A CLASSIFICATION AND COMPARISON OF EVALUATIVE ACTIVITIES

by

Kathleen A. Archibald

March 1976
Working Paper 262

TABLE OF CONTENTS

FIGURES

FOREWORD

This working paper was written in 1968. It was directed to the issue of designing an evaluation system for federal manpower programs. Reactions to it at that time were extreme: it was either loved (mostly by sociologists) or hated (mostly by economists). My own reaction was somewhere in between and I decided that a revision of the paper for publication would take more time than I was willing to devote to it.

The paper dates from that period when evaluation of government projects and programs was considered by many a relatively new endeavor. Many of the issues discussed are now "old hat." Some of the material included, however, has not yet received much attention in the now voluminous literature on evaluation. For this reason, I assume, the paper continues to be cited.

I am pleased that the publication program of the Institute for Urban and Regional Development provides an opportunity to make this paper more accessible. I have not attempted to update the paper, but I have shortened it and made some minor revisions.

Kathleen A. Archibald
Berkeley, California
March 1976

Evaluation, as the term will be used in this paper, is a method of determining how effective some program, project, or treatment is in meeting its objectives under operating conditions. The concept is a simple-minded one, but the task of evaluating is messy from a scientific point of view and a mixed blessing from an administrative point of view. This paper looks at various ways of evaluating social service programs in general and job training programs in particular.

DISTINCTIONS BETWEEN EVALUATIVE RESEARCH AND RELATED ACTIVITIES

We will distinguish evaluative research from three other kinds of activities: (1) compliance control, that is, checks for administrative and legislative conformity, (2) subjective appraisals of success, and (3) other kinds of research. In adopting the term "evaluative research," we follow Suchman's suggestion that "evaluation" be used to refer to "the general process of judging the worthwhileness of some activity" and "evaluative research" to evaluations that utilize scientific methods [1].

While distinguishing between evaluative research and other activities is only a semantic exercise, in this case it may be a useful exercise, since certain activities frequently called evaluation do not in fact increase knowledge about the relationship between operating programs and their success in meeting objectives. People may call these activities whatever they wish, but it is worth distinguishing between those which relate effectiveness to program operation and those which do not, between those which provide only vindication and those which provide verification [2]. Subjective appraisal and evaluative research are both methods of evaluation; compliance control is not.

Compliance Control. It is all too humanly easy to confuse conforming performance with successful performance, particularly where success is hard to measure, difficult to achieve, or both. The term evaluation is often used to refer to data collection which permits a check on a project or program to make sure it is obeying all rules and regulations

that are supposed to govern its operation. Such conformity may be necessary for the program's continuing existence, but it is neither necessary nor sufficient for effective performance in meeting objectives. The use of such conformity as a measure of effective performance is an example of a common characteristic of large organizations -- the "ritualization of means" [3]. Where it is difficult to perceive the relationship between means and ends, people and organizations tend to transfer their allegiance from ends to means.

Checks for administrative and legislative conformity tend to encourage the ritualization of means and, to the extent they do, they are the very antithesis of evaluation. The useful thrust of evaluation is to encourage people to look at the relationship between means and ends and improve the means to better reach the ends. The thrust of checks for administrative conformity is to focus attention on means and on only those means which are requirements imposed from above. This suggests that the two activities be carefully distinguished, perhaps organizationally as well as verbally.

Subjective Appraisals of Success. Subjective appraisals of success refer to evaluations based on judgment and opinion alone. It may be the judgment of experts, of participants in the project, of various sectors of the community, or of the mass media.

The good opinion of some of these people, for instance, participants in the project or politically powerful members of the community, may be important and even necessary for the continued existence of the project. Subjective appraisals are very useful in this respect, that is, as indicators of needed popularity and goodwill. But this is vindication not verification. Programs may be popular and respected for reasons that have little to do with effectively achieving stated objectives and subjective evaluations of program success should be treated with some caution.

Since it is difficult and expensive to measure actual achievement in many action programs, formal or informal subjective appraisal is by far the most common kind of evaluation. In many instances, subjective appraisal probably does correlate highly with actual success; that is, it provides a valid proximate criterion of success. In other

instances -- and who knows what proportion, subjective appraisals are misleading, but this risk can be decreased if attention is paid to political, professional, and social constraints on various potential judges. The most common error tends to be that of asking the opinion of those who have some vested interest in the success of the program. Our faith in "professional opinion" puts us most off guard in this respect; we seldom discount or control for professional vested interests. Professionals asked to evaluate a project are called upon to make a judgment about colleagues they may be loath to deprecate and about methods they have come to take for granted as efficacious.

While it is easy to point out the dangers in subjective appraisal and suggest caution in its use, it is by no means easy to suggest an alternative that for the same cost and within the same length of time can produce as useful and relatively reliable results. As will be pointed out later, much of what currently passes as evaluative research has little more claim to reliability than subjective appraisal, and is more costly in time and money. Rigorous evaluative research, on the other hand, seldom can address itself to the plethora of relevant issues that need to be taken into account for certain evaluative purposes; subjective judgment can. Finally, management information systems which promise a golden shower of data need to be taken back to the assay office since there are serious questions as to whether the benefits of mining that ore are worth the costs, monetary and non-monetary. This leaves subjective appraisal as not only a common form of evaluation, but also a very useful one.

Subjective appraisal may also be used as one of several proximate criteria of worthwhileness in evaluative research. In a research context, reference is usually made to subjective ratings rather than subjective appraisal since evaluative research will tend to structure a judge's response so that it can be quantified whereas subjective evaluation goes after lengthier, qualitative opinions. In evaluative research, attention would, or should, be paid to the reliability of such ratings, their correlation with other proximate criteria, their relationship to ultimate criteria, and to the relativity of implicit scales used by different judges.

Evaluative Research and Other Research. We suggest that the term "evaluative research" be limited to attempts to measure the effectiveness or adequacy of a program, project, or treatment under operating conditions. By this definition, evaluative research pays attention not only to the desirability of a particular project or treatment but also to its feasibility: its expense compared to resources available and to other alternatives, the ease or difficulty of implementation, and its abitity to withstand subversion by routinization.

Applied research may look at the success of a particular treatment or technique under experimental conditions only. This kind of experimental investigation of treatment components can become part of a broader strategy of evaluative research when, subsequent to experimental success, an estimate or test is made of the effectiveness and adequacy of the treatment under relatively routine operating conditions. In public service and social action programs, "routine operating conditions" include a multitude of intervening variables that affect the eventual outcome of any particular technique or treatment. If we wish to estimate or evaluate real-world success, these confounding variables must somehow be taken into account.

THE DESIGN OF EVALUATIVE RESEARCH

There are several ways of looking at the basic structure of evaluative research. The simplest model is a listing of questions suggesting what one may want to find out in an evaluation. A comprehensive set of questions has been provided by Sainsbury for the evaluation of community mental health services [4]. These are presented here with appropriate changes to make them applicable to the manpower training field:

(1) Who needs what kind of services?
(2) How does the introduction of services alter needs?
(3) To what extent are the original and derivative needs met?
(4) What are the effects (intended and unintended, negative and positive) of the services provided on the clients?
(5) What are the effects on other people involved (client's family, the personnel involved in the program)?
(6) What are the effects on organizations involved (employers, community organizations, pre-existing trainnng programs, etc.)?

(7) What does the program cost?

(8) Are there specific and local questions that should be taken into account?

This is an exhaustive set of questions and few evaluations in any field have attempted to cover all of them. Priorities must be set in each particular instance, with attention to the marginal cost and utility of collecting additional information.

Design Flow Chart. Figure 1 is a flow chart, adapted from Greenberg and Mattison [5], illustrating the basic experimental design and thus the basic logic of evaluative research. Variations in design are derivative from it and can be appraised for rigor in comparison to it [6].

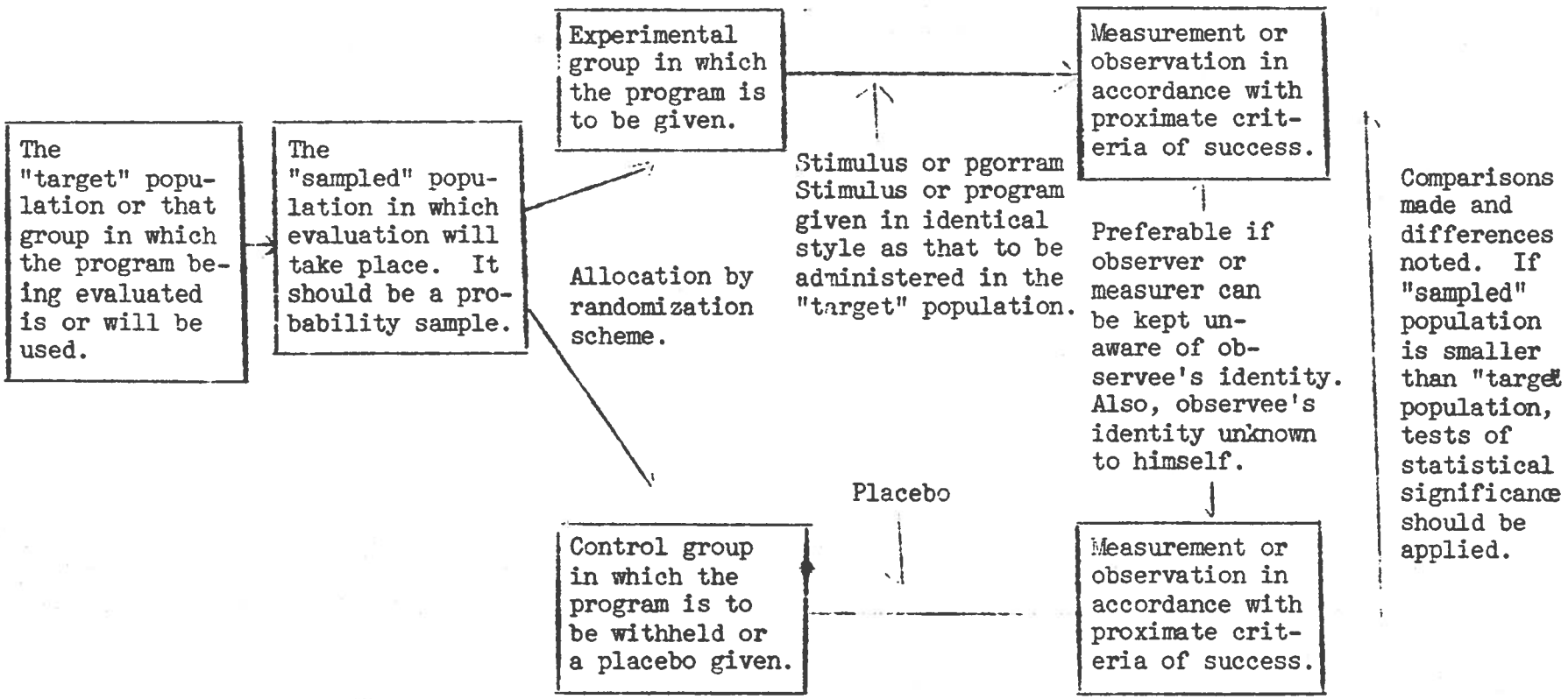
An examination of the flow chart suggests the practical difficulties of conducting evaluative research, in particular, the problem of a control group or some equivalent and the problem of providing a stimulus or treatment that can be said with any certainty to be identical to th t to be generally administered to the target population. This latter involves the problem of adequately specifying treatment components, to be discussed in the next section, and also such difficulties as controlling for the experience, ability, interest, and personality of the staff delivering the services.

Hand in hand with these technical difficulties in conducting rigorous evaluative research go a number of social difficulties. Evaluative research in an ongoing program tends to be a disruptive and potentially threatening experience for the practitioners. The researcher must cope with political and interpersonal relations within the program and there will be considerable pressure on him to compromise his research design, for with a sacrifice of some elegance he can expect increased cooperation and interest on the part of the program staff [7].

The deletorious effect of these technical and social difficulties on the design of evaluative research is demonstrated in a study by John Mann. He examined 181 evaluation studies in the fields of psychotherapy, counselling, human relations training, and education. These 181 studies were chosen principally because of their methodological superiority from an initial pool of 600 studies of "relatively high metnodological caliber."

FIGURE 1

FLOW CHART OF EVALUATIVE RESEARCH DESIGN



They represent the best of evaluative research in four fields where considerable evaluative research has been done. His findings on the methodological errors in these methodologically superior evaluations were:

> ...extremely damaging to the cause of evaluative research. With two or three exceptions, the errors are of a major character. In other areas of research in the behavioral sciences, any of them would probably render a study unfit for publication. They are not errors within subtle experimental refinements. Rather, they reflect the abuse of scientific procedures....These findings raise grave doubts as to whether any conclusions can be drawn from such research [8].

Most other writers, while agreeing that evaluative research is generally of poor quality, do not come to the pessimistic conclusion above but instead call for more and better evaluations. None of these other writers, however, has done as careful a study as Mann of the methodology and findings of evaluative research. While Suchman reviews much of the literature in his recent book he does not ask what has been found out by evaluative research nor does he tabulate methodological errors in past work and estimate their seriousness. Further, many of Suchman's examples are from the evaluation of drugs in medical research, a relatively simple type of evaluation and one that is comparable to the experimental evaluation of treatment components rather than to whole project or whole program evaluation [9].

Before discussing the implications of these methodological difficulties, we will look at one more model of evaluation that focuses attention on some additional problems.

An Evaluation Model. Another way of looking at evaluative research is in terms of key variables and the relationships between them. Figure 2 represents an ideal basic structure of evaluation of manpower training.

This model serves as a useful reminder on several points. First, it points out the need to look at unintended proximate effects as well as intended ones. In the manpower training field, and the war on poverty generally, there is considerable reason to believe that the costs or undesirable effects of program failure can be quite high. Such negative side-effects are anticipatable. They can be derived from the broad theoretical point of view, that has almost become a cliché in American society.

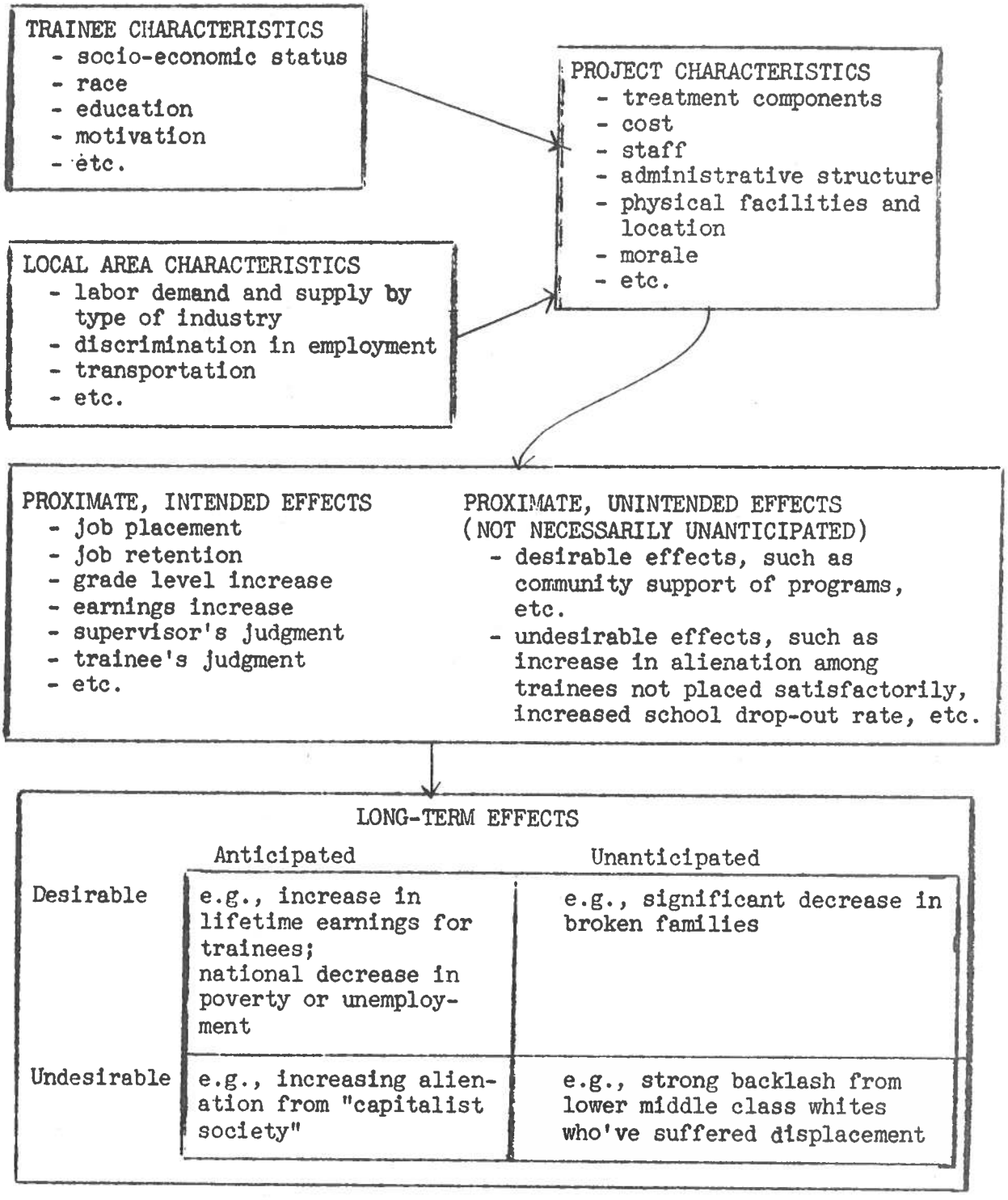


FIGURE 2

MODEL OF MANPOWER TRAINING EVALUATION

that it is unwise to raise expectations and then not fulfill them. Or they can be derived from many bits and pieces of specific data and knowledgeable opinion [10]; or from a simple estimate of the likelihood of failure and an analysis of its probable consequences [11]. Even if a program is successful in placing, say, 70% of its graduates in satisfactory jobs -- an uncommon success rate -- the 30% who have undergone training and are not placed may represent the most socially significant group. It is not unlikely that within that 30% lies most of the potential for increases in crime, welfare dependency, and riots. While the identification of that 30% and an estimate of the consequences of failure in their case will not solve the problem, it could provide a first step in that direction by raising a warning flag and encouraging a search far more relevant programs. Despite the potential importance of undesirable side-effects in the manpower training field, the great majority of evaluation studies in the field have failed to search them out.

Second, the model reminds us of the obvious but too conveniently forgotten point that the "real" objectives of manpower training are the long-term ones. The goal of manpower training programs is not immediate job placement nor even job retention but improvement in lifetime earnings or job satisfaction or both from an individual point of view, and a reduction in poverty or unemployment from a national point of view. Essential to adequate evaluation and planning of manpower programs are examinations of the relationship between proximate and ultimate criteria of success and analyses of probable long-term effects. The methodological desires of researchers and the vindication desires of program administrators both lead to an overemphasis on proximate criteria relative to long-term consequences.

Brief mention should be made of the reference to unanticipated consequences under long-term effects. We assume that there may be program consequences that are neither anticipated nor intended when the program is first planned and put in operation. It can be argued that the evaluation of an established program should include a search for such unanticipated effects [12].

Finally, and most importantly, by placing emphasis on variables, the model suggests that one of the gravest difficulties in current evaluations, in manpower training as in other social action fields, is the

specification of relevant project variables. Most often a project is treated as a single undifferentiated variable; a research subject either participated in it or he did not. Such evaluations have limited utility since they do not permit the identification of means of improvement. The difficulty is not solved by the rough breakdowns sometimes used, breakdowns into counselling, skill training, remedial education, and so forth. Each of these services may vary across a wide range. They do not represent basic, replicable components; each is complex enough that a comprehensive description would be a major undertaking [13].

Mann's study is extremely pessimistic on this point with respect to evaluation as usually conceived. We have already mentioned his sad findings on the methodological errors in "good" evaluation research, but he is even more critical with respect to the intrinsic difficulties of adequately handling project variables in the usual evaluation situation.

> ...regardless of the quality of the research itself, there remains the fact that it does not seem to produce any positive conclusions, except that change is consistently demonstrated in a certain fixed proportion of the studies. This is the heart of the problem. Evaluative research is intended to distinguish among methods of changing behavior, determining the most successful procedures, and to clarify the process of behavior change itself. To do this the demonstrated change must be related to other significant variables, such as the content area investigated, instruments used, and methods tested. None of these relationships can be clearly established. One is driven, therefore, to the inescapable conclusion that evaluative research shows no prospect of reaching these goals.
>
> Such a damaging conclusion requires an explanation. The most reasonable explanation the author can offer, on the basis of his direct examination of the evidence, is that evaluative research is not undermined so much by the problem of its execution as by the methods it attempts to evaluate. The ingredients of evaluative studies are inappropriate to scientific methodology, which, like any good recipe, requires the use of specific pure elements that are combined in known proportions and in a fixed time schedule. Virtually all methods evaluated by the studies reviewed were of such complexity as to defy description in terms of a limited number of carefully specified variables. ...If the methods that are tested cannot be precisely described, then the results can never be cumulative, since no one can state what was tested [14].

Manpower training programs may not cover as complex a set of operations as psychotherapy, but Mann's analysis included an examination of educational evaluation as well; and education, it can be argued, entails a

simpler set of treatment components than a comprehensive manpower training program. Also Gage reviewed evaluation studies on teaching effectiveness, a mere subset of educational evaluation, and comes to conclusions somewhat similar to those of Mann [15].

Mann concludes that evaluative research is a "scientific blind alley." For the areas he looked at he states: "It has failed to validate itself in practice and the sooner its failure is accepted and recognized, the easier will be a transition to another approach to the same problem" [16]. What Mann suggests instead of the traditional approach is that experimental analysis of treatment components in a laboratory setting, using factorial designs to test the relative effectiveness not only of individual components but also of each unique combination. This rigorous experimental approach offers the hope of solving the problem of specifying relevant project variables and also promises to avoid the seemingly inevitable methodological weaknesses that arise when evaluative research is conducted in an operating project.

Mann and Gage are in the minority in their wish to abandon traditional research approaches to evaluation. Given this disagreement among writers on the topic, what kinds of conclusions can be drawn about the value of evaluative research? Should we buy Mann's pessimism or should we look to those authors who see better evaluations as one of the best hopes we have for improving public service programs? Should all evaluative research aspire to the levels of scientific nobility that many social scientists seem to expect of it? Are there some humble tasks that can be performed by scientifically clumsy, but warm and well-meaning, evaluative research? Or does the shakiness of such work completely undermine its utility? Can even sloppy evaluation help inform intuitive judgments? Does it at least force us to think about the right questions? Or might it lead us further astray?

Questions such as these cannot be answered in a general way. If we are worried about the dangers of believing in invalid evaluations, we can take solace in the fact that there are few recorded instances of evaluative research leading to program change. If we are interested in capitalizating on any potential that evaluative research may have, however, we may make some progress by being more specific, both about the types of evaluation and the various purposes and functions it may serve.

PURPOSES AND FUNCTIONS OF EVALUATIVE ACTIVITIES

Evaluative activities, or activities that masquerade as such, can serve at least six distinguishable purposes or functions. They may help to vindicate or justify an ongoing project or program; for instance, evaluation showing the program is worthwhile in one way or another may be important in getting the political or financial support needed for its continuance. Vindication is very often a necessary activity -- without a budget no objectives are going to be reached. Evaluation may also be useful for salesmanship or diffusion, a somewhat different function than vindication. Salesmanship occurs when there is a desire to have others adopt or implement a particular treatment or type of project. A third function of evaluation is verification: is a program really doing what it is supposed to do and without too many adverse side-effects? Here the success of a program is compared to the effect of no program. A fourth function is improvement. The verification question is "What are we doing?" The improvement question is "How can we do better?" To obtain information relevant to improvement, a treatment or program has to be compared to alternative methods, real or hypothetical, of achieving the same objective. A fifth possible function of evaluation is to increase understanding of the way in which a treatment or program leads to success or failure. A sixth possible function of evaluative activities is to ease anxiety or insecurity. This is best done via the ritualistic collection of seemingly relevant but in fact innocuous data since a relevant evaluation may prove disconcerting rather than soothing. This sixth function is actually a special case of vindication. It is self-vindication for those who need to believe in or have a feeling of control over a program they are working in or responsible for.

It should be noted, for the sake of closure, that if evaluation has the power to help produce the six consequences listed above, it also has potential for producing just the opposite effects. If evaluation is useful in showing that a program is worthwhile, it can also indicate that another program is worthless. If evaluation is badly done, improvement attempts may have counterproductive consequences, and so forth.

These six functions help to make some sense out of the varying evaluations of evaluative research. Mann and Gage are interested in improvement and in increases in understanding. They want to be able

to identify the specific variables that account for success or failure and they want knowledge about these variables to be cumulative and generalizable. These are admirable ambitions, but vindication and verification are important too, from a decisionmaker's point of view if not from a scientist's. The methods that make Mann unhappy because of his interest in information that will lead to improvement recommendations are quite serviceable if one is only interested in verification. Further, while Mann's suggested experimental approach does offer the best hope of findings that can lead to the improvement of projects, it will nevertheless often be necessary to test out treatment components in an operating setting after they have shown promise in the laboratory setting. The reasons that lead you to the laboratory in the first place, that is, the complexity of an operating project, make it necessary, in a second step, to leave the laboratory and its ceteris paribus seclusion and find out if the promise of a treatment component holds up under the bombardment of intervening variables encountered in a real-life setting.

The real trick in developing a sensible evaluation strategy from an action or decisionmaking point of view lies in the way in which these six functions are combined and assigned priorities. A single evaluation could seldom if ever serve all purposes. If an increase in understanding is wanted, it is highly likely that vindication will have to be forfeited. If improvement is wanted, increased anxiety among project personnel may very well be the price. The terrain of evaluative research would look very different, and there would be little disagreement among writers on the topic, if increases in understanding were inexpensive to obtain and produced as inevitable by-products feasible improvement recommendations, verification, vindication, and anxiety reduction.

Any attempt to explore the actual possibilities of combining these functions harmoniously has to take into account the organizational structure that will sponsor and utilize evaluations. The next section will attempt to relate manpower training evaluation to different decision levels within the Office of Economic Opportunity and the Department of Labor and to specify what kinds of evaluation are most appropriate at each level.

EVALUATION AND ORGANIZATIONAL DECISIONMAKING

The typology that follows as Figure 3 attempts to take the decisionmaker's point of view not the scientist's. Because funds for the war on poverty are in short supply and because evaluative research has often been a weak tool, the typology and the subsequent discussion rests on the assumption that evaluative research and systematic data collection must earn their keep. To be more specific, we asuume that evaluative research should only be supported when: (1) it will provide solid evidence useful for decisions to be made in the future; (2) it will show that a program is or is not worthwhile for vindication purposes; and/or (3) it is specifically required by legislation. The focus in the discussion to follow is on the decisionmaking and vindication uses of evaluation. The evaluation strategy suggested meets most of the evaluative requirements specified in the Economic Opportunity Act of 1964 as amended in December 1967, but these legislative requirements will not be explicitly discussed.

The term "project" refers to local service units, such as East Los Angeles Youth Training and Employment Project and the Manpower Programs Laboratory of Mobilization for Youth in New York. The term "program" refers to functional units nationally, such as the Job Corps and the Neighborhood Youth Corps. "Director" refers to the top management levels of the Office of Economic Opportunity and the Department of Labor, including their respective research and evaluation sections, Research, Plans, Programs, and Evaluation (RPP&E) and the Office of Manpower Policy, Evaluation and Research (OMPER).

At each of these three levels, different kinds of decisions are made and, in general, different factors or variables become important at each level. In the typology and following discussion, only those decisions to which evaluation can contribute will be considered. We will discuss the three levels one at a time in an attempt to specify the kinds of evaluation most appropriate at each level.

Project Level. At the project level, decisions are made relating to adaptation and improvement of local activities. Such decisions may involve the fit between local needs and project activities, the overall effectiveness of the project, the relative effectiveness of various components of the project, the abilities of the staff, etc. These are the

FIGURE 3

A TYPOLOGY OF THE RELATIONSHIPS BETWEEN EVALUATION AND DECISIONMAKING

DECISION LEVEL	PROJECT -- local	PROGRAM -- national	DIRECTOR -- national
KINDS OF DECISIONS	Project adaptation and improvement.	Overall improvement at project level; bringing worst projects closer to best; introduction of new treatments in projects, including new projects; initiation and termination of projects.	Improvements at program level; allocation among programs, incl. initiation, continuation, and termination decisions.
NEED TO JUSTIFY AND VINDICATE EXISTENCE TO WHOM?	To program level; to community; to clients; to professional colleagues.	To agency top management; to national media.	To Congress, Bureau of the Budget and the White House, to national media.
DIFFERENCES WITHIN EACH LEVEL, I.E., HORIZONTAL DIFFERENCES, IMPORTANT FOR EVALUATIVE PURPOSES.	Implementation of treatments or means (staff differences, technique differences, structural differences); community served; kinds of clients.	Sub-objectives, including needs they are trying to meet; content of projects mounted and supported (treatments, components or means); target populations served; political support.	Broad objectives, including needs they try to meet; political and financial support.
KEY "EMERGENT" PROBLEM IN EVALUATION (AND PROBLEMS INESCAPABLY SUBSUMED FROM LOWER LEVELS).	Problem of feasibility, i.e., setting reasonable standards and aspirations given local conditions.	Problem of comparability of independent and intervening variables; (problem of feasibility).	Problem of comparability of dependent or criterion variables, i.e., negative and positive effects, benefits and costs, compared across programs, (problem of comparability of independent and intervening variables); (problem of feasibility).
SPECIFIC "NOBLE" FUNCTIONS OF EVALUATIVE RESEARCH	Immediate and continuous feedback device, so that standard can become improvement over past performance. [Changes within only.]	Aid in setting standards of achievement at and for project level; aid in deciding on new treatments and project designs and in persuading projects to adopt them. [Changes within and below.]	Aid in deciding how worthwhile a program is relative to cost. [Changes below only.]
KIND OF EVALUATIVE RESEARCH INDICATED	Routinized data collection permitting periodic comparisons of performance on a few key variables; quick and inexpensive operations research; soft pre-post designs using project as its own control when change is introduced; evaluation of the project by its clients.	Best-worst comparisons among projects to identify meaningful variables; rigorous experimental testing of treatment components followed by evaluation of components in routinely operating projects; surveys to identify unmet needs; work on side-effects.	Cost/benefit studies; relationship between proximate and ultimate criteria; studies of total negative and positive impact of all programs, including unmet needs; routinized data collection on numbers entering and completing programs, numbers placed in jobs, costs, and subsample survey of job retention.
TIMING	Evaluation procedures introduced in early stages of project.	Experimental work and need surveys at beginning of program; other work only after projects have settled down to routine.	Routinized data collection introduced at beginning of projects, other work after things running routinely.

internal concerns of the project but, while the typology primarily deals with these internal functions, it is worth considering first the factors which lower the likelihood of any evaluative activity at the project level.

To keep the money coming in, the project has to be able to justify itself to its continuing sponsors, in this case, the program level. To keep clients coming in, the project has to reasonably well thought of among the target population. A project might be extremely effective, but without continued financial support and client interest it would fail. These are the necessary conditions for project success, and it would be surprising if projects did not respond first to these necessary conditions for success and only with time or money left over to the much more difficult challenge of improving project performance. Vindication activities are likely to take precedence over evaluations suggesting improvements unless the two are either intimately linked (for instance, by orders from above) or completely separate so that those responsible for improvement activities can remain completely independent from those worrying about short-term stability.

Even more important in assessing evaluation potentialities at the project level is the fact that verification of performance through evaluative research is not the only way of vindicating a program. In fact, it tends to be the most expensive and treacherous way of attempting vindication. Although solid evidence of effective performance would have an admirable influence on sponsors and clients willing to weigh the evidence, one can ask how many projects feel confident enough of both their own effectiveness and the rationality of their clients and sponsors to risk investing in hard-nosed evaluative research when there are cheaper and surer ways of justifying their continued existence. For instance, if the local community and the professional community think highly of a project, there is a good chance of its next year's budget being approved. The pressures on a local project suggest that securing goodwill may be considered more important than verifying actual performance. Such a priority is quite realistic from the project point of view, particularly if its year-to-year existence is precarious.

It is probably quite unrealistic to expect a project to report out, to any unit to which it has to justify its existence, reliable and valid evaluation data on its own overall performance. Without strength of character and puritanical honesty at the project level, the data . . .

collected will be "processed" so as to present a complimentary profile. Even with strict honesty and nonpragmatic ethics, the collection of data on individual clients is unlikely to get sufficient attention if it conflicts with or even competes for time with provision of services, justification of budget requests, cultivation of goodwill, and maintenance of morale. Thus data collection is liable to get short shrift at the project level unless project personnel perceive it as useful for their own purposes.

If reliable data on project performance is wanted at the program level, it does not seem sensible to rely on the project to collect and report it out on a routine basis. It is difficult to conceive of appropriate safeguards that could ensure reliability given the incentives for fudging that derive from the program's control over the project's fate [17]. Further, there seems to be little need at the program level for routine data on the performance of individual projects or, more precisely, insufficient need to justify the expense, disruption, and acceptance of unreliable data that routine transmittal of data from project to program is likely to involve. The argument here is certainly not against all data collection at the project level, but against the routinized collection of extensive data on individual clients at the project level for reporting to the program level. It is thus an argument against routinized data systems like SDC's Manpower Management Information System.

An additional factor supporting this argument is the invasion of privacy entailed in a nationally prescribed management information system that tracks individuals through training programs and into employment. It is not only a matter of the intrinsic value of privacy; it is also a matter of practical consequences. For the individual trainee, there is the possibility of his records being made available to persons or agencies that he would prefer to withhold information from, and for the agency interested in the data, there is one more reliability problem in that questions resented as invasions of privacy are liable to be falsely answered. It would not be surprising to find collusion between counsellors and trainees in falsifying information on such records [18].

The program level does need data about projects; what is doubtful is whether they need (or, for that matter, will even look at [19]) routine data on all individuals going through all projects. We suggest

that the data a project chooses to report when requesting funds combined with subjective appraisals and with evaluative data the program level itself collects on a sampling basis is sufficient to meet the needs of decisions made at the program level [20]. This will be discussed further in the section on programs.

Routine data is more useful at both the project level and the director level than at the program level. If the transmission route is from project to director level, the incentives on the project to fudge can be minimized. The director level does not control allocation to individual projects and is not interested in the performance of an individual project. This could, for the project personnel, take much of the threat out of the data collection effort: the director's lack of interest in individual project performance could be made clear by letting the projects in one region aggregate their data before reporting it to the agency.

The routine data of interest at the director level, as will be suggested later, consists of a few simple figures. The interesting thing about the simplicity of the director's information needs is the leeway it permits to build management information systems that conform with the interests and needs of individual projects. Management information systems for training programs have so far been designed to meet the presumed information needs of the program and director levels with little attention to the utility of this data at the project level. Yet such systems are dependent on project personnel for the collection of reliable data. We have already suggested that if such data is reported to the program level on a project-by-project basis, there is actually a disincentive to conscientious data collection. Even if it is only reported to the director level, there is no incentive for careful data collection unless the data prove to be useful at the project level. If project personnel have a stake in the data collection and it represents no threat, reliable data will be gathered.

If projects are structurally quite similar, this might suggest designing a data system that meets project needs and, of course, incidentally provides the few figures needed at the director level. If projects vary considerably among themselves, they could be permitted to do what they wanted with respect to data collection, other than the figures needed nationally. The money that might otherwise be used to develop an

elegant nation-wide management information system to collect data that may be useless, because of their unreliability if not for other reasons, could instead be spent to provide information-processing consultants to projects on request.

Such a system gives the projects a great deal of leeway in their data collection procedures. More importantly, the leeway with respect to information collection facilitates flexibility in other endeavors, whereas a routinized data system specified from above will tend to constrain flexibility in other areas.

When we consider the internal uses of evaluation, and forget about the external use for vindication, we find that there is a distinctive emergent problem at each decision level. At the project level, key decisions concern adaptation and improvement. Evaluation can make significant contributions to these internal concerns. The distinctive problem of evaluative research at the project level is that of feasibility, a problem that arises because of the differences among projects (Row 3 in the typology). How can a project derive for itself realistic standards of achievement? Is a 50% job placement rate good or bad given local conditions? Does a 70% rate in another project mean that the project is better or that its clients and labor market are better? There are some ways of dealing with such comparability questions but the research needed is expensive and inappropriate for sponsorship at the project level. Such research should be handled at the program level, and programs can help set standards of achievement for projects. But improvements at the project level often need to take local idiosyncracies into account and evaluation can be of help on this. A project can handle the feasibility problem relatively satisfactorily by using its own past performance as its measuring stick; improvement over past performance becomes the goal.

Evaluation sponsored and conducted at the project level is best seen as an immediate and continuous feedback device keeping tabs on current performance and attempting to improve upon past performance. It can let the staff know how they are doing and provide some hints as to how they can do better. The kinds of evaluation that can be usefully conducted within one project are those that quickly provide usable data to project personnel: for verification purposes, routinized data collection permitting periodic comparisons of performance on a few key variables;

and for improvement purposes, quick and inexpensive studies or, when an important change in procedures or content is introduced, simple evaluative research using a pre-post design that considers the project as its own control.

Evaluation's main contribution at the project level is in helping to maintain adaptability and flexibility. If it is to do this, it needs to be an integral part of the project's operations and not an insulated appendage of little perceived interest to the practitioners. This suggests that evaluation at this lowest level should be introduced early in the life of a project before structures and processes have become routinized and while the staff is still open to the possibilities of change and interested in fresh starts. If evaluation can be routinized along with other procedures, flexibility and an interest in innovation may well become institutionalized within that project.

Trainee Ratings as Proximate Criterion. One proximate criterion that has received little serious attention in manpower training evaluations to date is the rating of a project by trainees in it. Such ratings which may attract more attention from now on, thanks to the December 1967 amendments to the Economic Opportunity Act, could possibly provide a useful criterion variable at each level. We discuss them here since their utility is more evident and immediate at the project level.

The advantages of collecting subjective ratings by trainees at the project level are the following:

(1) A flow of clients into and through a project is a necessary condition of project success. Satisfied trainees and graduates provide the best advertising a project can have. Trainee ratings provide a means of monitoring the level of satisfaction.

(2) Evaluation at the project level is most sensible if used as a rapid feedback device facilitating adaptability. Ratings by trainees serve this function well because they have a face validity that is difficult for project personnel to discount. It is difficult to continue justifying business-as-usual after finding a majority of your trainees dislike the program or consider it useless. Such findings provide far more change leverage than other criterion measures which are more easily explained away [21].

(3) We know there is a communication and understanding gap between ghetto dwellers and the middle class. The ultimate success of the poverty program as it stands is largely dependent upon the way in which the poor perceive the various activities. Their evaluation of projects is important and assumptions about their evaluations made by middle class professionals are quite likely to be wrong. Specific questions to trainees about what they want and what they are getting can provide a useful channel of communication.

(4) Evaluation by trainees provides the best safeguard available against unintended and unrecognized paternalism in projects.

(5) Work done in education has demonstrated that "student evaluation is a useful, convenient, reliable, and valid" criterion measure. If trainees respond to the evaluative task as students in formal educational settings do, then trainee ratings, as a scientific measure, might be as good as or better than other proximate criteria in use. With respect to reliability, it was found that "if 25 or more students ratings are averaged, they are as reliable as the better educational/mental tests at present available." With respect to validity, perhaps the most interesting finding in the educational field is the substantial agreement between current students and 10 year alumni on the relative importance of certain teacher characteristics and in the rating of instructors [22]. It could be argued that this is reliability <u>in extenso</u> rather than validity. Since we do not have a surfeit of ultimate criterion measures in educational and training fields, however, people's opinions of what has been important to them should be given some credence as a criterion.

Most of these comments concerning trainee ratings of the services they receive hold at the program and agency levels as well. There is clearly a need to explore the potential of trainee ratings as a proximate criterion.

There is also a more general point that can be made about finding out what the clients of training programs want. Manpower training programs have cut into the supply and demand situation in such a way that they typically have not taken advantage of and worked with the "natural" processes at work on either side of the equation. The two natural mechanisms determining the allocation of people among jobs are (a) an individual's desire for a specific kind of job and (b) the availability of

certain types of jobs. An individual may actively search out a particular job, willingly and consciously taking on the hardships and risks involved, or he may leave the active role to employers and rather passively accept the first thing available as long as it meets his minimal wage requirements. Essentially this is the difference between a vocation and a job. Most people fall somehwere between the two extremes. Manpower training programs have in the past not really utilized either mechanism. They gave their trainees neither the active freedom of responsible choice nor the passive security of training for a job of certain availability. The move to on-the-job training, with programs involving the private sector like JOBS, takes advantage of the latter mechanism. There is also room and reason to take advantage of the former mechanism by presenting to the trainee more alternatives and aiding him to make his own responsible choices in terms of his abilities and interests. The way this can be done is by taking seriously trainee statements as to what they want out of work and life and helping them to make a realisitc appraisal of the steps necesaary to get there. If the training project sets the direction and goals for its clients, as is often the case, responsibility for success or failure can be attributed to the training project. If the client sets his own direction and goals, after receiving relevant information as to their feasibility from project personnel, the burden of responsibility is shifted to his shoulders where it belongs if projects are to avoid the heavy costs of paternalism. More extensive use of trainee evaluations and opinions at all levels is a first step in this direction.

Program Level. Decisions at the program level concern projects. They are of two major types: those concerned with improving projects and those concerned with initiating, continuing, and terminating projects. The first involves improvement, understanding, and salesmanship functions of evaluation; the second, verification functions. A program also engages in vindication activities where information is directed upward and outward (to, for instance, the news media) rather than downward to projects. And it may engage in activities designed to improve performance or reduce anxiety at the program level itself.

Since the program level is in a position to look across all projects and compare them and since it exerts considerable leverage on

projects through budgetary control, it is the ideal location for evaluative research that has as its objective major improvements in project performance. Evaluation at the project level can lead to improvements, but these tend to be minor ones mainly involving adaptations to local conditions. The program level, on the other hand, can tackle the larger, more difficult, more expensive evaluative research that attempts to get results applicable across all projects.

The distinctive difficulty in doing evaluative research of this type is the problem of comparability of independent and intervening variables. The independent variable problem, that of adequately specifying the nature of the services delivered, was discussed in an earlier section. Similar difficulties are involved in trying to measure or control for the effects of intervening variables, in this case, local area characteristics. These difficulties affect any study that attempts to evaluate the relative effectiveness of operating projects.

It is possible, however, to design an evaluation strategy for the program level that largely avoids these difficulties, and meets most of the needs of the program level. This strategy consists of three linked parts: (a) a comparison of those projects believed to be the most successful and those believed to be the least successful to verify relative success and to identify manipulatable variables that seem to account for success or failure; (b) rigorous experimental projects evaluating treatment components and other project variables that seem related to success on the basis of the best-worst comparisons or other evidence; (c) the evaluation within routinely operating projects of those treatment components that appear promising as a result of the experimental tests.

The best-worst comparisons would verify the success rates of projects presumed to be good and bad and in so doing should work with several proximate measures of success. Main attention, however, would be devoted to project variables thought to affect success rates. Thus best-worst comparisons directly confront the problem of treatment comparability in a way that surveys of single project effectiveness or of overall program effectiveness, like the Dunlop and Associates study of Out-of-School Neighborhood Youth Corps, do not. Best-worst comparisons turn the heterogeniety of projects to good advantage, since variability of an independent variable is desirable in accounting for change in a dependent variable.

Such comparisons would focus more on manipulatable variables than on ones not subject to change by planned intervention. The intention of such comparative studies is to learn how to improve projects. The main focus should be on treatment or project characteristics and not on trainee characteristics. Correlations between trainee characteristics and success rates are probably the most common statistics reported in manpower training evaluations -- and the least useful in and of themselves. Such correlations, if causality can be demonstrated as in the Underhill study [23], have a predictive function and are useful in increasing understanding. Unless they are used in the context of a study that also examines project characteristics, however, they are of no help in project improvement nor in other decisions.

As an aid in studies that are primarily concerned with the relationship between treatment characteristics and success rate, data on client characteristics can be useful in two ways. Information on trainee characteristics that are causes of success tells us what variables need to be controlled, in the event that it is not possible to make a random assignment of clients, in examining the relationship between treatment characteristics and success. This is the thrust of Underhill's work.

In the second place, information on target population characteristics is essential if we wish to tailor projects so that they better meet the needs of particular groups of clients. But in both cases, experimental methods provide a better handle on the problem than do statistical methods. While heterogeneous projects can be used as "natural experiments," one is far better off with real experiments that permit only a small number of treatment variables to vary and randomly assign clients from the target population or, if the interest is in specific tailoring of treatments, from defined segments of the target population. If one treatment regimen is being compared with another (the comparison of interest for improvement purposes) rather than with no treatment (the comparison of interest for verification purposes) there is no need to deny service to certain clients, one of the most common objections to experimental evaluation.

In summary, the bulk of effort and expense in best-worst comparisons should be devoted to the identification and measurement of relevant treatment variables and not to the investigation of trainee characteristics. To develop a solid evaluative research program that will lead to

the improvement of services delivered to clients, the best-worst comparisons have to be followed up by rigorous experimental research. This suggests that the sophisticated control of trainee characteristics, as suggested in the Underhill work [24], is probably not justifiable in cost/effectiveness terms given the experimental alternative, an alternative that is clearly available under the amended Economic Opportunity Act. For the best-worst comparisons one can be content with a soft research design if they are followed up with experimental work. Even if they are not followed up with experimental work, it seems difficult to justify the expense of a large survey and sophisticated statistical design concentrating on trainee characteristics so long as the variables subject to improvement are neither measurable nor controllable for the period of the survey.

The best-worst comparison provides a first step towards identifying relevant treatment variables. It confronts the problem of making comparisons across projects meaningful, but it has to wait on additional experimental work to solve the problem. If the meaningfulness of the best-worst comparisons is to be maximized, it would be important for the same people to collect data across all projects.

If, for instance, staff rapport with clients is to be investigated as a possible factor contributing to success and each project is examined by a different research organization, or even by different researchers within the same organization, it would be incautious to put much faith in the inter-project reliability of such measures. We do not have tried and true ways of measuring treatment variables, and the more difficult the measurement problem the greater the need to have the same persons taking the measurements if there is to be any basis for comparison. If the task of making all comparisons, say of 10 good and 10 bad projects, is too large for one research organization, the task should be divided by clusters of variables and not by projects. One group of researchers can examine success rates; another, community relations; another, variables related to staff; others, sets of treatment components; etc. The degree of coordination needed between such groups should not be an insurmountable obstacle.*

*Perhaps the most difficult problem it raises is on publication credit, since no group alone has publishable findings. This problem should not be underestimated, particularly when it is possible to pose a solution. The responsibility for investigating success rates could be divided so that each group collects data on one proximate criterion variable. This means that each group can have its own set of correlations to report.

The best-worst comparisons serve three useful purposes: (a) they provide information helpful in decisions concerning the continuation and termination of projects. The program level cannot afford to do rigorous evaluative research on each and every project under its jurisdication. Judgments about the value of projects will instead be based on information from a variety of sources, from "inspector-general" visits to the project, from subjective appraisal, from data provided by the project, including cost figures, etc. Rigorous evaluation of those projects which look the weakest provides a check on these information sources and on the judgments derived from them. It can also provide, if results are obtained in time, some solid evidence upon which to base the important decision of whether or not to terminate a project. (b) The best-worst comparisons will provide evidence and leads as to what variables account for success and failure, and thus provide some material for immediate recommendations to the weaker projects as well as useful suggestions for more rigorous investigation. (c) Findings on the performance of the best projects provides data that is more useful for vindication purposes than the findings of a study of overall effectiveness. It casts the best but nevertheless honest light on the program's efforts. It shows what kinds of results the program can achieve and, if this is done within the context of an evaluative research program directed at improvement, the program's objective of bringing all projects up to this level is credible.

Studies of the overall effectiveness of a program -- the verification function -- should be left to the director level. The objectives of programs overlap and this overlap should be taken into account in estimating the effectiveness of a program. Only the director level can do this. It is at the director level that decisions are made about allocation among programs and it is only at this level that information on overall effectiveness contributes directly to the decisionmaking process. The director has an interest in verifying the actual effectiveness of a program. The program, if it looks at its own overall effectiveness, will have an interest in vindicating its performance. Data on overall effectiveness has no utility for decisions within the program level, thus any such data collected will tend to be dressed up for export purposes. Data on the performance of the best projects under the program's jurisdiction, on the other hand, has internal as well as external purposes and it is

more likely that overall effectiveness data can make a good argument for program continuation without being dressed up.

The rigorous experimental evaluation of treatment variables, not of whole projects, should be the main focus of evaluative research at the program level. Such work seems to offer the best hope for long-run improvement of project performance as suggested earlier in this paper and as persuasively argued by Mann [25].

The salesmanship function of evaluative research and its main manifestation, the demonstration project, are of most interest at the program level. The demonstration project is typically seen as a way of introducing a new treatment or concept or project design. As a means to facilitate innovation, the demonstration project is probably over-rated and, in the manpower training area, almost certainly over-used.

It is difficult for a demonstration project to live up to its promise of combining creative product development with scientific evaluation of product performance. In the first place, the special situation of a demonstration project raises serious doubts about the transferability of the methods of projects working in a routine fashion. A demonstration project is quite likely to have better quality staff, staff that pays more attention to clients and has more esprit de corps because of the Hawthorne effect, more administrative freedom than the average project, and often more money than could be allocated routinely to equivalent projects. If successful, is it the treatments tested or is it the special effects that are accounting for the success? Second, to be successful as a selling device, it is not sufficient to write up and make available a report on the results of the demonstration project, as is most often done. Instead a rather elaborate campaign involving publicity, on-site visits, pep talks, consultation, etc. has to be planned around the demonstration project if project administrators are to be persuaded to implement the new concepts [26]. Third, demonstration projects are almost never planned in a way that provides any cumulative impact on our general knowledge of treatment effects.

We would suggest that demonstration projects only be conducted under rare and carefully planned circumstances. There are only so many practitioners who are potential implementors and they do not have the time to be careful and selective consumers of demonstration projects. When

offered a large variety of demonstrations, they are quite likely to ignore all of them. It is much better to come up with one or two good solid demonstration projects, accompanied by a well-planned implementation strategy, than with fifty different ones of varying quality.

There are other kinds of "naturally occurring demonstrations" that should by all means be utilized. For instance, in doing best-worst comparisons, the program level may find some unique features that seem to account partially for the success of one of the better projects. An active attempt should then be made to encourage the adoption of this technique by other projects.

Finally, a firm distinction should be made between demonstration projects and experimental projects. Experimental work is aimed at improvement and increased understanding; demonstration projects are aimed at salesmanship and, sometimes, verification. They have quite different functions and a confusion of the two imperils the appropriate use of both. There is currently a crying need for more rigorous experimental work on manpower training and a surfeit of demonstration projects of dubious value.

There are two other kinds of evaluative research that the program level should be involved in. Both these overlap with interests at the director level. One is the determination of unmet needs. Such work involves attention to possible gaps in programs. An example is the recognition, occurring after the first year of MDTA institutional programs, that remedial education should be part of the curriculum. The program level should also investigate the need to expand the program to make it available to more of the same target population or to cover new target populations.

Finally, the program level should pay some attention to the possible side-effects, both beneficial and harmful, of its activities. An examination of side-effects could easily be incorporated into the best-worst comparisons. The search for adverse side-effects raises the same threatening self-incrimination potential that verification of overall effectiveness does, and prime responsibility for worrying about adverse consequences might best be exercised at the director level. The program level is closer to the problem, however, and is in a better position to develop hypotheses at an early stage about possible harmful side-effects. Also they are in a better position to plan the research on harmful side-

effects in such a way as to be helpful in developing recommendations for ameliorative measures. Work on beneficial side-effects, particularly local and proximate ones, provides the program level with another source of findings useful for vindication and public relations.

Director Level. Evaluation at the director level can contribute to two main kinds of internal decisions: (a) allocation among existing programs and (b) decisions to institute new programs. Both these uses of evaluation derive from the director's responsibility for the achievement of the agency's broad objectives. There is another, and probably the most important, use of evaluation at this level, not related to internal decision-making, and that is the vindication use. The final responsibility for justifying programs and projects rests with the director, and it is at this level that the verification function of evaluative research becomes an important aspect of vindication. At other levels the two can be separated, but the director's office needs relatively sophisticated, hard-nosed evidence on overall effectiveness for submissions to the Budget Bureau and, increasingly, to Congress.

Allocation among programs is also intimately linked with vindication since budget requests are made, and must be justified, on a program by program basis. Thus the director level needs to concern itself with the overall effectiveness of each program in order to justify both its allocations and its total budget request.

If we focus on the matter of allocation, it can be seen that the director level not only inherits the problems of feasibility and independent variable comparability from the project and program levels, but in addition has to cope with the problem of the comparability -- or incomparability -- of dependent variables. Programs have different objectives and allocation decisions make an implicit or explicit judgment about the comparative value of these differing objectives. While this need to compare programs having differing objectives should be of concern to evaluators, it remains impossible to put most such comparisons on an objective, scientific basis. Allocation among programs with differing objectives will be determined by judgment and bargaining. Analysis, including evaluative research, can be a useful aid in both judgment and bargaining, but it has to take its place alongisde other aids such as wisdom, experience, and political clout.

There is general agreement on the above point and there would be little need to discuss the matter further if it were not for the fact that one of the most important benefits of manpower training programs happens to be a neat, quantified variable: increase in earnings. Economists are likely to be seduced by this into hoping that it will be possible to make inter-program comparisons on the basis of a wholly rational scorecard. Such comparisons are only possible when some of the objectives of two or more activities are identical and all the objectives are in some way, even if grossly, measurable. Neither condition is likely to hold for manpower training programs the way they are currently set up. Benefits as measured by increased earnings can be roughly compared, but one way in which the objectives of programs differ is in the kinds of target populations served. Seemingly comparable monetary benefits to clients are in fact noncomparable because the clients differ -- and we are not even sure of the relevant ways in which they differ. A ramdom assignment of clients to two programs would solve it, but then we are talking of comparing experimental programs and that is a different matter. A second problem with cost-benefit comparisons is that not all benefits of programs are measurable, in even the grossest way, and some of these nonmeasurable benefits are important [27].

This is not to say that cost-benefit studies of manpower training programs are useless. They are both useful and meaningful, but their utility lies primarily in vindicating programs and not in making choices among programs [28] nor in deciding between a training program and transfer payments. If a cost-benefit study shows that discounted future earnings are greater than costs it provides some rationale for public investment. It is a useful device for justifying a program, particularly a program with high unit costs like Job Corps. But a cost-benefit ratio of less than one does not tell us in and of itself whether the program is worth supporting, since no one would argue that the government should sponsor every activity that has a cost-benefit ratio less than one. On the other hand, as Levine has pointed out [29], a cost-benefit ratio greater than one for a training program does not provide a foolproof argument to replace the training program with transfer payments since again we are not dealing with true alternatives. Training programs have a different, and generally preferred, objective and that is to raise people above the poverty line through their own earning power.

Since the introduction of program budgeting there has been a continuing debate about the role of analysis in government. Arguments are often phrased in an either/or fashion -- bargaining vs. analysis -- as if we had to pledge allegiance in perpetuity to one or the other. This is not the case. The relevant questions for those interested in the potential contributions of analysis to improved decision-making are (a) under what conditions should a decision be based primarily on analysis? and (b) under what conditions can analysis be useful in a decisionmaking process largely determined by bargaining or politics?

Cost-benefit studies of manpower programs fall in the latter category: they are an aid in a political decisionmaking process rather than a basis for rational decisionmaking. They can also be a useful device for persuading programs to attempt to improve performance. Again this is a political use. It is actually the other side of the vindication coin, since the argument to program personnel would point out that the program cannot be justified to the Budget Bureau and Congress unless it begins to show a better return on investment.

It should be noted at this point that the relationship between the director level and programs is not identical to that between the program level and projects. While the director has leverage vis-a-vis the programs and will be interested in using this leverage to improve programs, the director level is too distant from the point of delivery of services to concern itself usefully with the means of improvement. Its concern will be with the need for improvement, leaving the programs discretion in deciding upon the means of improvement. In the program-project relationship, on the other hand, much of the communication downwards should be focussed on the means of improvement.

Cost-benefit studies vindicate a program by showing that it is at least not an unreasonable thing to do from a public investment point of view. There are other ways of vindicating programs, however, and these other ways can often be of more help in allocation decisions than cost-benefit studies. These other methods focus on needs and on the adequacy of programs in meeting needs. There is a simple approach that looks at the performance of training programs and compares it with estimates of unemployment and underemployment in target areas. And there is a complex approach that attempts a broad gauge, long-term analysis of the inter-

relationships of relevant institutions and includes a questioning of the basic assumptions of training programs.

We will discuss the simple approach first. It is a matter of verification, verifying that programs are meeting the needs they are supposed to meet. Some of the data needed for this can be collected by routine methods at the project level, as previously discussed, and transmitted, after aggregation at the regional level, directly to the director level. The data needed -- numbers entering and completing projects, simple demographic data on trainees to show that projects are serving appropriate target populations, costs, and, if possible, numbers placed in jobs -- are easily gathered at the project level without undue paperwork and are also useful at the project level.

In addition to these simple figures on overall performance routinely collected, there is a need to conduct sample surveys on success rates to complete the verification job. In some projects, job placement data is easy to collect, in others it is not. If it is not easy to collect, the data are likely to be unreliable and not worth having. Besides, job placement is not even a good proximate measure of success, let alone a good ultimate measure. Job retention and increased earnings are more valid as success measures and these should be collected at the director level, using representative sampling among projects. This will provide an average success rate, which will be useful for comparison with success rates of good and bad projects examined at the program level.

These sample surveys should collect data on participants' opinions of projects, as well as on jobs held and earnings, both to meet legislative requirements and for examination of such ratings as proximate measures of success. They should also collect data on demographic characteristics, to be related to success rates and compared with the demographic profiles of the areas in which the projects operate, as a means not of explaining success rates but of identifying unmet needs. If a project has collected "before" measures on other variables that could be used as criteria of success, such as verbal or cognitive ability, attitudes, etc., it might be useful to get "after" measures on these.

As well as a concern with the verifications of short-run perfcr-- mance, the director level has the main responsibility for worrying about the attainment of long-run goals. One question in need of attention here

is the relationship between proximate and ultimate criteria of success. This is the key predictive relationship of relevance to manpower training programs. Work on this problem should include the planning of long-term follow-ups, and other ways of tackling the problem. Statistical relationships that hold for the population in general, as for instance, the relationship between education and lifetime earnings [30] can increase our confidence in certain proximate measures. It would also be useful to explore the potential of retrospective studies for identifying proximate criteria. Such retrospective studies would fall under the heading of "deviant case" analyses. For instance, the probability of occupational "success" for a Negro raised in an urban ghetto is low, but a minority do make it. An interview study of those who have made it might provide some useful clues with respect to proximate criteria useful in predicting long-term success, as well as, perhaps, some clues about causal variables that could be useful in program planning.

We previously referred to the complex approach to examining the adequacy of programs in meeting needs. This approach is concerned with the analysis of programs and alternatives in the broadest sense. Studies of this sort would attempt to determine the total impact of all agency activities, or of all agency activities in particular categories, for instance, the overall impact of the war on poverty or the overall impact of all training and job placement activities, and the potential impact of conceivable alternatives. They differ from other verification studies in (a) examining the effects of all relevant activities and institutions, not just agency or government supported activities, and (b) in questioning the basic assumptions on which agency activities are based [31], this being done in the context of looking for institutional alternatives that might perform better. As a corollary of examining institutional interrelationships and alternatives, such studies would pay attention to unmet needs and to the unintended side-effects -- short and long term, negative and positive -- of existing programs.

In the manpower training area, the question would be: Given the current poverty population, skill structure, and occupational structure, and projections into the future of the three, what is the significance of existing training programs and of possible alternatives? This is the most difficult kind of evaluation to do well but, due to its capacity to generate

new lines of thinking about alternatives, may have the greatest potential impact. Doubts about the efficacy of the current organization of training efforts are sufficiently widespread to make it worthwhile to explore major alternatives and supplements, such as programs directed to job development and man-job matching systems, including worker re-location schemes; plans, like the recently proposed California system, resting upon more individualized services to clients and incentives for "job agents" based on the post-training employment records of their clients [32]; utilization of existing systems -- public vocational education, private training institutes, educational and training activities within prisons, etc. -- in new ways; providing government help for the upgrading of working class and lower middle class whites in industry so that there are more "career ladders" open for blacks brought into entry level jobs and some economic reasons for management to be interested in their advancement.

Such studies would use quantitative data where possible and qualitative data where necessary. The unit of analysis for such studies should be either the whole country or a particular geographic area [33]. This permits an examination of the interrelationships between OEO programs and other institutions and estimates of the indirect effects of OEO activities and of the kinds and degree of unmet needs. All other evaluative research suggested in this paper, and almost all of the evaluative research that has been supported by OEO, uses the program or project as the basic unit of analysis. From that perspective it is almost impossible to question the basic assumptions underlying the programs. There is a need for such questioning at the director level, and the impact survey using a community as the unit of analysis, along with national statistics, seems to offer the most potential for getting sufficiently outside the framework of current operations to be able to examine their contribution to long-range goals [34].

NOTES

[1]Edward A. Suchman, Evaluative Research, New York: Russell Sage Foundation, 1967, p. 31.

[2]The distinction between vindication and verification is from F. Stuart Chapin, Experimental Designs in Sociologicll Research, New York: Harper and Bros., 1947, p. 177.

[3]Robert K. Merton, Social Theory and Social Structure, Glencoe, Illinois: Free Press, 1957, p. 199.

[4]Peter Sainsbury, "Research Methods in Evaluation," in Richard H. Williams and Lucy D. Ozarin, eds., Community Mental Health, San Francisco: Jossey-Bass, Inc., 1968, p. 214.

[5]Bernard G. Greenberg and Berwyn F. Mattison, "The Whys and Wherefores of Program Evaluation," Canadian Journal of Public Health, Vol. 46 (1955), p. 298, as cited in Suchman, op. cit., p. 92.

[6]K. K. Mathen, "Matching in Comparative Studies in Public Health," Indian Journal of Public Health, Vol. 7 (1963), pp. 161-169, as cited in Suchman, op. cit., p. 105.

[7]John Mann, "Technical and Social Difficulties in the Conduct of Evaluative Research," Appendix A, in Changing Human Behavior, New York: Charles Scribner's Sons, 1965, pp. 177-189. See also, in the manpower training area, Martin Moed, Irwin Feifer, and Leonard P. R. Granick, "Viability of Program Experimentation in Services for Upgrading the Employability of the Culturally Disadvantaged Youth," paper presented at the 75th Annual American Psychological Association Convention, Washington, D.C., September 4, 1967.

[8]Mann, "The Outcome of Evaluative Research," Appendix B in ibid, p. 204.

[9]Suchman, op. cit.

[10]Among others: David Wellman, "The Wrong Way to Find Jobs for Negroes," Trans-action, Vol. 5 (April 1968), pp. 9-18; Aaron Wildavsky, "The Empty-Head Blues: Black Rebellion and White Reaction," The Public Interest, No. 11 (Spring 1968) pp. 3-16.

[11]For instance, the likelihood of the New Careers program failing to provide anything but menial, dead-end jobs at minimum wages seems to be rather high unless the introduction of trainees into hospitals, schools, etc. is done with the whole-hearted support of the professionals in these institutions, including their support for the kind of re-structuring of tasks and qualifications in their fields that will open up new career ladders for those entering at the bottom with little formal education. New Careers programs have been introduced where this kind of professional support is lacking. In such cases, a wait-and-see-what-happens-at-the-end-of-the-training-program attitude has been taken, rather than a concern with the consequences of failure and an active attempt to lower its likelihood.

[12]See Herbert H. Hyann, Charles R. Wright, and Terence K. Hopkins, Applications of Methods of Evaluation: Four Studies of the Encampment for Citizenship, Berkeley: University of California Press, 1962, for a study that attempts to handle unanticipated consequences.

[13]For instance, see MFY's attempt to precisely describe a remedial education "delivery system." Despite the care exercised in this attempt, it is doubtful that they are dealing with replicable variables in this educational package. See Mobilization for Youth, Inc., Division of Employment Opportunities, "Eight Month Report to the Office of Manpower Policy, Evaluation and Research," December 16, 1965, to August 15, 1966, Section XI; and also their "Fourteen Month Report," pp. 100-145.

[14]Mann, op. cit., p. 210.

[15]Nathan L. Gage, "Paradigms for Research on Teaching," in Gage, ed., Handbook of Research on Teaching, Chicago: Rand McNally, 1963, pp. 113-120.

[16]Mann, op. cit., p. 210.

[17]For the kind of self-serving carelessness in data collection that one can expect, see Ida R. Hoos, Retraining the Work Force, Berkeley: University of California Press, 1967, pp. 208, 270.

[18]David Sudnow, "Interactional Considerations for the Evaluation of Manpower Training Programs," unpublished paper, June 1968.

[19]The data required to be collected by the Manpower Development and Training Act of 1962 is reputed to have remained in a warehouse until Garth Mangum took an interest in analysing it in 1964 for the Clark Subcommittee on Employment and Manpower.

[20]If the program level feels a need to run a close check on the performance of certain projects, then they would be better advised to take a lesson from industry and send "head office" personnel out to collect data on a routine basis at the local level. Herbert A. Simon et al., Centralization vs. Decentralization in Organizing the Controller's Department, New York: American Book-Stratford Press, 1954: cited in Harold L. Wilensky, Organizational Intelligence, New York: Basic Books, 1967, pp. 61-62.

[21]One study showed sixth-grade teachers given feedback from student ratings changed in direction of pupils' ideal teacher as measured by students in subsequent ratings. (.H.H. Remmers, "Rating Methods in Research on Teaching," in N.L. Gage, ed., Handbook of Research on Teaching, Chicago, Rand McNally, 1963, p. 367.) This is not an earth-shaking finding, but remember the scarcity of any kind of hard evidence on the efficacy of applied research.

[22]Remmers, ibid., pp. 367-368.

[23]Ralph Underhill, second phase of pilot study of poor youth conducted for the Office of Economic Opportunity by the National Opinion Research Center.

[24]Ibid.

[25]Mann, op. cit.

[26]Office of Manpower Policy, Evaluation and Research, U.S. Department of Labor, Putting Research, Experimental and Demonstration Findings to Use, pp. 31-35.

[27]For a more extensive discussion of the problems and possibilities of benefit-cost evaluation in manpower programs, see Thomas R. Glennan, Jr., Evaluating Federal Manpower Programs: Notes and Observations, RM-5743-OEO, Santa Monica: Rand Corporation, September 1969.

[28]As Robert A. Levine has implied in "Evaluation of Office of Economic Opportunity Programs -- A Progress Report," Proceedings of the American Statistical Association, 1966 Social Statistics Section, p. 343.

[29]Levine, loc. cit.

[30]Michael E. Borus, The Economic Effectiveness of Retraining the Unemployed, Research Report to the Federal Reserve Bank of Boston, Number 35, July 1966.

[31]Unpublished paper by Phyllis L. Cohan.

[32]Assembly Bill No. 1463, California Legislature, 1968 Regular Session.

[33]An example of this sort of work is provided by Neil H. Jacoby, U.S. Aid to Taiwan, New York: Frederick A. Pragger, 1967.

[34]Two approaches which move in this direction: Oakland's Partnership for Change, by J.M. Regal, June 1967 and The Castlemont Survey: Summary of Results, 1966, Department of Human Resources, City of Oakland, California; and John T. Doby, Design for a Comprehensive and Systematic Evaluation of the Community Action Programs Operating in Atlanta, Georgia, Emory University, 1966.