Evaluation and Program Planning, Vol. 3, pp. 91–94, 1980
Printed in the U.S.A.

0149-7189/80/020091-04$02.00/0

WHETHER EVALUATION – WHETHER UTILIZATION

O. LYNN DENISTON

School of Public Health
University of Michigan

In "Evaluation: Manifestations of a New Field," Flaherty and Morell (1978) suggest there is such great variety among the various endeavors termed evaluation that it is too early to define the characteristics of evaluation. Rather we should search for the common elements of these various endeavors and let that commonality define the process. At the 5th Annual Conference of the Evaluation Network, Kirkhart (1979) moderated a symposium on "Making Evaluation Results Useful; Knowledge Utilization Re-Examined," to some extent at least, a reaction to the oft cited complaint that results of many evaluations are ignored. Such inattention to evaluation results may be partly responsible for Tornatzky's (1979) argument, in "The Triple-Threat Evaluator," that evaluators should do more than evaluate; they should work toward modifications of the programs they evaluate, improving the effectiveness of current, helpful programs, "performing euthanasia of social programs that are ineffective, or harmful, or both." He goes on to say he "finds it difficult to separate the general role of the evaluator from the specific roles of researcher, *and* innovator, *and* politician-administrator."

Fox and Rappaport (1972) suggest programs need not be evaluated, but if they are, "good" evaluation can be utilized:

> On a different level, some mental health professionals argue that formal evaluation studies are inappropriate to the mental health field and that program decisions are better based on professional intuition than on such studies. We contend that, although program decisions must ultimately be based on judgment *regardless of whether* evaluation is conducted, good evaluation can provide information that can contribute to the decision making process. (emphasis added) (p. 172)

I would like to suggest here that there is no question about whether to evaluate, the question is which, if any, of the many evaluations is most trustworthy. Further, there is no question about whether utilization; the question is whether the utilizer did select, and properly utilize the most trustworthy evaluation. That leaves a question as to whether a profession of "evaluator" is needed with specially trained personnel to decide which if any current evaluation is "correct," or conduct adequate evaluations if current evaluations are all inadequate. I will suggest we do not, but the exchange of ideas among a wide variety of professionals about evaluation is indicated.

As a first step in these arguments, I would like to suggest a definition of evaluation which might find acceptance by a majority of currently interested people. If so, it could be a step beyond the position advocated by Flaherty and Morell (1978). I will then suggest a definition of program evaluation as a special case.

What definition of evaluation might gain acceptance? I suggest that *measurement, coupled with comparison of the obtained count or score to a criterion or standard* are necessary and sufficient conditions for evaluation. The fact the evaluator has preferences or values which identify the criterion or standard is the basis of the label "evaluation." The question, "How wide is your chair?" calls for a measurement, and a response such as "18 inches" indicates a measure has been taken. On the other hand, the question, "Is your chair wide enough?" calls for an evaluation and typical responses would be "yes" or "no." Such responses require first a measurement, then comparison of the score to some criterion, frequently one's own width.

Two different people might agree a chair was 18 inches wide yet disagree as to whether it was wide enough. More generally, two evaluators of the same object or program may reach different conclusions because they have different standards. But we should also recognize two measures of the same object or condition may obtain different scores.

If a group of people, in a classroom setting, are asked

Requests for reprints should be sent to O. Lynn Deniston, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109.

"How wide is your chair?" most answer in terms of inches (feet, centimeters, meters) but their answers are quite variable. When a ruler is available, their responses are much less variable and the scores are seen as more "objective."

If a group is asked, "How is the weather today?" there is a different type of variability. Several dimensions are utilized; degree of hot-cold, wet-dry, windy-still, clear-cloudy. Those using the hot-cold dimension differ in whether they utilize words (hot, chilly, frigid), or numbers, (17°). Upon further questioning, it turns out some who report in "degrees" measured by looking at a thermometer, others by what they heard on the radio during breakfast.

Thus, variability among scores upon measurement may arise because of the nature of the measurement process used, the uniformity of use of similar techniques or many other influences. Thus it should be no surprise that different evaluators often arrive at different conclusions since they may have measured different dimensions, or the same dimension in different ways, or compared the same score to different standards.

How frequently do people evaluate the width of chairs and the weather? At every new opportunity. We cannot see or sit in a chair without evaluating its width, color, and other dimensions; we cannot be in a weather without evaluating it, if we have senses of sight, feel, smell.

How frequently do people evaluate programs? At every new opportunity! Does that mean programs are just like the weather? No, they have an additional characteristic (assuming that we are not yet, or very persistently, trying to do anything about the weather).

Consider this conception of a program (our focus here is on human service programs, but the idea should be more general): *a purposeful response to one or more perceived problems with the intent of preventing, reducing or eliminating the problems.* What's a problem? We suggest, *"a situation or condition of people or the environment which in the opinion of responsible program personnel, would exist in the future, and be undesirable"* (Committee on Evaluation and Standards, 1970). This defines programs as based on values, reflecting the preferences of people, based on predictions about the future, and on beliefs about the possibility of a different future in the presence of program interventions. The intervention will require expenditure of time or other resources which would have been devoted to alternative programs, and the interventions may have "side-effects" other than those intended. The extra dimension of program evaluation, beyond measurement and comparison, is the requirement to establish the causal relationship between program interventions and the magnitude of problem(s) which lead to the program.

From this perspective we can see programs, like the weather, being multidimensional. Relevant dimensions for program evaluation include:

1. Appropriateness — Were "good" values utilized in defining the problem(s) which led to creation of the program? Is it alright to perform such interventions, regardless of their effects? The comment, "the ends don't justify the means" indicates program activity is inapproprite, in the opinion of the observer.
2. Adequacy — To what extent did the program eliminate the problem(s) which gave rise to the program?
3. Effectiveness — To what extent did the program have as much impact on the problem(s) as was intended?
4. Efficiency — To what extent did the relationship of program costs to effects meet expectations?
5. Side effects — What were the nature (in terms of desirable and/or undesirable) and quantity of program effects beyond those intended?

Is it possible to be aware of the nature of a problem which led to creation of a program — with the goal (aim, mission, purpose, objective, etc.) of reduction or elimination of the problem — and not evaluate appropriateness? I believe not. Each of us has values which we invoke automatically when faced with an idea, and for most such ideas, we care; we have preferences, which reflect our values.

But what if we don't know the purpose of the program? If we know of a program's existence, But not its purpose, we will impute a purpose or goal. We may know the name of the program — "the abortion program"; "the right to life program" — and from that impute a goal which we will deem appropriate or inappropriate. Or we may know the nature of the intervention, e.g., the pasteurization of milk. If we believe, as public health authorities in the early 1900s did, the purpose is to allow the sale of filthy milk, we might think the program inappropriate. But if we thought the purpose was prevention of communicable disease, we might judge the program appropriate.

Thus, it seems that anyone who knows the purpose of a program or knows something which allows imputation of a purpose, cannot avoid evaluating its appropriateness. The exception would occur when one had no value, no preference, for such a purpose. Do we sometimes, or even frequently, not utilize or act upon these evaluations? I think not. At the least, we will not voluntarily work for, or become clients of, programs judged inappropriate. At the most, we will devote our energy to eliminate, or change the purpose of, the program.

What about other dimensions of evaluation? Among people aware of a program, some will have direct, definitive, or "proven" (Stanley, 1964) evidence of program effect. But they may differ in the dimension measured, the protocol used, the size or composition of

the group of subjects measured, the evidence used to attribute program causality or the standards used for comparison.

What about those without direct evidence? Many will conduct a "presumptive" (Stanley, 1964) evaluation of adequacy and effectiveness. If they know of program activities, they will apply their hypothesis about relationship of such means to known or imputed ends. Some will even impute means on the basis of program names, and conduct a 2-step, presumptive evaluation. Do we sometimes or even frequently, not act on, or utilize the results of such evaluation in decision making? I think not. It may appear that program personnel are not utilizing the findings of an evaluation, but that will often be because they are acting in response to a different evaluation, or to a different interpretation of the findings.

I have tried to argue that program evaluation occurs, and those who do the evaluation do utilize their findings, to the extent possible. The person who sits in a chair and finds it too narrow, or too hard cannot avoid that evaluation, and will investigate, at least, the availability of alternative chairs. Whether action results will depend on the identification of a preferred chair, and the "costs" of moving. On a more "programatic" level, the therapist who has a lot of missed appointments will be hard put to not become aware of that fact. Thus evaluation will occur — a count of kept appointments and a comparison to total appointments — and if the therapist cares at all, will utilize that finding in exploring alternative ways to enhance appointment keeping.

Will there also be evaluation of the effectiveness of therapy on those who kept their appointment? I think so. But it will often be much more "subjective" and the therapist may well feel less certain of the validity of this measure. However, lacking anything better, the therapist will continue current techniques, or try different ones, based on this finding.

So if all this evaluation, and utilization of findings is going on, why do we find all the attention evaluation is being given? Why do we have the development of courses, of curriculums, of books and journals?

I believe it is because improvement is seen as needed, and possible. Occasionally, interested people doubt some aspect of their own current evaluations. Perhaps more frequently the concern is for an evaluation that will be accepted by a "significant other" — the budget committee, potential clients, etc. — who see the present evaluations as suspect or in error. So when we talk about evaluation, it is because of concern for improved evaluation, not whether evaluation (or perhaps, improved ability to decide which of current, conflicting evaluations should be trusted).

How can these improved evaluations come about. Will this desire for improved evaluations of programs lead to the development of a profession of evaluators? I think not; an adequate argument is based on the first criterion for a profession of "a monopoly over an esoteric body of knowledge which is considered important for the functioning of society" (Morell & Flaherty, 1978). The first is that many evaluations have been done which are widely considered adequate, in absence of a profession or professional evaluators.

Do we need the development of a profession of evaluators who will decide what values ought to predominate, and guide the definition of problems? Or, to provide the values to be used in judging the appropriateness of program interventions? I think not. Philosophers and theologians should continue to contribute to those judgments along with the myriad of others, and in democracies, we should continue to believe the individual, the client of social programs, should have some role in evaluating appropriateness of programs, often expressed through the choice of whether or not to become a client.

There is little criticism of the evaluation of the World Health Organization's smallpox eradication program. There has been little (but perhaps it's too early) criticism of the evaluation of the borderline hypertension control program (Hypertension Detection and Follow-Up Program Cooperative Group, 1979). These evaluations were performed by physicians, epidemiologists, statisticians, and others, not "evaluators" in any professional sense.

If evaluation as I have defined it above is accepted, I find it impossible to claim techniques of measurement, and of determination of causal relationships, as a monopoly for the evaluator. Too many other disciplines have a prior claim, at least as applied to their content area. We are more apt to find professionals from many fields specializing in evaluation.

If we need improved evaluation, but the development of a profession to do it is not the answer, where does the chemist, the social worker, the nurse, the engineer who has responsibility for evaluation but feels inadequate, turn? I believe there is a need for associations, journals, books, courses, workshops, etc. on evaluation where people from a variety of disciplines and professions can gain ideas to improve their evaluation work.

I believe the improvements will focus on areas such as:

- development of more valid, reliable, accurate, precise or acceptable measures, where current measures are unsatisfactory on one or more of these dimensions;
- development of cheaper measures where current assessment procedures are considered too expensive;
- development of random sampling schemes, with acceptable stratification levels and adequate sample size where current data are based on self-selected voluntary testimonials or otherwise biased samples,

or samples too small to yield population estimates of acceptable precision;
- development of a social and political environment to allow randomized experiments when that can be warranted, or conduct of improved quasi-experimental designs to allow more precise linkage of program efforts to effects than current evaluations will allow.

These I believe, are the major areas where improvement of current evaluation is needed.

In sum, it seems to me program-evaluation, as one form of a larger activity termed "research," can no more become the exclusive province of one profession than can research; it must be practiced by all members of all trades, disciplines, professions, or any other grouping of people whose efforts are aimed at making a difference.

Program evaluation is a process of answering several different questions about programs, and may be done well, yielding accurate answers, or poorly, yielding inaccurate answers. But be done, it will. And the results will be utilized. If inaccurate, poorer decisions about future programs will be made.

At times people "cheat" when evaluating programs, just as they sometimes do when doing research. But they know it. At other times people don't do as well as they would like, or could, because of lack of knowledge or skill. I believe there is a role for "content free" evaluation methodology resources, and for evaluation specialists within various "content" groups. But I can't see a professional "content free" evaluator.

Concern with valid, reliable measurement of phenomena, and discovery and quantification of causal linkages among phenomena do not belong to any single profession, but to all.

In the case of community wide programs, such as water fluoridation to improve dental health, few would quickly accept the professional evaluator as the "one hand" Boulding (1969) described:

> "And yet the politicians shudder
> To think of one hand on the rudder,
> Because nobody can agree
> On whose the guiding hand should be. . ."

REFERENCES

BOULDING, K. From Session II. The integration of European capital markets, the Ditchly Bank anthology. *Michigan Business Review,* 1969, *21,* 17.

COMMITTEE ON EVALUATION AND STANDARDS. Glossary of evaluation terms in public health. *American Journal of Public Health,* 1970, *60,* 1546-1552.

FLAHERTY, E. W., & MORELL, J. A. Evaluation: Manifestations of a new field. *Evaluation and Program Planning,* 1978, *1,* 1-10.

FOX, P. D., & RAPPAPORT, M. Some approaches to evaluating community mental health services. *Archives of General Psychiatry,* 1972, *26,* 172-178.

HYPERTENSION DETECTION AND FOLLOW-UP PROGRAM COOPERATIVE GROUP. Five year findings of the Hypertension Detection and Follow-up Program: 1. Reduction in mortality of persons with high blood pressure including mild hypertension. *Journal of the American Medical Association,* 1979, *242,* 2562-2571.

KIRKHART, KAREN E. Symposium: Making evaluation results useful; Knowledge utilization re-examined. *Evaluation News,* Fall 1979, *13,* 23.

MORELL, J. A., & FLAHERTY, E. W. The development of evaluation as a profession: Current status and some predictions. *Evaluation and Program Planning,* 1978, *1,* 11-17.

STANLEY, D. T. Excellence in the public service: How do you really know? *Public Administration Review,* 1964, *24,* 170-174.

TORNATZKY, LOUIS G. The triple threat evaluator. *Evaluation and Program Planning,* 1979, *2,* 111-115.