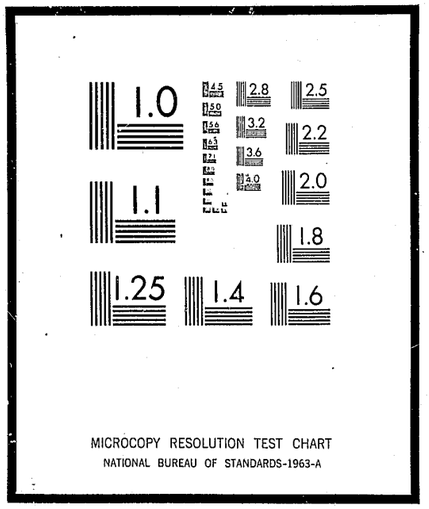

MEMORANDUM
RM-5743-OEO
SEPTEMBER 1969

EVALUATING FEDERAL MANPOWER PROGRAMS: NOTES AND OBSERVATIONS

Thomas K. Glennan, Jr.

The Research reported herein was performed under contract with the Office of Economic Opportunity, Executive Office of the President, Washington, D.C., 20506. The opinions expressed herein are those of the author and should not be construed as representing the opinions or policy of any agency of the United States Government.

The RAND Corporation
1700 MAIN ST • SANTA MONICA • CALIFORNIA • 90406

Published by The RAND Corporation

PREFACE

The RAND Corporation, with the sponsorship of the Office of Economic Opportunity, has carried out a number of studies intended to contribute to the improvement of the evaluation of manpower programs. This Memorandum, based on a critical examination of manpower program evaluations that have been carried out in the past as well as RAND's own experiments in program evaluation, considers the methodology of evaluation. But the use of evaluative data in the planning process depends only in part upon the capabilities of the methodologies of evaluation. The nature of the organizational pressures for and against evaluation and the decisionmaking process that might utilize evaluative data must also be considered. This Memorandum attempts to synthesize these factors. It is intended to clarify some of the problems that surround the use of program evaluation by Federal agencies, particularly in the area of manpower training.

The manuscript has profited from the comments of many people. Within RAND, the critical reviews of Anthony Pascal and Anne Summerfield were particularly valuable. An immense debt is owed to present and former members of the staff of the Office of Research, Plans, Programs and Evaluation of OEO, particularly Robert Levine, Walter Williams, and John Evans. The author, of course, is solely responsible for errors, omissions and the opinions expressed.

SUMMARY

The Office of Economic Opportunity has placed considerable emphasis on the evaluations of the programs funded under the Economic Opportunity Act. Although the major part of this evaluation effort has focused on project monitoring, a significant and controversial effort has been made to measure the impact of various programs in terms of the objectives established for them in the legislation or their administrative guidelines.

Impact evaluations of social action programs have had many shortcomings. In the case of manpower programs, the major problem has been finding a reference or control group with which to compare program participants. Because of this problem, many evaluations have lacked credibility and hence have been disregarded in the policymaking process.

Even if the control group problem did not exist however, the analysis of the data has varied from study to study rendering these studies incomparable and limiting their utility for comparing program outcomes. Much of this difficulty seems to be traceable to a failure to adequately and realistically specify program objectives. But a significant proportion can be traced to necessary but arbitrary assumptions that are made somewhat differently by each evaluator because of data availability or methodological bias.

It seems clear that the quality of impact evaluations could be improved somewhat if program managers had a greater interest in such evaluations. For a variety of reasons (both good and bad), such managers have been reluctant to have their programs evaluated. As a consequence, evaluations that have been carried out have not adequately taken into account the types of decisions that could be clarified with evaluative data. Evaluations have tended to become a weapon to be used in bureaucratic wars rather than a rich source of information to support detailed program design and funding decisions.

Since program evaluation remains largely in its infancy, the benefits to be derived from it remain to be demonstrated. This

Memorandum suggests several steps that should contribute to improving the usefulness of evaluative data. A careful examination (through actual use) of longitudinal study designs should be made. Similarly, imposition of a more strongly experimental structure on the initial operations of a new program or upon demonstration projects should be encouraged. Both of these steps would probably improve the validity of the results of evaluation and in some instances provide information that cannot be obtained by the currently used retrospective studies.

It would be very useful to establish a set of analytical conventions for carrying out benefit-cost studies of manpower programs that would ensure that the results of separately conducted studies would be as comparable as possible. A precedent exists in the so-called Green Book that guides such calculations for water resource projects.

Future evaluation efforts can be made more reliable and more simple if good information systems exist at the local project level. Currently, there are almost no examples of useful project information systems. National systems, without good local data, tend to have limited usefulness. A major effort should be made to develop and support such systems at the local level.

Perhaps the most important step to be taken in making outcome evaluation more useful however, is to bring the evaluator closer to the policymaker. Evaluations should be framed with important policy issues clearly in mind. Evaluation data should be so organized that they can be utilized to answer questions that were not thought of prior to data collection efforts. Evaluation should be a continuing activity.

The history of program evaluation does not provide clear evidence that these objectives are obtainable. But the history of policymaking in social action programs provides little evidence to suggest that good decisions can be made without such evaluation.

CONTENTS

I. INTRODUCTION

Five years have passed since the signing of the Economic Opportunity Act and an explicit declaration of a War on Poverty. Even more time has elapsed since the first specific social action programs were undertaken in an effort to help the poor or ameliorate the adverse consequences of the workings of our economic and social system. Much is known about the inputs to these programs. We know how much has been spent. We have a fairly good idea of how many people have participated in these programs and the characteristics of these people. We know remarkably little about the effects of these programs. Indeed, in many instances, we do not know or cannot agree about the dimensions by which to measure these effects.

The experience of the present and past social action programs should be the best source of information to guide our future programs. Programs that are "working" should be sustained or expanded. Programs that are "not working" should either be curtailed or restructured. Within a program, the most effective features should be emphasized and the least effective discarded or modified. The most effective projects should be expanded, the least effective cut back or reoriented. The performance of new programs, suggested by research or demonstration activities, should be compared with that of existing programs.*

In fact, few systematic efforts to extract information from existing programs and demonstration activities have been made. Those that have been made have generally had severe conceptual and methodological shortcomings. As a result, the decisions about program design and upon relative funding of these programs have usually been based upon hunches, anecdotal evidence, and political bargaining. Perhaps this is the best that could have been expected. Certainly a well-defined and reliable scheme for extracting timely information

*The term "program" in this Memorandum is used to designate a collection of local projects that are developed and managed according to a set of guidelines mandated by the Federal government. The Job Corps or the Neighborhood Youth Corps are examples of manpower programs.

on program effects did not and does not exist. Evaluation of social action programs is an art, and a not very well developed one at that.

It is interesting and useful to speculate on the reasons for the failure to pursue evaluation efforts more vigorously in the earlier days of the War on Poverty. Clearly, the initial efforts of OEO were, and had to be focused on, initiating a number of large and ill-defined programs. It was a time for innovators, activists, and operators not evaluators, and rightly so. In the first few years the programs were changing rapidly as the operators gained greater intuitive understanding of the possibilities and limitations of the program designs. Had evaluations been undertaken, they would have been largely irrelevant by the time they were completed.

From the beginning, OEO had an Office of Research, Plans, Programs and Evaluation (RPP/E) which had an ill-defined mandate to evaluate programs. In its initial years, the evaluation function of the Office of RPP/E was lodged in the Programming Division. In large part, the evaluations that were performed were carried out by the programs themselves although on occasions the Office of RPP/E took a strong lead in initiating particular studies. The Office did place considerable emphasis on developing information systems, anticipating that after a few years, when program operations had settled down, these systems would provide information that would support studies of program impact. In retrospect, this may have been a mistake. The information systems, developed without much guidance from "specialists" in evaluation, have failed to provide adequate and reliable information for studying the effectiveness of the programs.

Since most of the evaluation work was carried out by the programs themselves, it was natural that the evaluators focused on gathering information that would support program improvements. They sought out projects that seemed to be functioning smoothly, were using innovative techniques, or were experiencing great difficulty. The insights gained, usually in quite informal ways, were used to guide program operators in making changes in guidelines, in seeking new local sponsors or in justifying the program to Congress and the public.

With one or two exceptions, the analysts did not question the existence of any given program or whether the objectives of their program could be better achieved by other existing or potential programs. If they had, it is unlikely that the program operators would have chosen to continue supporting such analytical efforts.

This is not intended as a criticism. It is unrealistic and probably undesirable to ask a program organization to question its own existence. It is even more unrealistic to ask it to do so in its initial years. But it is not unrealistic to ask that someone in the government attempt to determine the relative effectiveness of the multitudes of Federal programs and make decisions on which ones to enlarge or cut back or to specify where totally new approaches are needed. Some individuals within the Office of RPP/E felt that this function in OEO should be strengthened and that RPP/E had the mandate to do so.

In the fall of 1966, a number of evaluations of program effectiveness were initiated by RPP/E. In the course of attempts to carry out these analyses, it became clear that the kind of information required to support decisions about which programs should be continued and expanded and which should be cut back or changed was not being generated. The next summer, a separate division of RPP/E was set up and, after several months, procedures dividing evaluation responsibilities between RPP/E and the programs themselves were developed. The OEO Instruction setting out these procedures suggested that there were three kinds of evaluations:

> Evaluations are categorized into three major types. The <u>first</u> is the overall assessment of program impact and effectiveness where the emphasis is on determining the extent to which programs are successful in achieving basic objectives. The <u>second</u> is the evaluation of the relative effectiveness of different program strategies and variables where the emphasis is on determining which of the alternative techniques for carrying out a program are most productive. The <u>third</u> is the evaluation of individual projects where the emphasis is on assessing managerial and operational efficiency.*

*OEO Instruction Number 72-8, March 6, 1968.

The project monitoring or "Type III" evaluation obviously should be the responsibility of the program manager. This function provides him with information needed to enable him to carry out his day to day management tasks. "Type II" evaluations are intended to support improvements in overall program effectiveness by identifying superior project designs, curricula, or types of project personnel. This information can be used to modify program guidelines or to suggest better procedures to project directors. Because information needed to structure such evaluations should be available at the program level and because the resulting information will be used by program managers, responsibility for Type II evaluations should also rest with the program manager.

Responsibility for overall impact or "Type I" evaluations is assigned to RPP/E. Type I evaluations are intended to help determine the relative impact or effectiveness of national programs as a (partial) basis for allocating resources to programs. A minimum of one percent of program funds are to be set aside for evaluation, with one-sixth of one percent being used by RPP/E for Type I evaluation. Although RPP/E has not yet completed a sufficient number of evaluations to support a final judgment, it appears that the establishment of the evaluation division represents an important step toward a more systematic examination of program experiences as a basis for program planning.

But the organizational history just discussed should not be cited as the sole explanation for the failure to mount more systematic evaluation efforts. The fact is that evaluation in practice falls far short of the ideal. It is easy to say that an agency should determine the impact of its programs. It is extraordinarily difficult to do so. Surely a part of the reason that more systematic impact evaluations have not been mounted is the lack of confidence that they can be mounted. The quantification of program outcomes and the meansurement of these outcomes pose significant conceptual and practical problems. Members of poverty populations are increasingly hard to survey.*

*In large part, this is the result of the intensive surveying that has already occurred in ghetto areas. Growing militancy among blacks has also increased the residence to being interviewed.

The impact of many of the programs is expected to be felt only over a period of years.

This Memorandum deals with evaluation and its potential use in the planning process. In particular, the focus is on the use of evaluation in manpower programs. Conceptually, this is one of the easiest areas of the War on Poverty in which to do evaluation. The purpose of manpower programs is to help people obtain better jobs, or maybe just any job. The increase in a man's or a woman's income as a result of participating in a program would seem a pretty good (even if incomplete) measure of the program outcome. Moreover, there is a sizable literature in economics dealing with the value of training and education that provides the theoretical underpinnings for studies to determine the benefits and costs of training. Despite these favorable factors, none of the overall impact evaluations that have been done to date should serve as a basis for planning future program activities. The few overall impact evaluations that have been completed are characterized by the use of very poor data and inconsistent analytical assumptions. This Memorandum will suggest ways in which evaluations can be made more relevant and useful. In Section II a benefit-cost framework for evaluation is developed. Section III examines several methodological problems associated with and limiting the quality of program evaluations. Section IV relates program evaluation efforts to the planning process, placing particular emphasis on whether straightforward impact (Type I) evaluations can constitute a useful input to this process. Conclusions and suggestions for potential program evaluators are contained in Section V.

This Memorandum treats the role of program evaluation in planning. Its tone will often seem to imply that program funding levels and program designs should be based solely upon evaluation data. This clearly is not and cannot be the case. Planning decisions are the result of a complex bargaining process. The outcome of such a bargaining process will reflect many factors, only one of which is information concerning past program performance. This is as it should be. Even the most sophisticated evaluations provide but crude guides to action. As will be seen, they consider only a part of the program

outcomes. They must usually utilize less than adequate data. They ignore many of the factors that must go into decisions on program funding levels and designs. Thus, program evaluations must be viewed as only one of a number of inputs to the planning process.

But two points need to be made. It is my judgment that program evaluations can be improved and, if improved, should play a larger role in the planning process. Second, the process of carrying out evaluation projects is likely to have a useful effect upon program planners. It can force a more careful examination of program objectives, as well as provide clues about how to improve program operations.

II. BENEFIT-COST EVALUATION OF MANPOWER PROGRAMS

In this section, a number of issues concerning the measurement and interpretation of benefits and costs are considered. For the moment, it is assumed that the major purpose for carrying out benefit-cost evaluations is to support the allocation of resources among a group of national manpower programs. A subsidiary purpose may be the justification of requests for additional funds to be utilized by manpower programs. Evaluations carried out for this purpose fit into the category earlier referred to as Type I evaluations.

If all programs have exactly the same objectives, it is fairly simple, conceptually, to specify the questions that evaluations should answer. Suppose, for example, that the sole objective of all manpower programs is to increase the national output. If this is the case, the evaluations should determine which program is providing the greatest increase in national output per dollar spent.

To accomplish this, the economist utilizes a form of analysis called benefit-cost analysis which attempts to support judgments concerning the economic efficiency of a program. The effects of the program must be translated into increments in national or collective output. This increment in output is then compared with the costs. The program producing the greatest increment of output per dollar is the most efficient in the sense that the increase in national output per dollar of input is greater than that of all other programs. Presumably, if resources are shifted from other programs to this program, total output will be increased.

Most program evaluations have attempted to obtain a measure of total program benefits or perhaps an average benefit per trainee. These evaluations have not focused on the problem of predicting the effects of an increase or decrease in program funding. Only under rather exceptional circumstances could measures of total or average benefits and costs be used to predict the effects of expanding or contracting a program.

The relevant benefit-cost ratios are those associated with marginal increments (or decrements) in the funding of these alternatives. Consider two programs, Program X and Program Y. Suppose a tentative decision has been made to add $50 million to these two programs and the problem is to choose whether to add it to X or to Y. The problem is not to determine whether the current benefit-cost ratio is higher for one or the other program but to estimate the benefit associated with adding resources to one or the other program. The problem is illustrated in Fig. 1 where the relation between benefits and costs for a program is portrayed. Suppose the program is currently operating at a level represented by cost C and benefits B. The average benefit-cost ratio is given by the slope of the line segment OA or $\frac{OB}{OC}$. If the decision is whether or not to add resources equal to CC', the relevant (marginal) benefit-cost ratio is that represented by the slope of AA' which is equal to $\frac{BB'}{CC'}$. The use of the average benefit-cost ratio could be quite misleading. This is illustrated in Fig. 2 in which two programs are compared. At current funding levels (C for both programs), Program X, represented by line OAA'D, has a higher average benefit-cost ratio than the program represented by Oaa'd. That is, $\frac{OB}{OC}$ is greater than $\frac{Ob}{OC}$. However, if the same increment of resources (CC') is added to both programs, the incremental or marginal benefit-cost ratio is greater for Y than for program X. We see that $\frac{bb'}{CC'}$ is greater than $\frac{BB'}{CC'}$; thus, the increment of resources should be allocated to program Y.*

There are a number of reasons for expecting marginal benefit-cost ratios to differ from average benefit-cost ratios. For example, as a youth program is enlarged it may reach deeper into the ranks of the disadvantaged. Such youths may require more services in order to achieve a given increment in income, or putting it another way, they may derive less benefit for a given quantity of services.

If a program requires administrative and professional personnel who are in scarce supply, increases in program activity levels should

*The optimal allocation of resources occurs when the marginal benefits associated with a small increase in expenditures are the

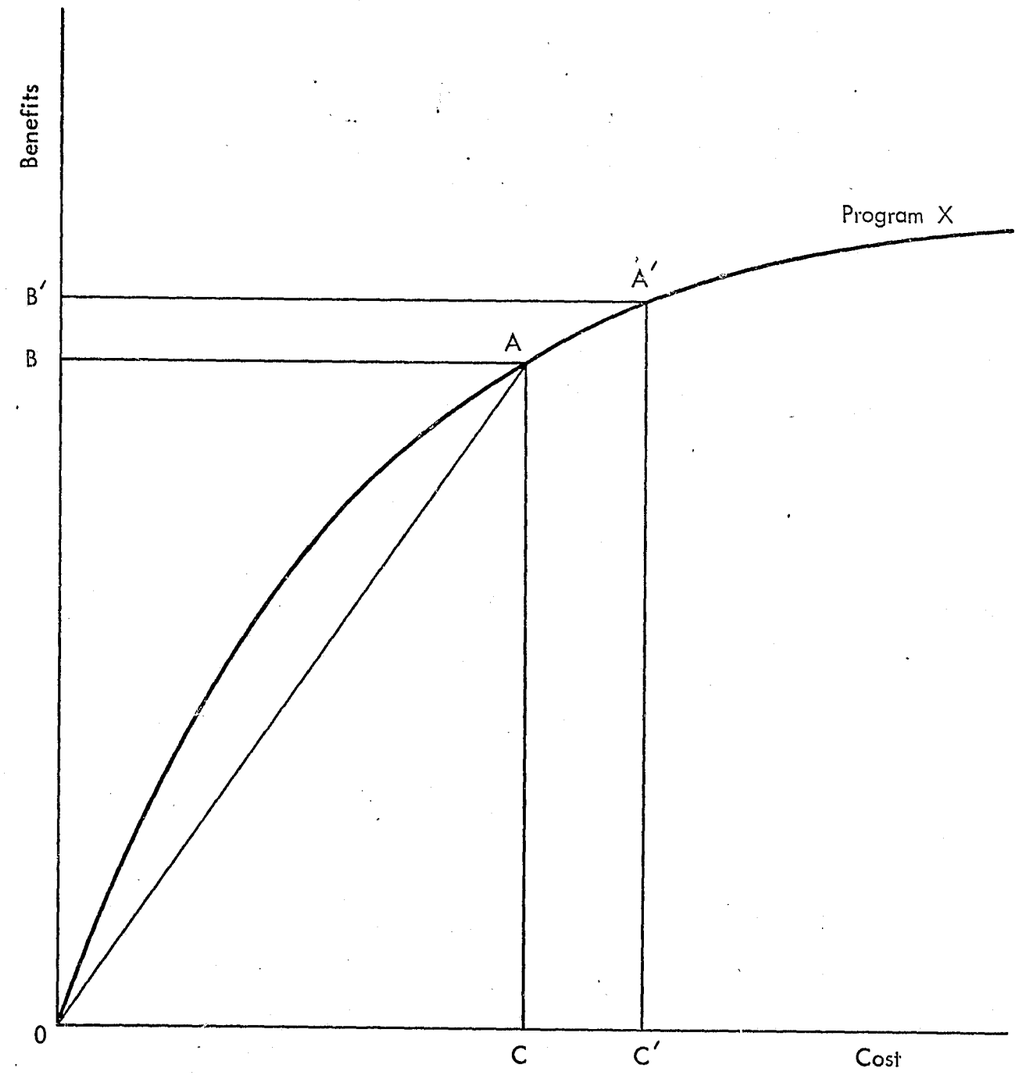

Fig. 1—A benefit-cost curve for a hypothetical program

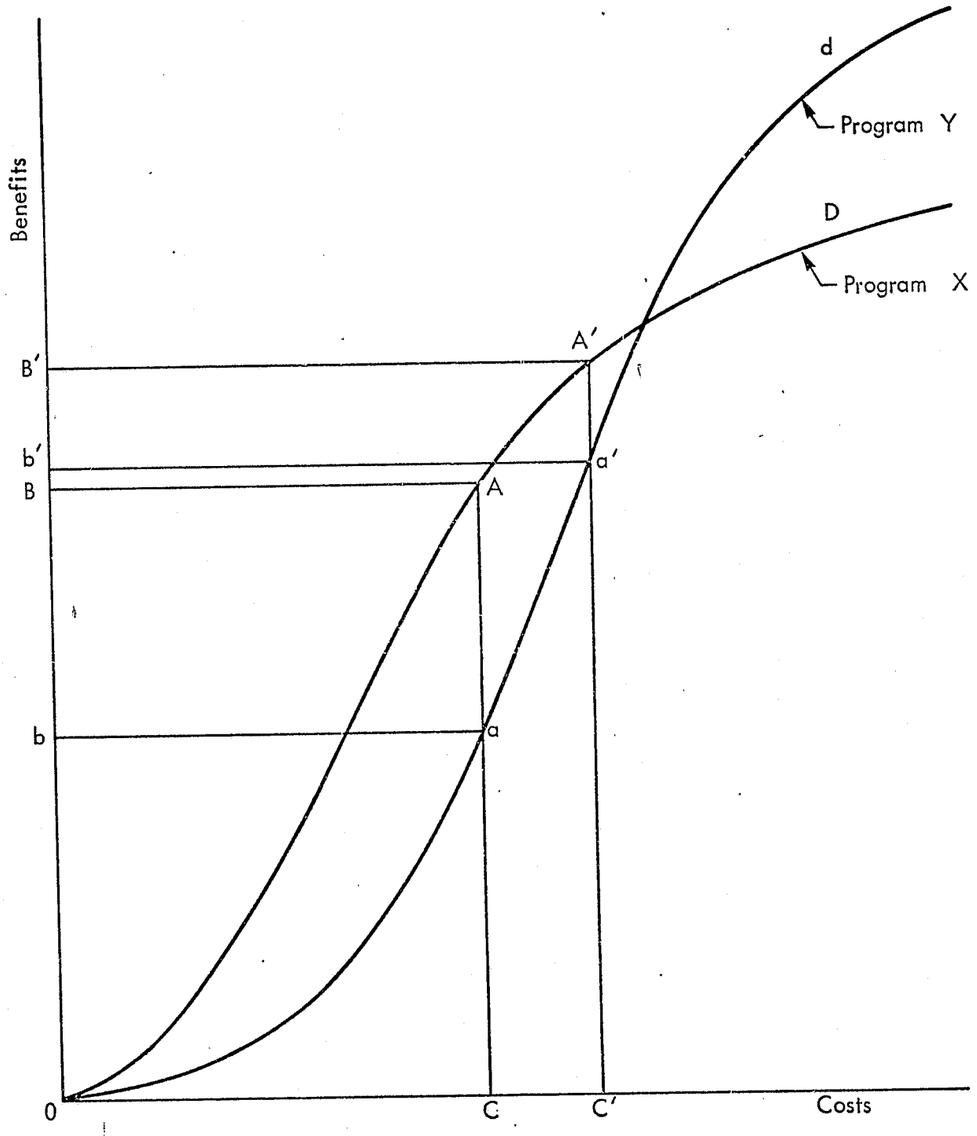

Fig. 2—Benefit-cost curves for two programs

be associated with either the hiring of lower quality personnel or the necessity of paying more for equivalent quality personnel. If the added personnel are of lower quality, the benefits to program enrollees associated with specific number of hours of professional services should decrease.

To my knowledge, no explicit attempts to estimate marginal benefit-cost ratios have been made for social action programs. However, the possibility that there will be decreasing marginal returns to additional investments in a program is frequently recognized. One reason for the steady movement toward on-the-job training programs rather than institutional training programs is the recognition that there is a shortage of high quality vocational teachers available in the nation's public school programs, whereas industry potentially has many skilled workers capable of teaching the necessary skills in their plants.

A crude analysis, closer to the type considered in this Memorandum, has been carried out by the Job Corps. Experience has shown that any increase in Job Corps enrollment would be likely to consist largely of 16-year olds. It appears to be difficult or impossible to attract increased numbers of older youths. The Job Corps appears to be less successful in dealing with 16-year olds than with older youth. They stay a shorter time. Because there are a large number of initial costs (for health services, clothing and testing, and so forth) the cost per month of Job Corps experience for a 16-year old youth is higher than for other age groups. Since expected benefits are thought to increase with the length of the Job Corps experience, the benefits accruing to 16-year olds are expected to be less. Hence, the marginal benefit-cost ratio for an increase in Job Corps activities should be lower than its average benefit-cost ratio.

This example suggests that attempts to measure the costs and benefits for different segments of the client population may have a

---

same for all programs. In the example shown in Fig. 2, the allocations shown are not optimal. In general, we have insufficient information to describe these curves accurately.

significant payoff for policy planning activities. Such data would support not only decisions concerning gross resource allocation among programs, but also decisions on program guidelines and target populations. If information on costs and benefits associated with providing services to different segments of the population were available, guidance could be provided which, if followed, would increase the benefits associated with the program while holding costs constant. In terms of Figs. 1 and 2, this would be equivalent to shifting the benefit versus cost curves upwards. Such evaluations would combine the functions of Type I impact evaluations and Type II evaluations aimed at program improvement. This combined evaluation might be called a Type I-plus evaluation and will be discussed in Section IV.

DISTRIBUTION OF COSTS AND BENEFITS AMONG ECONOMIC AND SOCIAL CLASSES

Statements about the economic efficiency of a social program do not take into account who pays for the program and who receives its benefits.* Clearly, in the Poverty Program the issue of who receives benefits is a crucial one. The introduction of these issues complicates benefit-cost analysis because of the necessity of weighing gains and losses of one group (the poor) against the gains and losses of another group (the non-poor).

When programs have objectives that go beyond simply maximizing the return on public investments irrespective of who receives the benefits, a simple benefit-cost ratio is an insufficient indicator of program outcome. Several alternative approaches to this problem have been suggested. Perhaps the most frequently advanced idea is the use of a system of weights reflecting the relative value society places on increases in the well-being of specific groups in society. For example, a given increase in income to very poor families might be considered more significant or valuable than a similar increase

*In many respects, my comments on treatment of distributional objectives parallels that of Rothenberg. See Jerome Rothenberg, Economic Evaluation of Urban Renewal, The Brookings Institution, Washington, D. C., 1967, particularly Chapter II.

in income to a "barely" poor family. An increase in the income of the barely poor is in turn more valuable than a similar increase in income of the non-poor. Or increases in the income of Negroes may be valued more highly by society than increases in the income of whites. If such a set of weights could be specified, a new figure of merit for the program's impact could be formed that consisted of the weighted sum of the benefits to differing segments of society. A similar weighted sum of the costs would also be needed.

It is difficult to conceive of a feasible way to arrive at an explicit set of weights. Clearly, however, a set of weights is implicit in the actions of Congress and various executive departments.* Because of the difficulty in arriving at a set of weights, the best the evaluator can do may be to simply portray the costs and benefits of a program for different subgroups in society. Thus, for example, analysis of poverty program outcomes might consider two groups, the poor and the non-poor (roughly speaking, these latter are the taxpayers). The benefits to the poor would include increased earnings resulting from program participation plus other increases in income from sources such as welfare or training allowances. Costs to the poor would include earnings foregone while in training plus out-of-pocket expenses for transportation or baby sitting services.

For the taxpayers the primary benefit of the program is probably the satisfaction that is derived from seeing the welfare of the poor improved. The value of this satisfaction is hard to determine. However, there are also tangible benefits. The increased earnings of the poor may be accompanied by a decrease in welfare payments, by decreases in crime against the taxpayer or more generally, decreases in the cost of social services from levels that would have existed in the absence of the program. The cost to the taxpayer is the cost of the program including training allowances (if any), net of decreases

*For a discussion of the need to integrate distributional effects and efficiency in assessing the cost and benefits of a program, see Burton A. Weisbrod, "Income Redistribution Effects and Benefit-Cost Analysis" in Samuel B. Chase, Jr. Ed., Problems in Public Expenditure Analysis, The Brookings Institution, Washington, D. C., 1968.

in other payments such as welfare that result from the existence of the program.

When the outcomes of programs are portrayed in terms of their consequences for various segments of society many of the questions concerning the treatment of elements of costs and benefits are simplified. Transfer payments such as welfare payments, for example, are usually not considered either a cost or a benefit in benefit-cost analyses because such a transfer simply represents a shift of consumption from one group to another. No consumption is foregone by society as a whole. However, it is clear that such transfers have significant consequences for different groups in society and form an important effect of most social action programs.

Although a tabulation of costs and benefits to various segments of society are important, it is clear that the policymaker is likely to want a figure of merit for the program that summarizes its performance. This desire is part of the reason for the popularity of the benefit-cost ratio. The construction of such a figure of merit should depend upon the objectives of the program. For manpower programs targeted on poverty populations, the following formulation might be used. Basically the objective of the program is the increase in the economic welfare of the target population. The costs are the foregone consumption of the rest of society. With such a formulation the benefits are:

(1) the increased earnings (net of taxes) of the target population resulting from participation in the program

(2) plus the net increase in transfer payments to the target population during participation in the program

(3) less decreases in transfer payments to the target population because of higher earnings subsequent to program participation

(4) less losses of earnings from work that would have been performed if enrollee had not been in program

(5) less losses of earnings of poor individuals displaced by trainees.

The sum of these changes is simply the stream of increments (or decrements) of real income both during and after the program which are attributable to the program.

The costs should include:

(1) The direct costs of the program including subsistence payments

(2) less any decreases in other transfer payments occasioned by the existence of the program

(3) plus losses of income of the non-poor if they are displaced by the program enrollees

(4) plus any decreases in income to the non-poor that occur because trainees are temporarily withdrawn from the work force

(5) less long term decreases in transfer payments because of the higher earnings of target population resulting from program

(6) less net external benefits which accrue to the non-poor and are not reflected in earnings of target population

(7) less the increases in taxes paid by the target population on earnings increments resulting from the program.

Numerous assumptions must be made in order to obtain estimates of many of the cost components. This is particularly true for items 3, 4, 6, and 7. For example, increased taxes paid by program participants have value to the non-poor only if they result in lower taxes for the non-poor or the support of other government programs that benefit the non-poor. Calculation of such quantities depends upon assumptions concerning level of economic activity, the reaction of the government to increases (or potential increases) in tax revenues, and the distribution of the benefits of government programs among the poor and non-poor.

Costs and benefits occur over a considerable period of time. In order to compare costs with benefits, both streams are discounted back to the present time using some value of discount rate. The proper value of discount rate to use has been the subject of considerable debate, a debate I do not choose to enter.* It is worth noting, however, that the relative ranking of programs will not be affected by the choice of a discount rate unless the temporal patterns of costs and benefits differ between the programs. The absolute ratio of benefits to costs will be significantly affected by the choice of discount rates.

Because of the many assumptions that must be made, the probability that an evaluation by one investigator will be comparable to that of another is not high. Comparison of two programs using figures generated by two different analysts is usually unwise. Two practical suggestions to improve this situation can be advanced. First, whenever practical, programs having similar or overlapping objectives should be simultaneously evaluated using identical assumptions (and if possible identical data collection efforts). Second, efforts should be made to develop an agreed upon set of conventions for the evaluation of social action programs similar in concept to those contained in "Green Book" for water resource projects.

NON-MONETARY BENEFITS

The discussion has proceeded as if all program benefits could be reflected in monetary terms. This is clearly not the case. There are benefits to the poor that are not measurable in dollar terms. Improvements in self-image, improved access to public services because of better knowledge, less alienation from the world of work or from other segments of society, better health or improved reading and computational skill are but a few of the non-monetary benefits that

*The choice of a proper rate of discount is extensively discussed in Economic Analysis of Public Investment Decisions: Interest Rate Policy and Discounting Analysis. Hearings Before a Subcommittee on Economy in Government of the Joint Economic Committee of the Congress of the United States, 90th Congress, 2nd Session, 1968, Washington, D.C.

are thought to accrue to participants in various manpower programs. To some extent some of these may be positively associated with income increases. Hence, comparison of programs in terms of their impact on increasing incomes will implicitly consider these factors. There is no simple way to include those factors that are more directly associated with program experience in the calculation of benefits.

If two programs have the same monetary benefits relative to costs, it might be possible to choose between them on the basis of the probable relative impact on other non-monetary benefits. For situations where the benefit-cost ratios differ, the judgment is much more difficult. Consider, for example, a comparison of the Job Corps and the Neighborhood Youth Corps (NYC). Suppose the youths from both programs gain the same benefits in terms of increased income. The youths from Job Corps receive extensive medical and dental care, considerable counseling, remedial education and some vocational skills, all in a residential environment. The youths in NYC, on the other hand, receive only work experience with generally limited amounts of remedial education and counseling. The Job Corps costs about four times as much per trainee as the NYC. Hence, with the assumption of equal monetary benefits, the benefit-cost ratio of Job Corps would be one quarter of NYC's. How much of this difference can be attributed to the failure to adequately account for the improved individual welfare associated with good health or reading capability? This is a matter of judgment that is now made, in the case of manpower programs, by an ill-defined set of decisionmakers in OEO, the Department of Labor, the Budget Bureau, the White House, and Congress.

This problem must be carefully separated from the one in which these non-monetary program outcomes are thought to lead to subsequent increases in income. The benefits described in the previous paragraph are what the economist calls "consumption" benefits to program participants leading to improvements in his current well-being. However, many of these benefits, such as health status, reading skills, or degree of alienation from various groups in society may be related to long term work experience. Improvements along these lines may

improve the capacity of the individual to find and keep a job, but this improvement may not be clearly discernible in the proximate work experience of the individual. In this case if the Job Corps provides the individual with capabilities that become useful only after some work experience or when the youth is older, then comparing the monetary benefits of the two programs only on the basis of proximate work experience is inappropriate. Unfortunately, there is little basis for determining the impact of many factors, such as health, upon the lifetime earnings of an individual. The analyst has to retreat to the rather unsatisfying activity of specifying the size of the improvement in employment or wage rates that would be required to equate the benefit-cost ratios so that the policymaker can more easily make a judgment about the probability that such a future difference can be expected to occur.

CONCLUSIONS

This section has touched on a few conceptual problems associated with benefit-cost analysis. A glance at any group of evaluations of manpower programs will be sufficient to indicate the great variety of ways analysts have approached the problems noted here. This variability has rendered the studies incomparable and to some extent has discredited benefit-cost analysis.* Steps should be taken to reduce this variability, perhaps by establishing conventions under which benefit-cost or cost effectiveness studies of human resource programs would be conducted.

*For example, three evaluations of the Job Corps using essentially the same data yielded estimates of benefit-cost ratios ranging from .3 to 5. See Lillian Regelson, "Applications of Cost-Benefit Analysis to Federal Manpower Programs," a paper presented at a meeting of the Operations Research Society of America, Denver, June 1969.

III. THE MEASUREMENT OF BENEFITS AND COSTS

In the previous section elements of a conceptual framework were established for comparing the costs and benefits of undertaking a manpower program. It was implicitly assumed that data on both the costs and benefits were available and that the major task of the evaluator was specifying what data to aggregate to obtain meaningful measures of costs and benefits.

Although it is true that many evaluations utilize questionable assumptions in calculating costs or benefits, the major difficulties seem to lie in empirically estimating these figures. Data produced routinely as a by-product of program operations suffer from two major flaws. They tend to be unreliable. Data for many projects are missing or contain numerous errors. More serious is the fact that few projects follow enrollees after the training period and hence are in a position to report earnings or employment histories.* Hence, the fundamental data required to assess benefits of a training program, the earnings of the trainee, must be obtained by other means. In most cases, the other means is some form of survey.

In general, the increase in national output is measured by the increase in income of the trainee. The use of this measure can be justified by the assumption that wages are equal to the marginal product of the worker. Two further assumptions are required. First, wages should represent total compensation. If extensive fringe benefits are also "paid," the use of only wages understates the program benefits. Second, it must be assumed that the enhanced employment and income status of the trainee has not been at the expense of someone else -- that there is no displacement of workers by the trainees. This is a hard assumption to validate, for displacement is difficult or impossible to measure. Displacement should be less during periods

*The reporting system for the Manpower Development and Training Act includes data on work histories of enrollees subsequent to enrollment. These data are supposed to be collected by the Employment Service but the return rates are quite low.

of high employment (labor shortages) than during periods of economic slack.*

If the objective of the program being evaluated is to enhance the economic welfare of a target population, increases in income experienced by the trainee as a result of his training must be measured. However, the change in income is made up of many more factors than simply changes in employment rate and wages. Changes in welfare payments, unemployment compensation, and other forms of transfer payments that result from program participation must be measured. Taxes must be netted out. Decreases in economic welfare of other members of the target population who are displaced by the trainee should be accounted for if such displacement takes place.

Measurement of all these effects poses significant problems. How much of the change in the wage income of a trainee should properly be attributed to his training? In many instances, individuals can expect normal increases in their income. During periods of increasing economic activity, labor markets tighten and unemployment rates decrease; wages frequently rise. In such circumstances, the income of most of the work force may be expected to increase. Young workers just entering the labor force typically experience considerable unemployment and only low wages, partly as a result of laws that prohibit them from taking certain jobs. More important, perhaps, is the fact that a youth is trying out jobs in search for work that appeals to him, a process that often leads to unemployment. As he ages, his wages and employment increase. If a training program has a large number of youths, much of the observed increase of income of the trainees can be attributed to this maturation process.

The ideal measure of the increase in trainee income is a comparison of his actual income subsequent to training with what his income would have been without training -- clearly an impossible comparison.

*The displacement effect has an anologue on the cost side. Opportunity costs to society due to the withdrawal of labor from the work force depend upon the employment level. In conditions of high unemployment, opportunity costs should be much less than the earnings that would have been received by the trainee if he had not been working, since other labor stands ready to fill the demand the trainee does not meet.

In the absence of this measure, the best substitute is the work experience and earnings of a control group of individuals who are similar to the trainees in all respects except for the receipt of training. The most satisfactory control group is that formed when potential trainees are randomly assigned to either training or the control group. Such assignments are generally held to be socially unacceptable and I know of no case where such a procedure has been used to construct a control group for a large social action program evaluation.

Many other types of controls have been tried -- none of which is very satisfactory. These include:

(1) The program enrollees themselves before and after comparisons).

(2) Groups of individuals who signed up but failed to enter the program.

(3) Groups of individuals who stayed in the program only a short time.

(4) Groups of individuals having similar backgrounds who for one reason or another did not sign up for training.

The first type of control, the experiences of the enrollees prior to enrollment, has already been discussed. It has very limited credibility at times when labor market conditions change rapidly or in the evaluation of programs serving a large number of youths. The second, third, and fourth types of groups have grave problems of their own; the most pervasive and yet unanalyzable problem is the so-called self-selection problem. Because the trainee group chose to enter the program and the control group chose not to, the two groups may differ in systematic yet unmeasurable ways. In general, the dimensions of these unmeasurable differences are considered to be attitude and motivation.

RAND's experience in examining a comprehensive youth program illustrates this problem.* A retrospective survey of program enrollees

*L. P. Holliday, Appraising Selected Manpower Training Programs in the Los Angeles Area, RM-5746-OEO, The RAND Corporation, May 1967, pp. 8-9.

was made. Short term enrollees, those staying less than a week, were used as a control group. Their average stay was less than two days, By various criteria, those in the control group did better than the longer term enrollees. In seeking an explanation for this, the analysts reached the tentative conclusion that the "controls" were typically more motivated than the long term program participants. They left the program quickly because they felt they could do better elsewhere -- in this case, by seeking a job by themselves. Indeed, there was some suggestion that the program facilitated this by providing placement counseling.

In contemplating this finding, however, we decided that had the result turned out otherwise, we would have had little confidence in the result. There appears to be an equally plausible set of arguments that would hold that short term stayers or no-shows (the second and third types of control groups listed on the previous page) are less motivated and able. Perhaps the distribution of motivation and attitude for this group is really bi-modal. It includes both the least and most motivated individuals in the population served by the program. One or the other type may predominate in any particular case.

Of the four types of control groups listed, the most satisfactory appears to be the last, a group of individuals who have similar work histories but have never come in contact with the program being evaluated. The choice of such a group has been accomplished in several ways. The Somers study in West Virginia utilized a random sampling of individuals in the files of the employment service.* Earl D. Main used a control group of friends, neighbors, or relatives of the trainee whose names were obtained from the trainee. Page and Gooding used persons who filed regular claims for unemployment compensation who reportedly had similar demographic characteristics.**

*Gerald Somers, Ed., Retraining the Unemployed, The University of Wisconsin Press, Madison, Wisconsin, 1968, p. 26.

**Main's, Page's and Gooding's results are reported in Einar Hardin, Benefit Cost Analysis of Occupational Training Programs: A Comparison of Recent Studies, paper presented at the North American Conference on Cost-Benefit Analysis of Manpower Policies, May 1969, University of Wisconsin, Madison, Wisconsin.

Although the last type of control seems most satisfactory, it is by no means obvious that it eliminates the self-selection bias. For this reason, lingering and reasonable doubts about the validity of the estimates of program effects will remain.

LONGITUDINAL VERSUS RETROSPECTIVE STUDIES

Further steps can be taken to satisfy doubts about the adequacy of a control group if the study is longitudinal and the control group is actually chosen before the trainees whose experiences are to be examined enter the program. Such a prospective and longitudinal study of a major manpower program has not to my knowledge been made, although OEO is now in the process of implementing one.

All of the studies reviewed in the course of preparing this Memorandum were retrospective and most obtained their data at only one point in time. The major limitation of a retrospective non-longitudinal sample is the inability to measure attitudes at different points in time. As a consequence, a control group can be compared with an enrollee group only on objective factors such as age, race, sex, or work experience. Questions about current attitudes or expectations are difficult to phrase and interpret but there is even less reason to place credence in such questions when they refer to a much earlier point in time. Thus, in none of the benefit-cost studies examined were attitudinal questions used to control for differences between a control group and the enrollee group.

In a prospective and longitudinal study, of course, attempts can be made to ascertain the attitudes and expectations of the two groups and differences in these dimensions can conceivably be controlled in comparing the work experiences of the two groups. Such control, however, is hampered by the absence of any well-developed and accepted theory concerning the relationship of attitudes and expectations to job search and retention behavior.*

*A preliminary example of such a study is contained in two publications by Ralph Underhill: Youth in Poor Neighborhoods and Methods in the Evaluation of Programs for Poor Youth, The National Opinion Research Center, Chicago, Illinois, 1967 and 1968.

Longitudinal studies can have other advantages of course. If repeated interviews are made, they may result in more reliable estimates of the sample's work experience because the respondent is not asked to recall information over long periods of time. Program experiences can be monitored in greater detail than that provided by program records. But such studies have disadvantages also. Most important perhaps is their expense. They take place over a longer period of time which means the evaluation staff has to be kept intact for a longer period. Longitudinal studies are susceptible to sample degradation as members of the original sample are lost because of moves, death, or simply because they cannot be found. This may result either in small ultimate sample sizes or a larger initial sample.* Since a prospective, longitudinal evaluation of a large social action program has yet to take place, the importance of both the potential benefits and problems cannot be realistically assessed.

One frequently voiced complaint about prospective and longitudinal studies is that they require longer to complete, increasing the probability that the evaluated programs will have changed; the chance that the evaluation will be irrelevant is higher than would be the case in a retrospective study. This complaint must be examined carefully. If a program is to be evaluated on the basis of the experiences of individuals entering in or terminating from a program during a specified period of time, either type of study will provide data at approximately the same time. But, the decision to undertake the longitudinal study must be made much earlier -- sometime prior to when the enrollees whose experiences are to be examined enter the program.

Up to now, evaluation has been a fairly ad hoc activity. Once it was decided that an evaluation was to take place, there were substantial pressures to obtain information as soon as possible. If, however,

*Such loss also results in biases because the lost group may be different from those that are found. But these biases are also present in retrospective surveys with low response rates. At least in the longitudinal survey, earlier data can be used to compare the characteristics of the group that is lost with those that are found.

evaluation becomes more routine, longitudinal study designs become more feasible with a continuing succession of such efforts in being at any given time. With such a commitment, longitudinal studies could provide data to decisionmakers at least as quickly as retrospective studies.

Clearly the desirability of instituting such a continuing program depends upon a variety of factors. Since such studies have not been carried out, we have little evidence on these factors. What is the cost per subject in the sample? Does the probable increase in confidence in the validity of control group/trainee comparisons seem worth the increase in cost? What is the probability that a major program reorganization will render the evaluation results irrelevant? The current OEO evaluation of manpower programs should clarify many of these questions.*

THE PROJECTION OF BENEFITS

Whether or not a study is retrospective or prospective and longitudinal, it will examine work experiences only during a short period of time, perhaps six months to a year. It is generally felt that benefits are likely to accrue over a period of years and hence some techniques must be applied to project the proximate work experience into the future. The number of ways in which benefits have been projected is approximately equal to the number of studies that have been carried out. The assumptions concerning the projections are usually dependent upon the data available. For example, Cain in his study of the Job Corps has inadequate data on employment rates so he assumes a constant and equal employment rate for both control and enrollee groups. Differences in income are due entirely to wage rate differentials. On the other hand, Borus and Somers have observations on income and project the observed differences. Borus chooses to project these for 10 years and assume no benefits accrue to trainees if they leave jobs for which they were trained. Cain and Stromsdorfer project the earnings

*OEO is currently carrying out a longitudinal evaluation of five manpower programs. Program enrollees in ten cities will be interviewed several times during a period of 18 to 20 months. The evaluators also hope to examine the impact of local labor market characteristics on program outcomes.

over a lifetime, correcting for mortality and assuming a decreasing differential between the controls and trainees. This decreasing differential was used because longitudinal data in the West Virginia studies suggested that the differentials faded.*

Clearly, the method of projecting benefits will be important in determining the absolute value of the benefit-cost ratio. But, if the same methods are used to project proximate earnings for each of several programs, the choice of method will not affect the relative levels of the ratios of proximate benefits to costs. Thus, if the evaluation is being undertaken to examine the relative effectiveness of several programs in achieving an objective, there appears to be little to be gained in projecting the observed income differentials over a lifetime unless there is solid evidence that income differentials will behave differently through time for the different programs.

As soon as the need to aggregate program outcomes into a unique and undimensional measure is relaxed, it is possible to compare programs against a variety of criteria. For example, programs could be compared on the basis of their contributions to lowering unemployment or increasing wage rates or perhaps changing rates of family desertion. Typically, such analyses are called cost-effectiveness analyses and are appropriate when comparing alternative means for achieving the same ends.

THE EXAMINATION OF ALTERNATIVE DESIGNS

There are relatively few national manpower programs. Moreover, these have been established with only vague hypotheses concerning the combinations of services that are likely to be successful. It is tempting therefore to structure an evaluation in such a way as to provide insight on alternative designs. Suppose the projects examined

*See Glen G. Cain, Benefit/Cost Estimates for the Job Corps, op. cit.; Michael Borus, "A Benefit/Cost Analysis of the Economic Effectiveness of Retraining the Unemployed," in Yale Economic Essays, Volume 4, 1964, pp. 371-429; and Glen G. Cain and Ernst W. Stromsdorfer, "An Economic Evaluation of Government Retraining Programs in West Virginia," Chapter IX in Retraining the Unemployed, op. cit.

differ in the mix of services provided or the type of personnel utilized. Would it be useful to view these projects as a form of natural experiment that could be used to cast light on superior project demands and hence suggest changes that should be made in program guidelines?

Two major problems limit the value of the natural experiment. The first problem has to do with multiple causality. In RAND's examination of a comprehensive youth program there was some indication that successful labor market performance was inversely related to length of stay in the program. There are a number of plausible explanations for such a phenomenon. Perhaps the most reasonable is that youths with more severe problems tend to stay in the program longer and also to have worse labor market performance after they leave the program. Ascribing all of the poor labor market performances to the length of stay rather than to some unmeasured personal characteristics of the enrollee results in the conclusion that the program may be detrimental.

The same may be true for attempts to relate the success of a project to the mix of services it provides. To the extent that the mix of services reflects the peculiar and unique (but unmeasured) needs of the enrollees of the project, attempts to relate success of the project to the mix of services will be frustrated. It will be impossible to separate the impact of the service mix of the project on the labor market performance from the impact of the quality of the enrollees.

Multiple causality frequently plagues the social sciences. Basically, this problem arises because of the lack of a theory of human behavior that relates measurable psychological variables to various forms of human performance. In the absence of such theory, it will be impossible to seaparate the effects of the multiple causes in natural experiments. This has led to suggestions that more formal experiments be carried out. In such experiments a more systematic attempt

*See for example Glen G. Cain and Robinson G. Hollister, "Evaluating Manpower Programs for the Disadvantaged," a paper presented at the North American Conference on Cost-Benefit Analysis of Manpower Policies, May 14-15, 1969, University of Wisconsin, Madison, Wisconsin.

would be made to vary project inputs independently of the enrollee characteristics and so lessen the problems of multiple causality.*

The use of experimental projects as a means of systematic program development is likely to be more common in the future. Certainly OEO's experience with rapidly initiating large national programs on the basis of "theory" rather than proven experience would not support the contention that this approach to program development should be continued. Yet the value of experimental projects or social experiments as a means of program development and as a source of planning information remains to be demonstrated. Such experimentation will be expensive and may not lead to replicable designs. It will take considerable periods of time and require an uncommon cooperation between project operators and evaluators. While the use of social experiments for program development and planning remains an exciting possibility, it should not be viewed as a panacea for the planner seeking to improve program design.

SUMMARY

These comments on problems associated with carrying out meaningful program evaluations are intended to convey the impression that evaluators still have a long way to go before they can routinely produce evaluations that are unassailable and reproducible. In the near future, evaluation will remain an art. New efforts should be viewed in part as attempts to improve methodology.**

How then, in light of these potential and actual shortcomings of actual evaluations, should an agency proceed to utilize such evaluations in its planning efforts? This problem will be considered in the next section.

---

*The major current example of such an experiment is Project Follow Through which is seeking to try out a substantial variety of programatic approaches to helping disadvantaged youngsters succeed in the early years in school.

**Federal agencies that want to improve evaluation efforts would do well to promote continuity and quality of the staffs that carry out these efforts. In informal observations of the staffs carrying out evaluation, one gets some sense that each new evaluation effort starts out fresh with unfortunately little input from previous evaluations.

IV. THE RELATIONSHIP OF PROGRAM EVALUATION TO THE PLANNING PROCESS

The previous section considered the conceptual underpinnings of program evaluation, particularly as they apply to manpower programs. In general, the discussion assumed that the evaluator was simply trying to compare the benefits with the costs of the program. This perspective characterizes what OEO calls Type I or impact evaluations.

If a pure Type I evaluation of a single program is undertaken it will provide a figure of merit, a benefit-cost ratio for that program. What role can this piece of information play in the planning process? By itself, this ratio can do very little. If it is unsatisfactory, that is, if the benefits are low relative to the costs, it may result in initiating a search for better ways to achieve the program's objectives. But such a ratio provides no clue about how to find a better program.

If several programs exist and have the same (or at least overlapping) objectives, simultaneous Type I evaluations may provide information on the relative effectiveness of the two programs. If due attention is placed on distinguishing average from marginal costs and benefits, a planner would recommend a shift of resources away from the program with a low marginal benefit-cost ratio toward the program with a high ratio. As noted in the last section, however, obtaining information on marginal as opposed to average benefits and costs may require posing questions about the effectiveness of components of a program with different parts of the target population. This type of question requires what OEO calls a Type II evaluation.

The planner obviously has a much greater menu of alternatives than just increasing or decreasing the funding levels of existing programs. He can:

(a) Add new programs and/or delete old programs,
(b) Change the management of existing programs,
(c) Change the design of existing programs, including the mix of services and/or the target population,.

(d) Change the mix of local projects within the national program, as well as

(e) Reallocate resources among the programs.

In the context of the analytic structure presented in the last section, Type I evaluations can provide guidance only for decisions relating to reallocation among the programs. In many instances, however, decisions falling within the first four categories may be more appropriate. Such decisions require richer information than that provided by a "pure" Type I evaluation.

In light of the shortcomings of Type I evaluations for real world decisionmaking, is it worthwhile carrying them out? Does it really make sense to do studies that examine only the impact of a total program, not the impact of the program on subgroups of the target population or variations in impact as a function of variations in project design? The experiences of the Office of Research, Plans, Programs and Evaluation (RPP/E) with its initial major program evaluation is instructive in this regard, and suggest both the bureaucratic and methodological difficulties associated with program evaluation.

An evaluation conducted by the Westinghouse Learning Corporation and Ohio University sought to examine the national impact of Project Headstart, the major OEO-sponsored program dealing with pre-school education. Project Headstart itself has a substantial research and evaluation activity and several large national evaluations were conducted in the early years of the program. These evaluations had, in the eyes of RPP/E, a number of shortcomings, many of which were beyond the control of the evaluators or Headstart itself. They did suggest that children experience gains in cognitive and affective behavior during their exposure to the Headstart program, but they indicated that these gains might not be sustained once the Headstart youngsters entered public school. There has been a good deal of debate on this "fade-out" or "catch-up" phenomena. Advocates of Headstart argued that what was occurring was that other children in the schools were catching up with the Headstart children and, consequently, that the program was having a useful effect. Critics or skeptics suggested

that these gains were fading out, either because the program did not have sufficient impact or lasting effect or because the nation's public school systems were so unresponsive to the needs of disadvantaged youngsters that they could not capitalize upon gains the youngsters had made during Headstart.

One question that RPP/E wanted to answer was whether the total program had a positive effect. It is clear that in any program such as Headstart there are local projects that are very successful in preparing preschool youngsters to function more effectively in the school environment. There are also poorly run projects which provide almost nothing for the youngsters. Headstart sprung into existence in a great hurry, enrolling some 250,000 youngsters within a period of a few months in the summer of 1965. Any program that is inaugurated and expanded at such a rapid rate is bound to lack the kind of careful planning that might lead to relative homogeneity of the outcomes of the numerous projects. This, combined with the lack of proven theories concerning learning by preschool youngsters, gave rise to a reasonable doubt about the overall impact of Headstart.

In light of these possibilities and because the national Headstart program was not carrying out a national evaluation, RPP/E decided to inaugurate a nation-wide impact evaluation. The study was intended to provide indications of the performance of a national probability sample of youngsters who had participated in Headstart over a period of three years. The desire for fairly rapid results led the evaluators to choose an ex-post design. About one hundred projects were chosen randomly and students who had entered local public schools after some Headstart exposure were studied. The evaluation team attempted to find a group of comparable children to use as a control group who had not been in Headstart and who were currently in the same classrooms. In order to obtain reasonable national coverage with economically feasible sample size, eight students at each grade level (first, second, and third grades) from each project were examined, together with a comparable group of control students. Not all projects had existed for the entire three years so the sample size of third graders

is smaller than that of the second graders, which in turn is somewhat smaller than that of the first graders. The number of students examined for each of these individual projects is too small to allow one to reliably characterize the effects of an individual project and this was not intended to be the objective of the evaluation. Instead, the project results were aggregated to give a total national estimate of impact. While a determination of national impact was the major focus, subsidiary analyses examined differences between the program impact by regions of the country, by racial groups, and by whether the youngster was in the summer or the fall year program.

This study represented a classical retrospective "impact-only" design. It provided an estimate of the impact of Headstart along a large number of dimensions of cognitive and affective development. The impact of the program was determined by comparing the scores of the children who had had Headstart experiences with the scores of a control group selected from the same school system who had been eligible for Headstart but who had not been enrolled. The control group was matched to the experimental group on sex, racial or ethnic group, and whether or not kindergarden was attended. It was impossible to match the control group with the experimental group on socio-economic status because such data were not available at the time of sampling. However, extensive interviews were conducted with parents of both groups which, among other things, provided the socio-economic data required to match these groups. Covariance analysis was used to effect this match.

During the planning stage for this evaluation study, the Headstart organization argued that the study should not be carried out. They felt the design focused on too narrow a set of objectives, utilized instruments that could not properly assess the psychological and educational development of young children and, because it was ex-post, ran a significant risk of utilizing an imperfectly chosen control group. They felt adverse findings of dubious scientific worth were likely and that such findings would have unfortunate impacts upon the development of the program because of their effect on the morale of national and local Headstart organizations and on public support for the program.

The RPP/E response made several points. First, whatever the multiplicity of program objectives established by Headstart itself, the prime objective of the program is to help improve the functioning of the disadvantaged youngster in the school. The important functions are cognitive and affective (or attitudinal) development. Second, they admitted the possible shortcomings of the ex-post design but felt that an adequate control could be constructed or if not, that this fact would be detectable. Finally, they argued that OEO had the responsibility to make some judgments concerning Headstart's effectiveness and that the then current Headstart research program was not producing any data on the program's effectiveness.

It is important to note that the statements on both sides are potentially valid. They are also to some extent self-serving. A politically popular program such as Headstart which has reason to suspect an evaluation will turn out negatively is unlikely to want to be evaluated. An organization such as RPP/E that aspires to provide rational advice concerning the allocation of resources among programs based upon "hard" data will want an overall program evaluation done. Since the programming organization can only affect major resource allocations it cares little about time-consuming data collection and analysis efforts that seek to answer more complex questions than simply "is the program working?" On the other hand, the program managers will be concerned with all the nuances of program design and, if evaluation must be carried out, will seek a richness of information that will support decisions on program design. A well-publicized national evaluation that is negative may well have adverse consequences on the morale of program personnel and hence it is reasonable for program managers who feel they are still developing and improving the program to resist such an evaluation. But despite this possibility, a failure to assess the validity of claims of effectiveness breeds complacency and leads to the development of an entrenched bureaucracy committed to the status quo.

Headstart was developing a major longitudinal study which, in the minds of the program administrators, overcame most of the problems

of the quick retrospective design; they proposed that their study constitute the evaluation. This study involved only three or four sites and required five to seven years to complete. RPP/E felt that such a study, while potentially useful for program development, was neither timely nor sufficiently representative of the national impact to constitute an evaluation of the program.

The RPP/E study results were indeed unfavorable to Headstart. There was little indication that Headstart youngsters did any better than non-Headstart youngsters. Predictably, the study was attacked by advocates of Headstart on methodological grounds. Early versions of the study reached President Nixon's advisors and appear to have been instrumental in shaping the rhetoric of his pronouncements on the program. It is still too early to assess whether the evaluation will have a useful impact in forcing changes in the program design. It is clear, however, that the study provides little guidance on what changes should be made. A study that could provide such guidance would require a substantially more elaborate design, a larger sample size (hence more expense), and a longer execution period. No study that purports to be a national impact evaluation that will also provide clues to what program changes should be made has yet been mounted -- perhaps it cannot be.*

Conceptually and tactically, a simple impact evaluation is the easiest to carry out. The question posed in such an evaluation is simple: Does the program, as represented by a national probability sample, have a discernible impact along some specified dimensions, or does it not? Although there will be debate on dimensions, the sample design is relatively straightforward and the analysis is not terribly difficult. If, however, one wants to pose additional questions to be answered by the evaluative activities, the design becomes more complex. If it is desired to determine what kinds of treatments are effective, what types of teachers are effective, or what kinds of institutional

*The study did indicate that the summer Headstart program appeared to have less impact than the year-round program suggesting that reallocation of funds from summer to year-round projects would lead to improved outcomes.

environment seem most productive in achieving the project ends, one must not only describe these factors and determine which ones the child has been exposed to, but one must also design a sample that is of sufficient size and structure to allow meaningful statistical generalizations to be drawn. If as is generally the case in large social action programs, there is an absence of well-developed and concrete hypotheses about these factors, the design of the sample and of the survey instruments must proceed with a great deal of uncertainty. No doubt the sample sizes will be larger and consequently, the costs of evaluation will be substantially greater. These are not the only problems, of course. The length of time required to prepare for the data collection and subsequently to analyze the data collected, will be greater. The problems of effecting a bargaining agreement between the evaluator and the program to be evaluated will take longer because more agreements will have to be reached. Simplicity of design is lost and the possibilities of disagreement and argument over the design are substantially increased. Finally, the interpretation of the data is far more difficult and complex. Indeed, it is likely that there will be something in the data for everyone; for every conclusion one draws about the effects of the program, someone else can draw a different conclusion. It is hard to refute the Headstart evaluation's conclusion that no perceptible and consistent gains have been experienced by Headstart children in a national probability sample. But if that same study had been able to say that children of a certain background or sponsors of a certain type had been effective, surely the emphasis placed on the evaluation by proponents of the program would have been substantially different and might have obscured the overall pessimistic conclusion.

The qualities of impact-only evaluations designed solely to determine whether or not a program is having effects along relevant outcome dimensions can be summarized as follows:

- Impact-only evaluations are relatively easy to mount and interpret because only a single hypotheses is being investigated.

- o This simplicity translates into an ability to mount such an evaluation relatively quickly and to carry out the analyses associated with it relatively quickly.
- o Impact-only evaluations are politically dangerous because of their go/no-go quality. There is little capacity in the design to point out directions in which the program should be changed in order to improve its effectiveness, if indeed it proves ineffective. By the same token, it is the hardest type of evaluation for program managers to shrug off because of the straightforwardness of its conclusions.

The impact-plus evaluation contrasts with the impact-only evaluation in the following ways:

- o It examines a wider range of questions (in particular, what is working for whom) and consequently, it requires a longer set-up time.
- o The additional hypotheses to be tested mean that larger sample sizes are required, and with some survey designs these studies may require reductions in the number of sites examined and a reduction in the representativeness of the total sample examined.
- o Results are more equivocal and subject to many differing interpretations and hence are likely to be more politically acceptable but possibly less effective in producing change.

The choice between the two types of designs, or more properly along the continuum between the two types of designs, will depend upon the particular case. As a general rule, it would appear that the more complex impact-plus evaluation is appropriate to the early years of a program when adaptation is taking place. Later, the impact-only evaluation may be more appropriate, particularly when the evaluation is intended to support the allocation of resources among programs having similar objectives.

PROGRAM VERSUS PROJECT EVALUATION

The point of view explicitly taken in this Memorandum is that of a senior planner in a Federal agency who seeks to allocate resources among a number of programs. This is the usual view of evaluators at the Federal level. Bateman has commented on this:

> The development of PPB [Planning, Programming and Budgeting Systems] has, in most instances, been characterized by an almost exclusive concern with efforts to more optimally allocate resources among programs. Very little attention has been given to the problems of program management; that is, the organization of resources within a program to achieve the greatest effect. This is not unexpected in a department like Health, Education, and Welfare where the bulk of Federal financial resources are channeled through State and local administrative hierarchies, a circumstance which precludes extensive involvement in the day-to-day management and direction of program operations. In a sense, PPB has followed a natural course in emphasizing, through the legislative and budget process, the resource allocation issues among programs since it is precisely in those areas that able people have had the greatest power to produce identifiable change.*

In the National Manpower training effort, on the other hand, the need for rigorous project evaluation may soon become quite acute, while at the same time the value of program evaluations is reduced. The reason for this is the Department of Labor's attempt to decentralize the operations of the manpower program and to emphasize comprehensive local manpower programs. It will be important to seek methods of carrying out the evaluation of local projects.

It is beyond the scope of this Memorandum to treat this problem in any detail. However, several important points can be made. The comparison of the effectiveness of a number of projects requires far more data than an examination of overall program effectiveness. Not

---

*Worth Bateman, "New Techniques of Federal Program Management," in Federal Programs for the Development of Human Resources, A Compendium of Papers Submitted to the Subcommittee on Economic Progress of the Joint Economic Committee, Congress of the United States, Volume I, Washington 1968, p. 100.

only must each project be examined, but the records of a sufficient number of enrollees to characterize the local program's effectiveness must be collected. This puts a very great premium on finding an inexpensive means of following up on enrollees -- clearly interviews at $10 to $50 dollars apiece are much too expensive. At least two alternatives are available. One is the time-honored practice of using placement rates or other proximate criteria as the measure of program effects. This appears to be the measure used in the system Bateman has described. The problems with using such project-reported measures are well known. If a project knows that the placement rate is being used in judging its performance, it will tend to find ways to inflate this figure.* The more nearly the criterion approximates the true program objectives, the better it will be. The true objective of a manpower program is not high reported placement rates or even truly high placement rate; rather, its goal is continuing high employment rates or earnings for its trainees. Observations on employment or earnings clearly are superior to observations on reported placements.

One method of obtaining such information on a routine and continuing basis is to tap either the Social Security Administration or Internal Revenue Service files. These files have wide (though not total) coverage, are relatively inexpensive to search, and could provide a basis for a useful project evaluation system. Both the Department of Labor and OEO are investigating the use of such data.

Even after such data has been obtained, important conceptual problems remain. Local labor market conditions, patterns of discrimination, or geographic structure will materially affect project outcome. Very little work has been done on local labor market phenomena and even less work has been done on these phenomena as they apply to the types of populations served by manpower training programs.** In the

*RAND's experience in using placement data as a means of tracking individuals did not provide great confidence in the validity of these statistics. In one case, as many as 50 percent of the individuals reported as placed had never been at the firm. In many instances, individuals left after only a day or two of work.

**RAND is currently undertaking a group of studies of the impact of local labor market phenomena on the poverty population.

V. CONCLUSIONS AND RECOMMENDATIONS

On the basis of the foregoing analysis, this section seeks to advance a few words of advice to would-be evaluators. It seems far too early in the history of social action program evaluation to view these conclusions as more than suggestions for next steps to be taken, so perhaps the most important recommendation is that evaluations undertaken in the near future should be viewed as vehicles for the development of evaluation methodology as well as providing information on program outcomes.

It is important to reemphasize that my perspective is that of the planner who seeks information to guide his decisions. For the most part, my model of the planners' world is one where he has a group of program alternatives among which he can allocate resources. The role of evaluation, then, is to guide the resource allocation. But, as has been noted, this is too narrow a concept. If a program is not doing well, the solution may be to change its management or to improve certain types of services or to change its target population or even to redefine the objectives of the program. Indeed, as the Department of Labor moves toward increasingly comprehensive designs, such as the Concentrated Employment Program, there will be no obvious alternative programs in which to invest and the only decision alternatives are changed program designs.

In my judgment, planners and evaluators should informally conduct a "contingency" analysis before they start an evaluation. Such an analysis would pose questions about decisions to be made if one or another result is obtained. If such an analysis is carried out, I suspect one would seldom find an evaluation designed solely to estimate the overall impact of a program. If the evaluation shows a negligible impact, the consequences are simply too harsh for a government agency to take and the results will be suppressed if possible. A far more realistic approach is to propose a set of hypotheses that have implications for program planning. For example, it could be hypothesized that a program works better with youths than adults, males rather than females, or in loose labor markets rather than tight

labor markets. Answers to such questions provide guidance to the planner and, even if the overall program impact appears small, the planner is in the position to indicate how its impact can be improved.

An alternative approach is to sample program experiences in such a manner that exemplary projects can be identified. These exemplary projects then provide some information on the potential program effects as well as guidance on changes that can and should be made in the less effective projects. At present, this choice must be made on a subjective basis because little hard data are available on project outcomes.

There are no doubt other ways in which to structure program evaluations. I am convinced, however, that if the evaluator and the planner were to sit down and ask what will happen if the results show one thing as opposed to another, the quality of evaluation would improve, its relevance would increase and its results would be more likely to be used.

OBJECTIVES

The objectives against which the programs are to be evaluated have generally been inadequately specified. Much of the confusion over what is a cost and what is a benefit results, in part, from the lack of explicit statements concerning the program objectives. I have argued that two types of objectives dominate manpower programs. One is the increase of the national product through increases of the productivity of the labor force. The other is a distributional objective, to improve the welfare of one segment of society by increasing its labor market productivity. Legislative or administrative intent may suggest that secondary benefits be considered.

There is no reason why a program should be evaluated against only one objective. Indeed, there frequently is a mandate in the legislation supporting manpower programs to consider several objectives. As a consequence, specifications for evaluation efforts should include explicit statements of program objectives and, if possible, criteria by which to measure the degree to which the programs achieve those objectives. There is no need to insist that all program effects be

combined into a single measure of program impact. The world is a complex and messy place. Within reason, evaluations should reflect this complexity.

MARGINAL VERSUS AVERAGE EFFECTS

In principle, it is important to consider marginal rather than average effects. In practice, this is generally infeasible. Program evaluations deal with what exists (or more properly what existed at the time of the evaluation) rather than what might exist if the program were expanded or contracted. In putting evaluation results to use, however, marginal effects can often be taken into account. If increases or decreases in program enrollment will be limited to particular demographic subgroups, data on the average program effect on that subgroup are likely to be a better estimate of marginal effects than the average outcome for the program as a whole. Again, contingency analysis on the part of the planner and the evaluator should lead to better evaluation designs which, in turn, should improve the usefulness of the evaluation for making estimates of marginal impacts.

THE NEED FOR A SET OF CONVENTIONS

The results of the studies examined during the research for this Memorandum are not comparable. In part, this non-comparability results from differences in the data available to each analyst. More important, however, each analyst uses his own set of assumptions concerning such phenomena as displacement effects, opportunity costs, social rates of discount, and transfer payments. Clarification of program objectives should help resolve part of the problem, but there is still a great deal of scope for arbitrary judgment. Consequently, a set of conventions for carrying out benefit-cost or cost-effectiveness evaluations or manpower training programs should be developed and published.

DATA SYSTEMS

Little evidence has been found during this study to suggest that data systems currently used by manpower programs will ultimately

develop information that will support program or project evaluation. The data produced are unreliable and do not provide any useful output measures. As a consequence, in the short run, data produced by these systems should not play a large role in the planning for evaluations. Evaluations should rely upon sample surveys. The results obtained will be less subject to unknown biases.

In the long run, however, data systems should be designed with evaluation needs as well as the needs of local projects in mind. Much of the lack of reliability of information systems at the local level appears to come because these systems are of little or no use in program operations. Much of the data that must be fed into these systems are of little current use to project managers -- in part because they are not retrievable and are often out of date.

Several steps should be taken to improve the data systems. Their design should be explicitly tailored to the needs of local projects. If they fulfill a need, they will probably be used and the data will be more reliable. They should in all probability be automated. One of the factors that will facilitate their use is timeliness, which may most easily be obtained through modern data processing techniques. It may well be that the government should consider setting up regional computer centers that will support these systems. It seems likely that once such local systems are in being, they can be routinely tapped to provide the national reporting desired.

These developments still will not solve the problem of following up on the individual trainee. For this purpose, it appears that tapping into one of the national reporting systems such as those of the Social Security Administration or the Internal Revenue Service may well be the best approach.

LONGITUDINAL STUDIES

For several reasons, longitudinal studies have significant advantages for the evaluation of any social action program. When performed on an ad hoc basis, they have two significant disadvantages. They take a long time to perform and they are expensive. If the information

systems suggested above were developed and if evaluation was routinized, the disadvantages of longitudinal studies could be substantially reduced. The data could be timely and the costs significantly lessened. Such a set of changes will not come about without positive action. Whether the benefits of such designs are great enough to warrant such action is not clear. These benefits are, after all, emphasized particularly by deficiencies of the more commonly used retrospective designs. Whether longitudinal designs in practice will turn out to realize their theoretical advantages remains to be demonstrated. For this reason, the current OEO effort to evaluate several manpower programs using a longitudinal design should be carefully monitored.

SYSTEMATIC EXPERIMENTATION

For reasons that are quite similar to those militating for longitudinal studies, systematic experimentation with varying program designs has great appeal. Again, it is an appeal that grows largely out of the shortcomings of current efforts to learn from what I have called natural experiments. Certainly evaluations of current programs, even when the programs appear to have benefits exceeding their costs, leave many questions unanswered. Is there a less expensive design that will do much the same thing, for example?

Somers, in an introduction to a collection of studies that show consistently favorable outcomes for MDTA projects, states:

> These benefits of retraining programs are impressive. Their worth would seem to be well established. But there are still some nagging questions. One unanswered question is whether on-the-job training would provide even better benefit-cost ratios and whether methods can be found for encouraging on-the-job training of the disadvantaged. Another is whether non-training job development and human resource programs might do as well for the disadvantaged unemployed, without the higher costs of vocational training courses.*

Questions such as these suggest the need to systematically establish a set of demonstration programs that examine program design

*Somers, op. cit., p. 15.

alternatives that are not currently a part of our national programs. Such projects would not be the same as the current demonstration projects. These tend to be established as a result of a proposal from an individual or organization. There tend to be many idiosyncratic factors associated with the operation of the projects. Moreover, they are generally poorly evaluated. A program of systematic experimentation would seek to establish projects having a range of characteristics. There should be some replication of project designs to reduce the impact of particular personalities or localities.

Again, there is a current activity that deserves monitoring as a guide to the benefits of such an effort. The OEO-funded, and HEW-run, Follow Through program is attempting to simultaneously examine a variety of project designs utilizing, in part, a common evaluation design to assess the outcome.

A FINAL NOTE

It is clear that evaluations of program outcome at the national level have not been of sufficient quality to justify their use as a major input in planning a national manpower program. This is, as I have argued, only partly due to the very large methodological problems facing the evaluator. More important as an explanation are two organizational factors: (1) Most programs and most agencies are reluctant to be evaluated; (2) if they must be evaluated, they will seek to find evaluation designs that have the greatest probability of supporting the status quo.

But the evaluators themselves are also at fault. Too little effort has been placed on relating evaluation to the planning process. Too little concern has been given to identifying the decisions that evaluative efforts should be designed to clarify. Too often the evaluator has either chosen to or by default had to define his own evaluation objectives with little continuing interaction with the programs or agency. Too often the evaluation results in a final report whose summary and conclusion are read but whose data are left largely unanalyzed.

Solutions to these two problems (the reluctance to be evaluated and the irrelevance of the evaluations) push in opposite directions. The first argues for separating the evaluation function more sharply from the operating programs because otherwise the program managers will tend to render them useless. The second argues for bringing them closer in order to make the information generated by the evaluation more relevant and useful.

This is a quandry that requires close attention by senior agency administrators. There is no simple solution.

BIBLIOGRAPHY

Bateman, Worth, "New Techniques of Federal Program Management," in Federal Programs for the Development of Human Resources, A Compendium of Papers Submitted to the Subcommittee on Economic Progress of the Joint Economic Committee, Congress of the United States, Volume I, Washington 1968.

Borus, Michael E., The Economic Effectiveness of Retraining the Unemployed, A Study of the Benefits and Costs of Retraining the Unemployed Based on the Experience of Workers in Connecticut, unpublished dissertation submitted to Yale University Graduate School, New Haven, Connecticut, July 1964.

-----, "Time Trends and the Benefits from Retraining in Connecticut," The Development and Use of Manpower, Proceedings of the 20th Annual winter meetings of The Industrial Relations Research Association, Wisconsin, 1968.

-----, "A Benefit/Cost Analysis of the Economic Effectiveness of Retraining the Unemployed," in Yale Economic Essays, Volume 4, 1964.

Cain, Glen G., Benefit/Cost Estimates for Job Corps, Department of Economics and the Institute for Research on Poverty, University of Wisconsin, September 1967.

-----, and Robinson G. Hollister, Evaluating Manpower Programs for the Disadvantaged, a paper presented at the North American Conference on Cost-Benefit Analysis of Manpower Policies, May 14-15, 1969, University of Wisconsin, Madison, Wisconsin.

Center for the Study of Unemployed Youth, A Study of the Meaning and Effects of the Neighborhood Youth Corps on Negro Youths Who are Seeking Employment, Graduate School of Social Work, New York University, March 1968.

Committee on Administration of Training Programs, Report of the Committee on Administration of Training Programs, submitted to the Secretary of Health, Education and Welfare, Washington, D. C., March 1968.

Dorfman, Robert, Measuring Benefits of Government Investments, The Brookings Institute, Washington, D. C., 1965.

Dunlap and Associates, Inc., Survey of Terminees from Out-of-School Neighborhood Youth Corps Projects, Final Report prepared for Department of Labor, Bureau of Work Programs, Washington, D. C., May 1967.

Hanson, W. Lee., Burton Weisbrod, and William J. Scanlon, Determinants of Earnings: Does Schooling Really Count?, Economics of Human Resources, Working Paper 5A, Department of Economics, University of Wisconsin, April 1968.

Hardin, Einar, Benefit-Cost Analysis of Occupational Training Programs: A Comparison of Recent Studies, a paper presented at the North American Conference on Cost-Benefit Analysis of Manpower Policies, May 1969, University of Wisconsin, Madison, Wisconsin.

Harris, Louis and Associates, A Study of August 1966 Terminations from the Job Corps, Conducted for the Job Corps, March 1967.

-----, A Study of the Status of August 1966 Job Corps Terminees--12 Months After Termination, October 1967 (revision).

-----, A Study of Job Corps No-Shows: Accepted Applicants Who Did Not Go to a Training Center, March 1967.

Hearings Before a Subcommittee on Economy in Government of the Joint Economic Committee, Economic Analysis of Public Investment Decisions: Interest Rate Policy and Discounting Analysis, 90th Congress, 2nd Session, 1968, Washington, D. C.

Mangum, Garth L., Contributions and Costs of Manpower Development and Training, Policy Papers in Human Resources and Industrial Relations, No. 5, National Manpower Policy Task Force, Washington, D. C., December 1967.

McKean, Roland N., Efficiency in Government Through Systems Analysis, New York, John Wiley & Sons, Inc., 1958.

Planning Research Corporation, Cost/Effectiveness Analysis of On-the-Job and Institutional Training Courses, prepared for the U. S. Department of Labor, June 1967.

Ribich, Thomas I., Education and Poverty, The Brookings Institution, Washington, D. C. 1968.

Rothenberg, Jerome, Economic Evaluation of Urban Renewal, The Brookings Institution, Washington, D. C., 1967.

Somers, Gerald G., and Ernst Stromsdorfer, A Benefit Cost Analysis of Manpower Retraining, Proceedings of the Industrial Relations Research Association, December 1964.

-----, Ed., Retraining the Unemployed, The University of Wisconsin Press, Madison, Wisconsin, 1968.

Suchman, Edward A., Evaluative Research, Russell Sage Foundation, New York, 1967.

Underhill, Ralph, Youth in Poor Neighborhoods, National Opinion Research Center, Chicago, Illinois, 1967.

-----, Methods in the Evaluation of Programs for Poor Youth, The National Opinion Research Center, Chicago, Illinois, 1968.

Walther, Regis H., A Study of the Effectiveness of Selected Out-of School Neighborhood Youth Corps Programs: Methodological Considerations in Evaluative Research Involving Disadvantaged Populations (Draft), Social Science Research Group, George Washington University, May 1968 (Mimeo).

-----, and Margaret L. Magnusson, A Study of the Effectiveness of Selected Out-of-School Neighborhood Youth Corps Programs: Retrospective Studies of the Effectiveness of Out-of-School Neighborhood Youth Corps Programs in Four Urban Sites, Social Research Group, George Washington University, October 1967.

END