which recognize the flaws in some evaluation designs and the benefits of others. The Supreme Court's Federal Judicial Center, for example, is developing policy on the use of randomized field experiments in legal settings to make clear the issues and precedents. The theme of quality, though, is not sufficiently well established to flourish without periodic reiteration. The task was undertaken in the Holtzman Report through two recommendations, one made to the Congress and one to the Department of Education. A third, concerning standards, is treated later.

Pilot tests and designs

We recommended that the Congress: 1) routinely consider pilot testing every major new program, major variations on existing programs, and major program components before they are adopted at the national level, using high quality evaluation designs, and 2) authorize the Secretary explicitly, in each statute that requires estimates of the program's effects on target individuals, to use high quality designs, especially randomized field experiments, for planning and evaluating new program components, program variations, and new programs.

The rationale for the first part of this recommendation is that higher quality evaluations are more feasible before the program is adopted at the national level. Political-institutional constraints are likely to be less severe, better designs can be employed, and conclusions then are likely to be less ambiguous. The introduction of new programs can be staged so that earlier stages are pilot tests for later ones. We stress formal tests of

new program components and new variations here because such evaluations are not a matter of common practice.

The second part of the recommendation, as well as the first, stems from our conclusion that better designs must be used if the Congress or the Department wants good estimates of the effects of programs on children. We do not advocate estimating those effects in all cases. Estimation is complicated under the best of conditions, despite simplistic announcements that the "program was successful because test scores went up" or that it was unsuccessful because they went down. Nor do we believe that designs that are high quality relative to statistical standards are always feasible or warranted for estimating program effects. We do advocate explicit authority in statutes for high quality designs, especially randomized experiments, to facilitate their use. We believe explicit statutory provision is essential because such designs are the best in principle, and that should be recognized. The authorization should provide for review of the use of these designs.

Tests of new program components, program variations, and new programs

We recommended that the Department of Education explicitly authorize the use of high quality evaluation designs, especially randomized experiments, in evaluating new program components, new program variations, and new programs, in all regulations that require estimating the effects of innovative changes.

The main justification is that high quality designs lead to less debatable estimates of programs on children than do low quality designs. They are less difficult to execute and are more feasible for pilot testing new programs, program variations, and program components, than for estimating the effects of ongoing programs. Explicit authorization would make the importance of good designs plain, and would provide a more clear opportunity for competent state education authorities (SEAs) and local education authorities (LEAs) to exploit them. We use the word "authorize" here rather than "require" to make clear that the evaluator is empowered to use an experimental design but need not do so if it is unwarranted or not feasible.

Remarks

The recommendations on randomized field tests were supported by some evidence on their feasibility and appropriateness. A judgement about feasibility in the particular case, we believe, should be based on precedent, for a number of field experiments have been undertaken in education and other

areas. It should be based, for complex evaluations, on pilot tests of the experimental procedure itself, for one cannot anticipate all problems engendered by the method. And it should be based on independent criteria such as whether the service is in short supply and randomized assignment is indeed an equitable method of allocating it. These criteria need to be explicated better, and they need to be linked to broader testing strategy.

Standardized Evaluation. In doing both, we can rely partly on Davis' presentation. He proposes that six conditions must be met in order to obtain decent estimates of the effects of programs. 1) The program has to have relatively clear goals and operating procedures, that is, it must be implementable. 2) The evaluator must be responsible for both the program operation and its evaluation, maintaining special control over evaluation. 3) The program must be implemented first in an optimal setting - field conditions, training, and the like being the best possible. 4) Schools must be selected for their willingness to participate in the research. 5) The research design must approximate laboratory models in terms of assignment and execution of evaluation. 6) The results of the evaluation must serve as a standard against which normal field operations can be judged.

Not content to just lay out conditions, Davis is attempting field trials under these conditions on a program that has never been investigated well, despite its attractiveness, in the U.S. or elsewhere. The program is, as I understand, a national tutoring effort in which university students get academic credit for helping children in grades five through nine. The regular program has minimal supervision and the more elaborate "optimal" version involves intensive supervision and more hours of tutoring. The more intensive version is distinctive in that "the program is a supervisory and guidance structure which is sensitive to the problems encountered by the tutors and can help in solving them, and the program is maximally flexible so that it can adapt to the specific conditions of each tutor-tutee relationship". Earlier evaluations of the ongoing program show mixed results. Consequently, a good field test of the ideal version of the program is a natural way to understand what the maxiumum effects of tutoring can be.

Apart from the conditions that Davis proposes, his general strategy of conducting a controlled experiment of an optimal program to gauge the maximum effects of subsequent or ongoing programs is an attractive one. It is generalizable to the U.S., at least in principle, and does not appear to have been suggested before, at least not as explicitly. There have been related suggestions however. For instance, the Riecken et al. (1974) volume on

social experimentation stressed the idea that in field tests of programs, one ought to assess not only program levels that are clearly feasible, but also some that are not practical at the national level. The argument is based on the premise that "practical" programs are often weaker than we expect them to be and that high dosage (optimal) programs, though impractical at one time, may be practical in the future. This is especially likely if one finds that the higher intensity does produce notable effects while low dosage "practical" programs have no detectable effect at all.

Testing Components and Variations. No theory of evaluation demands that the effects of an entire program be estimated, and few practitioners would regard such an unqualified demand as sensible. Yet professional vernacular, rhetoric, and legal mandates foster the view that wholesale evaluation is warranted, distracting attention from the possibility of testing components of programs. For example, one may find that running high quality tests of new parent education programs is not possible. But estimating the effect of alternative sources of information, of various ways to present the information or ways to prevent ingenuous use of information, and so on, may be possible in small high quality experiments. The strategy of component-wise evaluation has been exploited in the U.S. evaluations of the Emergency School Assistance Act, in research which preceded the development of Sesame Street, and elsewhere. Incorporated into evaluation policy, the idea broadens early options, and in the event of a major evaluation's failure, it is a device for assuring that at least pieces of the program can be assayed properly.

Analogs to this approach are not difficult to find in work at the Israel Curriculum Center, the High School Biology Project, and other projects. The process of identifying which components to evaluate seems to have been routinized best at the ICC, notably by exploiting the mapping sentence approach: Lewy's application suggests that focusing on the entity to be evaluated is integral to the continuous evaluation strategy. The Israeli adoption of high school biology curriculum (BSCS) programs developed in the U.S. is pertinent too. According to Tamir, the BSCS program was available in three versions. The blue version emphasized bio-chemical concepts and is relatively sophisticated. The green emphasized an ecological perspective. The yellow stressed a more conventional approach, but was somewhat more interesting and adaptable to the Israeli perspective. The important point here is that three variations of the program were developed. This seems an imminently sensible idea when there are major differences in perspective

about how something should be taught and major differences in teacher opinion about how it can be taught. Moreover, the variations' relative effectiveness are testable in principle. In practice, pieces of each are likely to be evaluatable.

Dan Davis's "standardized evaluation" is also consistent with this theme. To determine the maximum possible effect of an ongoing program which itself may be difficult to evaluate well, one may invent a very intensive variation on it, an optimal version, and submit this to very well-controlled tests. The idea can help to circumvent the chronic problems of estimating the effects of ongoing studies.

Finally, the plan being developed by Gershon Ben-Shakhar and Baruch Nevo for understanding the effectiveness of matriculation tests reflects some of the same spirit. Formal testing in a part of a large, complex education system and its evaluation, more or less independently of the rest of the system, is an idea worth exploring. The so-called Irish Study has had a distinctive advantage in this regard, since standardized testing is not common in Ireland and evaluators could introduce it on a trial basis and estimate its effects on teachers and students using randomized experiments and other evaluation designs (Airasian, et al., 1978).

6. CRITIQUE AND SECONDARY ANALYSIS OF EVALUATION RESULTS

The evaluation design, its execution, and the skills of original investigators are basic to the production of useful information. But they are not always sufficient. The pressure toward less than candid reporting is sometimes great, and it is not always clear that one can resist them. Egregious errors are made and corroborative or contrary evidence is ignored, for the time available for analysis is not always adequate. The benign skepticism necessary for the in-house evaluation generates a reasonable but parochial picture. A less benign, or at least more impartial, outside analyst could come up with different conclusions. Finally, data generated in social program evaluations constitutes a national resource and should be treated as such. The research can be expensive despite the production of data that are useful in short-term decisions. It behooves the evaluator to learn how to exploit the information repeatedly.

Partly for these reasons, we recommended that in statutory requirements for evaluation of major programs, the Congress: 1) require an independent, balanced, and competent critique of evaluation results that are material

to policy decisions, 2) require critique of samples of evaluations submitted by LEAs and SEAs in response to legal requirements, and 3) require that statistical data produced by national evaluations be made available for reanalysis.

A complementary recommendation was made to the Department of Education: 1) Incorporate into procurement procedures and policy the requirement that all statistical data produced in major program evaluations be documented and stored for reanalysis. 2) Create an administrative mechanism for deciding when simultaneous analysis by both the original evaluator and an independent analyst is desirable and feasible, and a mechanism for executing simultaneous independent analyses.

The text of the Report made it plain that we did not mean adverse commentary in using the word "critique". The idea is to ask for reasoned judgements about whether conclusions drawn from the evaluation are sensible and can inform decisions. The immediate reason for the recommendation is that such criticism is not routine but is essential to enhance the credibility of good evaluations, to properly identify poor evaluations as such and to provide feedback to federal evaluation units, contractors, and grantees about the quality of their work. There is no formal system for the competent critique of evaluation reports produced by local and state education agencies in response to federal law, yet many such reports could benefit from conscientious review.

The elements of a system for critique and secondary analysis should include: 1) an explicit institutional policy on the rapid disclosure of reports and access to the data underlying the reports, 2) a mechanism for independent critique or secondary analysis where possible during an evaluation, and where this is not possible, after a report is submitted formally, and 3) guidelines on the reporting and storage of information.

The Report recognizes the problem that criticism may be witless and counterproductive. The recommendation is based on the premise that the long-run benefits will offset the effects of self interested criticism and the burden that criticism imposes on the evaluator.

Remarks

No explicit attention to this matter is evident in the seminar papers. Rather, the theme is implicit in the spirited exchanges of opinion during the seminar, and in custom if I judge correctly the participants' stress

on presenting research results in the sometimes harsh climate of professional forums. The High School Biology Project, for instance, appears to have generated a large number of fascinating papers that are informative to researchers inside and outside Israel. As I understand it, the Office of the Chief Scientist too can serve as a device for independent critique, and perhaps secondary analysis as well. But the ordinary mission of the office, in advising and deciding on research projects, can dilute administrative independence. Similar offices in the United States, such as the Director of the research-oriented National Institute of Education, operate with a similar constraint: fiscal, administrative, and bureaucratic independence is a matter of degree.

Reasonable critique depends on standards, and standards were addressed in both the Holtzman Report and in the seminar papers. They are considered in the next section.

Analysis and Competing Models. The Holtzman Report did not examine the methods of data analysis used in evaluation. Its principle audience was nontechnical and most of the questions it addressed were answered in a nontechnical way. Other research produced by North Western, however, has examined technical issues in assuring access to and the quality of data for analysis and the nature of competing analyses that might be undertaken. The volume edited by Boruch, Wortman, and Cordray (1981), for instance, considers both policy and practice, and can be regarded as an explication of the Holtzman Report's recommendations on secondary analysis.

Itai Zak's paper is most pertinent to this level of detail. It represents a statistical tradition of trying to understand the structure underlying data, of establishing the extent to which a theory represented mathematically is consonant with the information. The tradition is represented most visibly in econometrics, but recent attempts by Goldberger, Jöreskog and others to link approaches in psychometrics and econometrics through structural equation models, and the quantitative sociologists' work on the latter have led to some remarkable advances in facilitating and understanding their use. Several points bearing on the general topic seem worth making. They are implicit in Zak's approach.

First, methods such as Wold's PLS and Jöreskog's maximum likelihood approaches permit one to relax the assumptions characteristic of conventional textbook approaches such as regression analysis. Simply put, they allow one to build more realistic models of reality. That this advantage is not trivial is apparent from policy-relevant research in the U.S. For example, the

Westinghouse-Ohio State University evaluations of Head Start, a preschool program for deprived children, resulted in estimates of program effect that were near zero and in some cases negative. At worst, it seemed that the program hurt rather than helped in their conventional covariance analyses. Secondary analyses of the same data, by Magidson (1977), capitalized on structural equation models similar in character to those that Zak uses. The new results, based on less demanding assumptions than the original work, suggest that the program had positive though small effects on the cognitive ability of children who participated.

The second point is that this benefit of new methods also produces ambiguity. Disparate models, theories of behavior if you will, may fit the data equally well. So, for instance, the Magidson results are being debated by other analysts who believe their models are at least as appropriate. This ambiguity is typical. And consequently it behooves the analyst to fit several models to his data, in much the same way that Zak has done.

Another point worth recognizing stems from the fact that these new model-fitting approaches have developed independently of methods in conventional randomized experiments. The two cultures here are different, but failing to recognize their linkages would be a mistake. It is possible, for instance, to express the ordinary analysis of the variance model that underlies randomized tests in terms of structural models. Lee Wolins and I have done so for one class of models but not much other work seems to have been done. Perhaps more important, the structural models lend themselves to internal analyses in experiments. That is, having discovered that a program has an overall effect, based on randomized design and conventional analysis, one may exploit the new methods in path analysis to better understand links between specific components of the program, specific types of participants, and specific outcome variables. Something of the sort has been tried in an analysis of prison parole programs by Rossi, Berk, and Lenihan (1980). More distantly related approaches are not uncommon in analyzing the results of field experiments on income subsidy programs for the poor (see Boruch, Cordray, and Wortman (1981) for other illustrations).

7: STANDARDS AND GUIDELINES

One of the justifications for the Holtzman Project was Congressional interest in whether evaluations could be subjected to uniform standards for judging their quality. Indeed, a variety of guidelines to judge quality have been developed by the U.S. General Accounting Office, the Evaluation

Research Society, and the Joint Committee on Standards for Program Evaluation, a group whose members include representatives of most professional associations with an interest in evaluation. Crude standards are also embodied in certain federal activity, notably the Joint Dissemination and Review Panel, whose mission is to assess the evidence on locally developed programs in order to determine whether the programs merit federal support for distribution to other local agencies. There is substantial overlap in topical coverage among guidelines. But they differ in detail. Our review led to the following recommendation: while recently developed standards and guidelines for evaluation should not be incorporated into law, they are sufficiently well developed to recommend that the Congress: 1) use such guidelines to understand what can reasonably be expected of evaluations, 2) direct that agencies use them as a guide where appropriate to developing criteria for judging evaluation plans submitted by local and state agencies, and 3) elicit assistance in the interpretation of guidelines from Congressional support agencies, such as GAO, that have been instrumental in their construction.

The main reason for recognizing that guidelines be recognized officially is that we believe they can be useful in explaining what is meant by evaluation to the public and its representatives, and in informing the public about what can reasonably be expected of evaluation projects. Guidelines may also assist in protecting the competent evaluator from incompetent criticism. They should certainly help one to identify inept evaluations.

We argued against incorporating such standards into law because neither evaluation law nor the standards are sufficiently well developed as yet to justify incorporation. Moreover, giving legal status to specific guidelines can impede the development of better guidelines, are almost certain to be applied inflexibly, and are likely to do more damage than good in other respects.

Remarks

Most standards are general and their proper application depends on circumstance. A survey of needs, for instance, requires a subset of criteria in a complete list; a randomized experiment requires a few of the same criteria but others must be employed as well. The Holtzman Report did not address this matter in detail. Instead, it focused on when guidelines should be used, notably in judging the merit of grant proposals and contracts

that contain evaluation plans and, once the program is evaluated, in judging the quality of the evaluative evidence.

The ICC evidently takes a somewhat different approach in ascribing the idea to "minimal evaluation requirements". As I understand Lewy's remarks, the concern is not only on when evidence becomes material but on the general kind of evidence as well. Expert judgement is regarded as essential for judging the quality of materials, observation of the teaching and learning processes are essential during early try-outs of the material in classes, and assessing cognitive achievements of children is essential at the end of the first try-out. The first two activities are the responsibility of the program developers and the last is undertaken primarily by the evaluator.

The ICC criteria are not incompatible with other, more elaborate guidelines. They can be regarded as a distinctive operationalization of items that appear in more general lists and, partly because of their brevity, are likely to be a useful operating rule in at least some large local education agencies in the U.S. And there are distinctive parallels to this minimalist approach in some U.S. evaluation agencies. The Officer of the Inspector General in the Department of Health and Human Resources, for example, undertakes fast turnaround studies that rely heavily on expert bureaucratic judgement and some on-site observation of processes in, for example, health services delivery (Hendricks 1981). But his effort is dedicated to administrative failures rather than to programs in general. There is some similarity also to recent Agency for International Development efforts to execute fast turnaround studies of foreign assistance projects. Here, as in the Israeli case, there is often little time to do much more than obtain expert judgement and crude observations.

Implicit in both the Holtzman Report and in Lewy's standards paper is the idea that guidelines for judging quality can range considerably, from the very permissive to the very general. The clearest illustration of both the idea and its exploitation comes from the engineering sciences. Here, the demand level of the standard depends on the uses to which the information is put. Relatively wide tolerances are permissible in local geophysical measurements since the information is used primarily by lawyers and construction engineers. Much closer tolerances are required in geophysical measures made for some scientific studies, on land erosion for instance. The same spirit is evident in the ICC's minimalist approach - expert judgement normally being less precise and certainly less verifiable than elaborate

observation. It is evident too but underexploited in the U.S. It is sensible to consider using one set of standards in special grants made for small innovative programs and a less restrictive set for programs that operate with regular budgets. It may be sensible to require larger, more capable school districts to provide information that accords with more rigorous standards and to require much less precise information from the less capable ones. The implications have not been worked out nor have the options been articulated well in evaluation management in the U.S. or abroad. Both tasks should be undertaken.

8. THE USE OF EVALUATION RESULTS

Whether the results of evaluation are used and by whom they are used was a fundamental concern of the Project. We uncovered a great deal of information on use, but the research was difficult. Not the least of these was lexical confusion. A federal director of research, for instance, announced that he did not perform evaluation at all despite a list of projects under his direction that included evaluations labelled as such. His superior, interviewed fifteen minutes later, claimed the contrary: that the division produced a great many evaluations all useful to management. We encountered a Congressional staffer who announced baldly at the beginning of an interview that his committee did not use evaluations, yet he later said that evaluation reports were used to guide committee hearings on programs.

This confusion, or at least inconsistency, underlay a good deal of debate about the utility of evaluation results. And so we defined "use" explicitly to mean: 1) applying these results in making specific decisions about law, regulations, budgets, or related administrative topics, and changes in substantive content of programs, 2) capitalizing on them to enhance understanding of issues even where a decision could not be made, or 3) exploiting the information to persuade others, as in political speeches, or to confirm one's own beliefs. The Project's efforts to document use and nonuse of evaluation reports focused on specific evaluations, and stressed the corroboration of evidence from different sources. The findings suggested, as one might expect, that some evaluations are used and some are not, and that use depends heavily on the planning of use, close relations between the user and evaluator, and willingness and capacity to use results.

Our first recommendation bearing on use of evaluation results was made to the Congress. We urged that its members:

1) direct the staff of relevant committees, the Department, and the GAO to routinely outline which institutions can reasonably be expected to use results of each major evaluation and how such results might be used, during the design stage of every major program evaluation;

2) specify exactly which evaluations have been used and why they were used, which have not been used and why they were not used, in authorizations and appropriations committee reports;

3) require evidence about specific changes resulting from evaluation, whenever the law requires state agencies to describe uses of evaluation; and

4) explore the feasibility of direct competitive grants and contract programs focused on improving the use of results at the local and state education agency levels.

The origins of the first part of the recommendation lie in the absence of any mechanism for planning use at the national level. Simply put, unless specific user groups are identified and some decision options laid out, evaluation results are less likely to be used. Indeed, if there is no clear way to link the evaluation with decisions or considerably better understanding, one can argue that the evaluation should not be performed at all. Specifying users and options will also help to make it easier to track utilization, and that, in turn, will help to inform judgements about how evaluation resources could be better allocated. The recommendation to cite useful and useless evaluations in federal reports and to require SEAs and LEAs to record specific changes has the same objectives: better understanding of use and better resource allocation. The suggestion to identify useless evaluation is not an invitation to criticize arbitrarily. We found that some local and state education agencies are capable and interested in inventing and testing better ways to use information. The suggestion to expand their opportunities for doing so is based on this.

A second, related recommendation was made to the Department of Education. The Report urged that evaluation unit staff and evaluation contractors be directed to 1) provide oral reports regularly as well as written reports on results of major evaluations, and on the uses to which results can be put, to relevant Congressional staff and support agency staff and the program staff within the Department; 2) create a system to periodically collect, synthesize, and report specific uses to which evaluation is put; 3) improve the Annual Evaluation Report by citing instances of use more specifically;

and 4) direct evaluation staff to meet regularly with Congressional staff to clarify information needs, feasibility of evaluation, audiences for results, and ways in which results can be used to modify programs.

This cluster of suggestions is based partly on the finding that the use of evaluation results is not tracked conscientiously and the belief that it ought to be tracked to learn how to perform evaluations better, and how to better allocate evaluation resources. The rationale for the last recommendation is identical to the one given earlier for the Congress on planning and executing evaluations.

The final suggestion under this topic focuses attention on assuring access to and better specification of reports.

We recommended that the Department 1) adhere to a clearance rule which makes evaluation reports automatically available after a fixed number of weeks; 2) specify completely the evaluation documents referred to in the Department's Annual Evaluation Report, the Federal Register, and policy statements; and 3) include, in every major evaluation report, a list of core recipients.

The recommendation stems partly from difficulties encountered in obtaining reports under review by the Executive Secretariat of the Department of Health, Education, and Welfare and other groups involved in the DHEW clearance process. We also found it difficult to identify reports precisely when they were cited as evidence of the usefulness of evaluation in developing regulations or policy. The absence of a list of core recipients of reports made it very difficult to identify potential user groups and to determine if reports were used. The consequence is that what is useless or useful is less verifiable.

Remarks

The basic idea that there needs to be a group constituted to reason from the data is implicit in both our recommendations and in Israeli operations. The form differs a bit, though.

Committees to Reason from Data. Consider, for instance, Kugelmass' description of the Van Leer study of primary schools, a massive undertaking to understand the process and products of primary schools in Israel, including controversial issues such as integration and religious vs. non-religious school systems. One distinctive aspect of this enterprise was that the Ministry of Education and Culture and the Chief Scientist took pains to

appoint a high level committee to understand the ways in which the results of this research could be used by the Ministry. The Chief Scientist, as well as heads of various divisions of the Ministry and of the primary school system, were involved in the committee, and a number of sub-committees were set up to recognize special interests and problems. A mechanism of this sort was not suggested in the Holtzman Report. But it does seem to be a sensible way to understand how policy implications can be educed from such data. The Holtzman Report assumed that negotiation and regular communications between the Congress and the Department would facilitate in understanding the uses to which the data could be put. The power that a blue ribbon committee has to do this, or to exploit suggestions made at lower levels could be used to make negotiation more effective.

Decision Options. The Holtzman Report marshalled a good deal of evidence on the use of evaluations in decisions at the federal level. But we obtained very little for local levels because of the difficulty of corroborating use. Nor did we classify the uses according to what decisions they might concern. Partly for this reason, the classification schemes developed by Lewy and Davis are pertinent. They are tidy ways to specify decision options. Moreover, we can exploit them to explain our own recommendation about making decision options clear before evaluation is undertaken.

For example, Lewy identifies three decisions that might stem from an evaluation done by an evaluation unit. The first is the selection of program components: what should be taught, what materials might be included in the teaching, and so on. He is careful to point out that the program developer is ultimately the one who must choose from among several alternatives examined in evaluation. The second decision is modifying a program. He suggests that it "may turn out that some element such as exercises, illustrations, or explanations contains certain flaws". It is up to the evaluator to call attention to these. The third decision option has to do with qualifying the use of the program. Here he stresses the "optimal" conditions under which the program might work or the minimal conditions for usage. This includes, for example, whether the program will work with little or no training of teachers rather than with a good deal of training, whether equipment, space, and the like are available, and so on.

This point is important for U.S. evaluators. The Holtzman Report pointed out the inability or unwillingness of various audiences to specify decision options before an evaluation is actually undertaken. Yet here Lewy implies that the process is almost a matter of course for the Israelis. More impor-

tant, Lewy recognizes some decisions that are independent of budget and therefore less controversial.

Davis is careful to explain decisions that could emerge on the basis of his special standardized evaluation plan and to urge that these decision options be recognized before evaluation is undertaken. His options depend on the results of the evaluation. For instance, the first possible result is that the standardized evaluation yields an estimate of no effect or of negligible effects for the optimal version of the program. The implications for decisions are that one can question the "efficacy of the program in any form" and may recommend either dropping it or revising it. The second possible result is that the standardized evaluation results in a large program effect and all subsequent field evaluations show negligible program results. It is at this point that the program director, according to Davis, has to think about alternative ways of improving the field version of the program. The third case is that evaluation of the optimal version yields a large program effect and "one of the generalized evaluations results in a moderate effect". Here Davis suggests that the difference is a matter of consideration and cost benefit analysis. The final outcome of interest is the standardized and generalized evaluation result in approximately the same moderate or large program effects. It is at this point that one goes to the next step and tries to understand how the cost of programs can be reduced.

Enumerating Uses of Information. Tamir's strategy of briefly identifying the specific findings of evaluation and their eventual uses is similar in spirit, although not in detail, to the way the U.S. Department of Education Office of Evaluation's annual report is composed. The uses he identifies are interesting.

For example, he cites a 1974 paper showing low student interest in the study of botany. The subsequent action was integrating the study of topics common to both botany and zoology. He cites a finding that students had trouble applying statistics in biology with the consequent action of preparing curriculum materials on the use of statistics in biology. Another finding was that students in agriculture schools achieved poorly on this version of the material, probably because it had no applications to agriculture. Action was taken to prepare special modules to "suit the needs and interest of agricultural students". Similarly, it was found that culturally deprived students whose families come from developing countries

achieved poorly and the consequent action was preparing a Hebrew version of the material of the test designed for the original course and designing supplementary modules as well as in-service teacher training.

Such items are very persuasive. More importantly, a listing of the sort provided by Tamir might serve as a good model of the sort of information which should be recognized in the university training of evaluators. It is the kind of information that can be provided to program development staff to guide them in making decisions about how to "reason from the data" and to senior executives and perhaps legislative staff to show how evaluation results have been used. And the model is likely to be useful at the local level to illustrate use at least and to foster invention of other approaches at best.

Definition of Use. None of the seminar papers put much stress on defining the use of evaluation results. But several recognize difficulties engendered by ambiguity in the word and a variety of definitions are implicit. Blass, for instance, recognizes, as the Holtzman Report does, that arguments about utilization may be gratuitous, citing the "usual almost ritual litany about the underutilization of research results". The implicit definitions range from examination of a report, judging from the Kugelmass paper: "The very process of bringing... senior decision makers (through the Office of Chief Scientist) into continued examination of the research... may be the most important product of the process of evaluations, and not necessarily specific decisions". The Holtzman Report recognizes the same kind of use in defining evaluation and exploits it in enumerating references made to evaluations in Hearings routinely issued by Congressional Committees charged with authorizing funds for education programs.

Tamir's listing is much more specific and implies a different category of uses, also identified in the Holtzman Report: specific decisions. The category is important but was rather difficult to examine. To be sure, some evaluations, such as the National Institute of Education's Compensatory Education Study are remarkable in that evidence on use is available from the public record, such as Hearings, and can be corroborated through interviews with legislative staffers and bureaucrats. Moreover, it was relatively easy to trace ties between items in the Study and subsequent changes in federal regulation and administrative practice. Others are not as easy. Overreporting of use is likely, for instance, if one talks only to the producers of a study. Underreporting is chronic partly because of memory lapse, the absence of written accessible records on use, and the time it may

take for a report to filter through several layers of bureaucracy. The centralization of Tamir's project may make tracking utilization easier than in a decentralized enterprise.

Factors Influencing Utilization. For Blass, the most important factors influencing use are the political situation, the organizational structure of the society, personal attributes of the evaluator or of the decision maker, especially of the latter, the state of the art in the scientific discipline, and the character of the issues.

The Holtzman Report did not frame the influences on evaluation this way. But the conceptualization seems natural for Israeli operations and for some U.S. activities. My understanding is that evaluation is more centralized in Israel, if the organization of the ICC, the High School Biology Project, and the Office of the Chief Scientist are any index. This stands in contrast to many U.S. efforts in that evaluation responsibilities are dispersed across levels of government - local, state, and federal - and agencies within levels of government. Judging from the Kugelmass paper, the Office of the Chief Scientist can relate directly to the political concerns of the Ministry of Education. In the U.S., the several layers of bureaucracy between the Secretary of Education and the Evaluation unit within the Department, for good or ill, probably affect the political attentiveness of the latter and the receptivity of the former.

Centralization does appear at times in special studies undertaken in the U.S. For example, the NIE Compensatory Education Study was created by law to examine, among other things, how federal funded programs for poor primary school children fared. It was centralized in that a team approved by Congressional staff was created to be answerable to the Congress alone. It was deliberately sensitive to political issues and accommodated them through continuous negotiation between the evaluation groups and Congressional interest groups. The personal attributes of the group leader, Paul Hill, including his prior experience as a bureaucrat and Congressional staff member, seem to have been very influential in producing useful results.

There are, of course, ways to characterize factors that influence the use of evaluation other than the one that Blass proposed. In some cases, they may be more productive. Consider, for example, that each of the seminar papers stresses the richness and the diversity of the process by which evaluation results are used or not used. They are important in this respect but, for the moment, let me propose a model which may help to understand better the order that underlies the diversity.

The following questions imply distinct events that determine whether evaluation results are used:

. Does the prospective user know about the evaluation results?
. If results are known, are they understood?
. If understood, are they believed?
. If believed, does the user have the ability and willingness to use them?

When all are answered in the affirmative, then results are likely to be used. But the first negative will reduce any possibility of use, especially if the potential user is the same at each step.

This way of describing the utilization process is useful by itself. For example, it suggests that simple probabilistic models may be helpful in understanding why some research on utilization is misleading, and how one might enhance utilization. The simplest such model posits that each event is independent and a probability is attached to each question's resulting in a Yes. If the probability for each is only 1/2, say, then the overall probability of a "use" occurring is $(1/2)^4 = 1/16$. If, as I suspect, the odds are lower on each Yes, say 1/4, then the probability of a user's knowing about results is 1/4, the probability of the user's understanding them is 1/4, and so on; the overall odds against results being used are 255 to 1. Not very promising.

Other models though are more realistic. If, for example, we suppose that the evaluation process is centralized so that the prospective user is also the producer of the information, then the probability of being willing to use the information is conditional on the prior events and the likelihood of use is closer to 100%. Indeed, the conditional model is a numerical representation of what happens when a liaison person is used in ICC evaluations in Israel and when a brokerage system is used to translate findings into usable results as Cooley (1980) did in the Pittsburgh school district.

The model does suggest that we collect data at each stage, to properly estimate odds on ultimate use of evaluation results. But little work appears to have been done on this problem in the U.S. Most research deals with the question: What does the process of use look like in particular case studies? It seems to me that the case studies are important to inform such models, as well as being important as qualitative descriptors of the process. But they require large samples of events across sites or of multiple events within site to obtain decent estimates of parameters in the model.

Notes

[1] The research on which this paper is based was supported by the Division of Evaluation of the U.S. Department of Education and the National Institute of Education. The discussion of Israeli work here is based on material presented at the Israel-US seminars on evaluation in June 1980. The material on U.S. policy adopts heavily from Boruch and Cordray (1980) and other documents cited in the text.

[2] See Verma's (1977) discussion of research on medicine in medieval India and Rabinovitch's (1973) fine description of rabbinic thought on evidence during the ninth through twelth centuries. The Syrian experiment is described by American researcher F.S. Chapin (1947); the English and Scottish tests are described in Cochran (1976) among others.

[3] Two other efforts, independent of these, are worth examining because their conclusions differ at times from ours. The first, undertaken by Rand Corporation staff, appears in a monograph by Pincus (Ed.). The second, prepared by members of Stanford's Consortium on Evaluation Research, is presented in a volume by Cronbach and others (1980).

REFERENCES

AIRASIAN, P.W., KELLAGHAN, T., and MADAUS, G. The Consequences of Introducing Educational Testing: A Societal Experiment. Boston, Mass.: Boston College, Center for Field Research and School Services, October, 1978.

BEN-SHAKHAR, G. and NEVO, B. Some Thoughts Regarding the Possibility of Evaluating a National Matriculation Examination System. Hebrew University, Jerusalem, March, 1981.

BLASS, N. Educational Research and Educational Policy-Making: Reflections on the State of the Art Based on Empirical Evidence in the Israeli Ministry of Education. Office of the Chief Scientist, Ministry of Education and Culture, Jerusalem, March 1981.

BLASS, N. Research Utilization and Scientific Knowledge Impact on Public Policy: Some Thoughts and Guidelines for Proposed Research. Office of the Chief Scientist, University of Education and Culture, Jerusalem, 1980.

BORUCH, R.F. and CORDRAY, D.S. (Eds.) An Appraisal of Educational Program Evaluations: Federal, State, and Local Agencies. Report submitted to the U.S. Department of Education. Evanston, Illinois: Psychology Department, Northwestern University, 1980. (To be published by Cambridge University Press).

BORUCH, R.F. and WORTMAN, P.M. Implications of Educational Evaluation for Evaluation Policy. In: D. Berliner (Ed.) *Review of Research in Education*, 1979, 7, 309-363.

BORUCH, R.F., WORTMAN, P.M., and CORDRAY, D.S. (Eds.) *Reanalyzing Program Evaluations*. San Francisco: Jossey-Bass, 1981.

CHAPIN, F.S. *Experimental Designs in Sociological Research*. New York: Harper, 1947.

COCHRAN, W.B. Early Developments in Comparative experimentation. In D.B. Owen (Ed.) *On the History of Probability and Statistics*. Base: Marcel Dekker, 1976, 1-27.

COOLEY, W.W. Improving the Use of Evaluation Results. Presented at the Annual Meeting of the Evaluation Research Society, Arlington, Virginia, November 21, 1980.

CRONBACH, L.J. and Associates. *Toward Reform of Program Evaluation.* San Francisco: Jossey-Bass, 1980.

DAVIS, D. Standardized Evaluation of Education Programs. Hebrew University, Jerusalem, June 1980.

HILL, P. Evaluating Education Programs for Federal Policy Makers: Lessons from the NIE Compensatory Education Study. In J. Pincus (Ed.) *Educational Evaluation in the Public Policy Setting.* Santa Monica: Rand Corporation, May 1980. (R-2502-RC).

KENNEDY, M. School District Use of Evaluation Research. Presented at the Annual Meeting of the Evaluation Research Society, Arlington, Virginia, November 21, 1980.

KUGELMASS, S. Considerations Toward a Policy of Evaluation Research: The Case of the Chief Scientist at the Israeli Ministry of Education. Office of the Chief Scientis, Ministry of Education and Culture, Jerusalem, 1980.

LEWY, A. Continuous Evaluation Services Attached to Large Scale Action Projects: The Case of the Evaluation Unit at the Israel Curriculum Center. Israel Curriculum Center (ICC), Ministry of Education and Culture, Jerusalem, 1980.

LEWY, A. Standards for Research Data. Israel Curriculum Center, Ministry of Education and Culture, Jerusalem, 1980.

MAGIDSON, J. Toward a Causal Model for Adjusting Preexisting Differences in the Nonequivalent Control Group Situation. *Evaluation Quarterly*, 1977, 1, 399-420.

RABINOVITCH, N.L. *Probability and Statistical Inference in Ancient and Medieval Jewish Literature.* Toronto: University of Toronto Press, 1973.

RAIZEN, S. and ROSSI, P. (Eds.) *Program Evaluation in Education: When? How? To what ends?* Washington, D.C.: National Academy of Sciences, 1981.

RIECKEN, H.W., BORUCH, R.F., CAMPBELL, D.T., CAPLAN, N., GLENNAN, T.K., PRATT, J., REES, A., and WILLIAMS, W. *Social Experimentation.* New York: Academic, 1974.

ROSSI, P., BERK, R.A., and LENIHAN, K.J. Money, Work, and Crime. New York: Academic, 1980.

STALFORD, C.B. NIE-Funded Studies of Local Evaluation. Presented at the Annual Meeting of the Evaluation Research Society, Arlington, Virginia, Nov. 21, 1980.

TAMIR, P. Continuous Evaluation as an Integral Part of Curriculum Development: The Case of the Evaluation of the High School Biology Project at the Israel Science Teaching Center. School of Education and Israel Science Teaching Center, Hebrew University, Jerusalem, May 1980.

VERMA, R.L. Graeco-Arabic Medicine in Medieval India - Its Hakims and Hospitals. *Hamdard Medicus*, 1977, 20, 26-40.

WHOLEY, J. Evaluability Assessment. In R. Rutman (Ed.) *Evaluation Research Methods: A Basic Guide.* Beverly Hills, CA.: Sage, 1977, pp. 13-37.

ZAK, I. Latent Variables in Causal Models. School of Education, Tel Aviv University, Tel Aviv, June 1980.