Evaluating Health Programs

O. LYNN DENISTON, M.P.H., and IRWIN M. ROSENSTOCK, Ph.D.

EVALUATION, by definition, is value laden, requiring the selection of certain qualities, attributes, or conditions; measurement of these qualities; and comparison of results with the underlying value system (*1*). Differentiating among several different uses of the concept of evaluation is important.

Evaluation is likely to be performed at the initial contact between a patient and a physician. The physician will select and measure a set of qualities (for example, weight, temperature, pulse rate, blood pressure, color, and alertness) and will compare the scores obtained with a norm that reflects his beliefs about the preferred condition of the patient. Such a procedure meets the formal definition of evaluation although it does not seem to describe all that is usually meant by evaluation of a program.

What is Program Evaluation?

Before the evaluation of a program can be discussed seriously, some agreement is needed about what a program is. We use the word "program" in many different ways (*2*). Neal (*3*) has suggested that people with a management-science viewpoint seem to agree that a program is ". . . a set of activities, a social enterprise, with certain inputs of resources and conditions, certain ways of organizing those resources and conditions and establishing relations among them and certain outputs with standards for evaluating them."

We prefer the following formal definition of a program: an organized response to eliminate or reduce one or more problems where the response includes one or more objectives, performance of one or more activities, and expenditure of resources (*4*). Neal's definition differs slightly from this one in that he does not allude directly to a problem whose elimination is valued. We, on the other hand, do not insist that standards for evaluating outputs (objectives) must exist in order for a program to exist.

What the two definitions have in common is that any size of enterprise or response could constitute a program. One could with equal validity label as a program this paper, this issue of *Public Health Reports*, the totality of the Public Health Service, or the work of a neighborhood health center.

Once agreement has been reached about what constitutes a program, the requirements for evaluation of the program can be specified. We believe that evaluation of a program should focus on the objectives (outputs, outcomes, goals) of the program in terms of their appropriateness, adequacy, effectiveness, efficiency, and side effects.

Appropriateness of the program is most directly related to value: the good-bad continuum. In evaluating appropriateness of the program, one asks if the objectives of the program are desirable. The decision depends on who answers the question and what his values are. Most people agree on certain values—peace is good, murder is bad—but considerable disagreement exists about other values: U.S. involvement in

Mr. Deniston is a research associate and lecturer in public health administration, and Dr. Rosenstock is a professor of public health administration. Both are members of the department of community health services, University of Michigan School of Public Health, Ann Arbor.

Vietnam, registration of guns, or keeping certain people biologically alive through heroic medical manipulations.

The dimension of appropriateness may be viewed in two ways. First, is the proposed program desirable or undesirable in an absolute sense? Second, and more difficult, is determining the degree of desirability or priority of a program in relation to other programs. Even if an objective is desirable, it is necessary to decide whether it is better than all other possible desirable objectives. Health workers probably agree that the eradication of measles, tuberculosis, and lung cancer are each desirable objectives, but they might disagree about which is the most important.

The critical question centers on who has the right to decide: the professional? the consumer? and which professional or which consumer? Even if these questions were answered, additional answers would be required concerning the objective and who is able and willing to describe it. If all these questions could be answered satisfactorily and appropriateness of the program thought of as a simple dichotomy (that is, good or bad), the evaluation could be straightforward. If appropriateness is a matter of degree, however, we have difficulty because our ability to measure the degree of value is not well developed.

Effectiveness and adequacy of the program are related; we separate them more for psychological than logical reasons. Adequacy is concerned with the extent to which a problem has been prevented or eliminated, while effectiveness is concerned with the extent to which an intended amount of attainment has occurred. Thus a program with an objective that the incidence of lung cancer be reduced by 50 percent and which attained that objective would be 100 percent effective but only 50 percent adequate since half of the problem still remained.

We believe that objectives should specify both what is to be attained (the valued condition) and how much is to be attained. The objectives of most current programs either propose the eradication of an existing problem or reduction of the problem by an unspecified amount. Eradication is usually unrealistic in that few people really expect it—at least in the short run. The second is unusable because it provides no basis for comparison of attainment with any value or expectation.

Efficiency is concerned with the cost in resources of attaining the program's objectives. Knowledge of effectiveness is thus prerequisite to knowledge of efficiency. Therefore, the definition clearly prohibits such often-stated conclusions as "we don't know how effective we were, but we were very efficient."

Thus four kinds of evaluation (appropriateness, adequacy, effectiveness, and efficiency) focus on the objective—the intended effect of operating the program. Another kind of evaluation focuses on other or side effects of the operation. We can never be sure that the operation will lead only to the intended effects. Side effects, either good or bad, nearly always occur. The thalidomide experience is one of the most familiar examples of undesirable side effects. The recent discovery of a highly selective and effective raticide while testing cancer drugs is an example of a good side effect.

Other Bases for Evaluation

Not all judgments about programs are based directly on the objectives of the program. Many judgments are based on data concerning the resources and activities rather than the objectives. Stanley (*5*) has aptly termed this approach to evaluation "presumptive." The operators of a program often presume that if the budget is of a particular size, if the personnel possess certain credentials, and if certain activities are performed, the program has some degree of effectiveness. The presumptive approach to evaluation in public health is best illustrated by the logic of the several evaluation schedules, appraisal forms, and "Health Practice Indices" published by the American Public Health Association between 1925 and 1950 (*6–8*). We are now more aware of the dangers associated with presuming that resources and activities invariably lead to desired outcomes.

Who evaluates the program? Everyone with any knowledge of a program evaluates it. Each evaluation varies with the knowledge of the evaluator, the criteria he selects as signs of success, and the data he uses to measure the criteria. Several groups have or will evaluate this paper: the authors, the reviewers, the readers, and those who read only the title of the paper but not its

content. If the evaluations differ, who can say which are correct?

Our approach to the evaluation of a program proposes that those people who decide that a program shall be created, or that one already in existence shall be continued, are the people who know the program's objectives. Thus we would say that evaluations based on objectives described by the operators of the program are correct. But many people who judge a program by the same objective frequently differ in their conclusions. Which of them are correct? It is not easy to decide.

Many times different conclusions are reached because of variations in the objectivity of the measures used. Objectivity means the extent to which clear-cut rules are formulated and followed for obtaining measures. In this sense, the procedures for performing laboratory tests are more objective than the procedures for making clinical judgments. We tend to accept the results of those who use the most objective measure, not because a high correlation always exists between objectivity and validity of measures but because we simply cannot say much about the validity of subjective measures.

Stanley (*5*) has suggested the terms "impressionistic" and "proven" to describe the ends of a continuum of objectivity. The act of measuring fever by placing the hand on the forehead tends toward the "impressionistic" or subjective end of the scale; when fever is measured by a certified thermometer, the "proven" or objective end is approached.

If we grant that all judgments are evaluations, there is no real lack of numbers of evaluations of programs. Rather, concern is with kind and quality of the evaluation. What is generally wanted is (*a*) clarity about the actual objectives of the program, (*b*) measures of attaining these objectives, rather than allocation of resources or performance of activities, and (*c*) measures that are more objective and more valid than the usual.

Relationship of evaluation to planning. Evaluation of the program is usually thought to be an assessment of its operation. We believe that in evaluating its effectiveness, not only the operation but also the accuracy of planning should be assessed.

Once decisions have been made about what problem should be attacked, the ideal process for planning usually would begin by obtaining information about the current state of affairs, referred to as baseline data, which rarely are sufficient to lead the planners of the program into action. Rather, baseline data are used for estimating what the status of the problem would be during or at the end of some planning period if no program were undertaken. On the basis of knowledge and expert judgment, one might estimate that, if left unchecked, the problem would increase, would stay about the same, or would diminish. This prediction of the future forms the basis for planning. Frequently, the projection is only implicit, but it is necessary. No one would use resources to eliminate a problem that he expected to disappear as quickly without using these resources.

Having estimated what the course of a problem would be without a program, one would next estimate how the problem would be affected if a program were undertaken. With cost-benefit analyses, one might have different estimates concerning the impact of the program. Such estimates could be based on the use of alternative programs by considering various levels of resources or different approaches to the problem.

Regardless of how one estimates the future status of a problem if a program were undertaken, that estimate constitutes the objective of the program and may be viewed in two ways: first, that the status of the problem be at a desired level during a given period or by a given time and, second, that the program produce the desired amount of change.

If both estimates of future status are made while planning the program, it becomes feasible to evaluate both accuracy of the planning and effectiveness of the program. When we evaluate, we measure the actual status of the problem and estimate what the status would have been without the program, using a control group, if possible, to make the estimate.

Although accuracy in planning and effectiveness of the program are related, the concepts are quite different, and the differences should be understood. Effectiveness refers to the extent to which specified objectives are attained as a consequence of program activity. Accuracy in planning refers to degree of correspondence between two estimates of what would happen to the prob-

lem if no program were undertaken. The two estimates are made at different points in time: the first before, and the second after, operation of the program. We use baseline data and expert judgment to make the first estimate and, if possible, control groups for the second estimate.

Since decisions to undertake or not to undertake programs at various levels of resource allocations are always based on estimates of what would happen without a program, it is important that health planners progressively increase their accuracy in planning. This can only be done by specifying each estimate for the future during the planning stage, checking the estimates against reality during evaluation, and then feeding that information back into the planning process.

Let us consider a hypothetical example (see chart) to illustrate the relationship between planning and evaluation. The top portion of the chart depicts a planning period. Point A represents the baseline status of the problem, point B represents the estimate of what the size of the problem would be if no program were undertaken, and point C represents the anticipated status of the problem if a program were undertaken. At D and E the status of the problem is estimated, given different levels of resource allocations or approaches to the problem that may have emerged from cost-benefit analyses, a computer simulation, or some other quantitative analysis. Assume that point C has been selected as the objective of the program. This means both that we want the absolute level of C to be at a particular point and that we wish to reduce the problem by the amount B minus C.

The bottom portion of the chart shows the results of an actual evaluation. Point C is determined by measuring the group exposed to the program, and point B is estimated by measuring a control group not exposed to the program or by some other method, several of which will be mentioned later. It can be seen that the absolute size of the problem following the program is lower than anticipated, but it is also evident that less change was produced than had been desired. That is, the actual difference between B and C is less than the difference planned between B and C. Looking merely at the absolute attained level of C gives cause for celebration, but looking at the difference between B and C gives cause for concern. Planning was not very accurate. When it is possible to make these various estimates, evaluation can contribute most to subsequent planning.

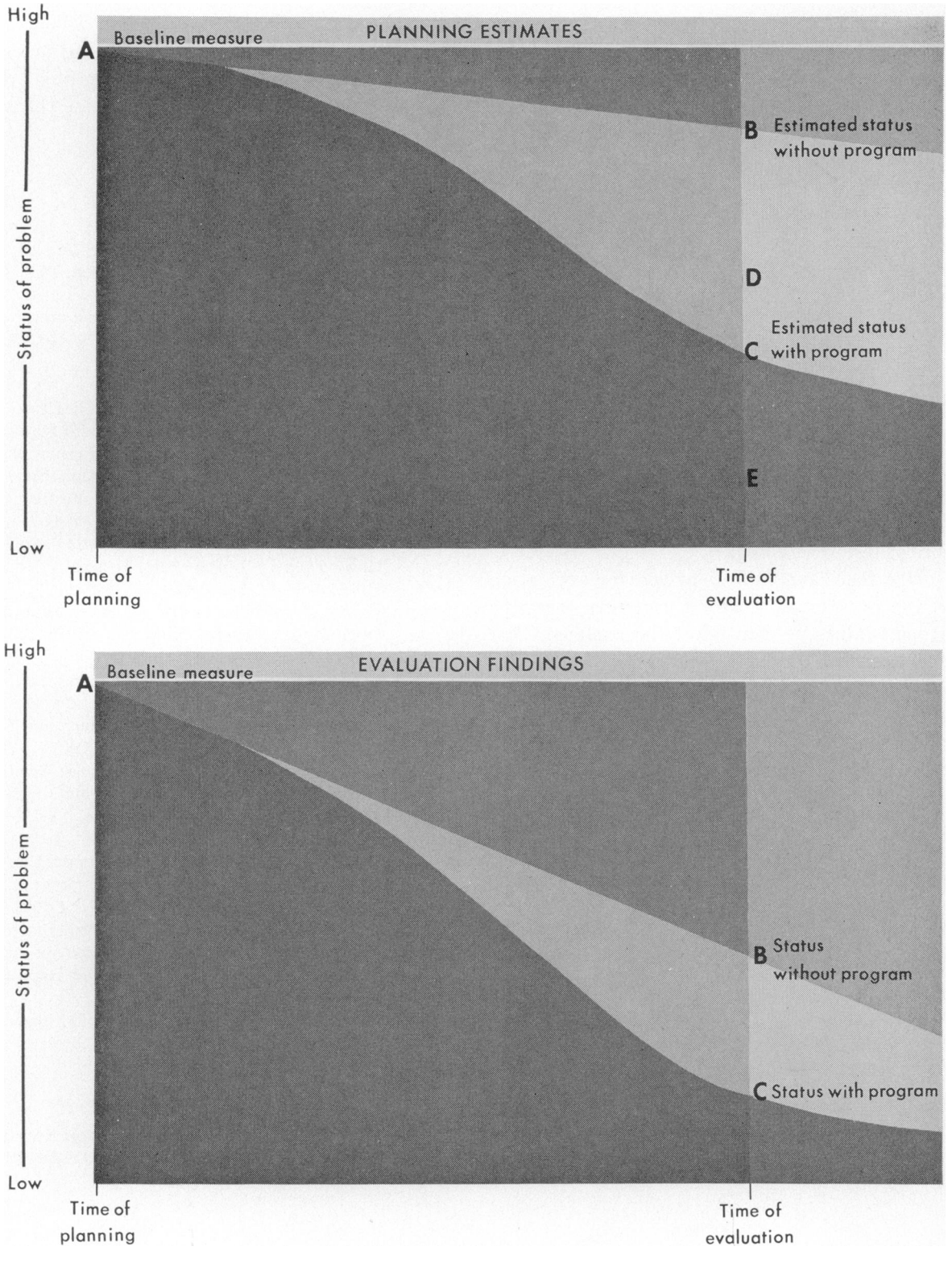
The typical but incorrect norm in evaluating the program is a before-and-after analysis; conditions at the time of evaluation are compared with conditions existing before the program. These two measures say little about the effect of the program itself. As shown in the chart, conditions might have improved without a program or they might have deteriorated. In one of our studies on food sanitation (*9*), conditions in restaurants that were inspected were worse at the end than at the beginning of a year; however, our analysis indicated considerable positive effect from the program. Without the program, conditions would have deteriorated far more than they did. A simple before-and-after analysis would have indicated a negative effect of the program.

Considerable attention has been given to estimating conditions at the time of evaluation had there been no program. We know that control groups provide the best estimates. Unfortunately, a classic experimental design often cannot be applied in real settings of programs; we cannot assign clients to treatment and nontreatment groups at random. Campbell and Stanley (*10*) have described a series of quasi-experimental designs that can be used for making this estimate. Most of these designs include measures of some group that is composed of persons not exposed to the program but who are believed to be similar in other important respects.

Circumstances often provide a basis for making this estimate even when none could be planned and built into the design for an evaluation. In the program for food sanitation, a staff vacancy that could not be filled and the subsequent lack of service to one district were the bases for estimating what would have happened without the program. In evaluating the housing for agricultural migrants (*11*), inadvertent failure to provide inspectional services to all housing areas during the first year of the program was the basis for the estimate.

Although the literature on research design devotes much attention to the problem of estimating how much of an apparent effect results from the procedures in the program rather than

Hypothetical example of relationship between planning and evaluation



from other causes, not as much has been written about how one goes about estimating future status while planning. Where the data are available over a considerable span of time, trend analysis can be performed. This procedure is useful for new programs, but if a routine program has been operating for some time, trend analysis is of little help in estimating what the conditions would have been in the absence of a program.

Perhaps the attention now being given to formal planning in universities and planning agencies will develop techniques for making such estimates. Or, if program operators begin to use the concept of accuracy in planning, additional methodologies may be developed.

Summary

A program is an organized response to eliminate or reduce one or more problems where the response includes one or more objectives, performance of one or more activities, and expenditure of resources.

Five foci have been identified for evaluation of the program. For appropriateness, were the proper values used to select the problem? For effectiveness, to what extent were objectives attained? For adequacy, how much of the total problem was eliminated? For efficiency, at what costs were the objectives attained? And for side effects, what outcomes occurred that were not central to the objectives of the program?

Programs are always evaluated, but the evaluations vary as to whether the measures are presumptive or direct and the degree to which the measures are impressionistic or objective. Valid and objective measures of program goals make it possible to assess a program systematically. Ideally, we should compare the actual status at the time of the evaluation with the status that would have existed had there been no program.

In the process of setting objectives, one should not only specify the desired amount of change but also the absolute level expected. If both estimates are specified when the program is being planned, subsequent evaluation can reveal both the extent to which an intended amount of change has occurred and also the accuracy of the planning estimates. When the findings are fed back into the planning process they should have the effect of increasing both the effectiveness of the program and the accuracy of planning.

REFERENCES

(*1*) Weckworth, E.: On evaluation: A tool or a tyranny. Presented at annual meeting of the American Public Health Association, Philadelphia, Nov. 12, 1969.

(*2*) Jackson, J.: Some issues in evaluating programs. Hosp Community Psychiat 18: 161–168, June 1967.

(*3*) Neal, F. W.: Doctors, dilemmas, data and decisions. Presented at a training institute on Program Evaluation in Mental Health Services, Portland, Oreg., Oct. 24, 1966. Mimeographed.

(*4*) Deniston, O. L., Rosenstock, I. M., and Getting, V. A.: Evaluation of program effectiveness. Public Health Rep 83: 323–335, April 1968.

(*5*) Stanley, D. T.: Excellence in the public service: How do you really know? Public Admin Rev 24: 170–174, September 1964.

(*6*) Committee on Administrative Practice: Appraisal form for city health work. American Public Health Association, New York, 1925.

(*7*) Committee on Administrative Practice: Evaluation schedule for use in the study and appraisal of community health programs. American Public Health Association, New York, 1943.

(*8*) Committee on Administrative Practice: Grading standards for A.P.H.A. evaluation schedule. American Public Health Association, New York, 1950.

(*9*) Deniston, O. L., and Welch, W.: Evaluation of performance of a food sanitation program. J Milk Food Technol 32: 115–121, April 1969.

(*10*) Campbell, D. T., and Stanley J. C.: Experimental and quasi-experimental designs for research. *In* Handbook of research on teaching. Rand McNally & Co., Chicago, 1963.

(*11*) Deniston, O. L.: Migrant camp conditions improved by inspection. J Environ Health 31: 338–346, January–February 1969.

Tearsheet Requests

O. Lynn Deniston, Public Health Practice Research Program, 122 South First Street, Ann Arbor, Mich. 48108