Evaluation methodology in a large scale program for high school students: delinquency, drug abuse, and attitude change.[1, 2]

William F Soskin, Ph.D.
Director, PROJECT COMMUNITY
University Extension
University of California, Berkeley

Robert L. Fisher, Ph.D.
Director of Evaluation and Research
PROJECT COMMUNITY
University Extension
University of California, Berkeley

PROJECT COMMUNITY
2508 Hillegass
Berkeley, California 94704
(415) 642-1855

[1]Paper presented at the National Conference on Criminal Justice Evaluation, Washington, D.C., February 21-24, 1977.

[2]The research reported in this paper was funded by grants from NIDA and NIMH.

Introduction

This paper summarizes a number of evaluation issues common to prevention programs in the areas of substance abuse and juvenile delinquency. The issues and research results we will discuss today evolved from our on-going evaluation of an experiential program designed for high school students and teachers called Project Community.

Problems of Evaluation

1. Program evaluation is almost always the interaction of two simultaneous experiments, one having to do with the invention and application of some new intervention process, and the other having to do with the invention and application of suitably sensitive measuring instruments to demonstrate change. If after some period of application of the service the evaluation unit turns up objective evidence of change in the behavior of subjects, then both groups can claim success. But if the results are negative, it is rarely clear which of these two experiments is not working. In the overwhelming majority of cases when results are inconclusive or negative, it is assumed that the intervention procedure has had little or no effect. Yet, at least as often it can be asserted that the techniques of measurement might have been inadequate for the task.

The reasons for this are probably many, but here is one that deserves our attention particularly. The evaluator is often the master of his craft, one that is backed up by a highly complex and seemingly arcane technology. It isn't easy to out-argue an expert who draws on complex mathematical formulae; mathematics is, after all, an exact science. One doesn't usually consult a physician and then argue with him about diagnosis and treatment. Or at least, one isn't expected to. So it is often assumed that when the evaluator brings his tool kit of skills and methods to bear, he is like the physician; he is the expert, and his status is respected. Too often, too, under

the guise of objectivity, the evaluator holds himself aloof, not recognizing or acknowledging that his role, too, is that of an experimenter.

A related problem is that under the circumstances of most evaluation, the evaluator doesn't have time to adequately carry out his side of the research. He often assembles and applies a battery of instruments developed by other investigators for somewhat different purposes. This lack of time brings us to the second general problem.

2. Federal agencies have the obligation to commit their limited funds in the most creative and promising and democratic ways. They have many claims on their money. Grants must be spread around. Many seemingly good ideas deserve support. Hence, grants tend to be of limited duration, usually two or three years. Many prevention programs have only that much time and no more in which to stand trial so-to-speak. Yet, a complex service program in a community is very different from a carefully controlled laboratory experiment, and it may well take several years simply to put a new practice into smooth operation. If the evaluator is a part-time person or team doing a simple Time-1 -- Time-2 measurement of change, there is hardly time at all for the evaluation unit to discover and overcome the shortcomings of its own procedures.

If we look at the many interrelating institutes like the Sloan-Kettering foundation, the National Cancer Institute of N.I.H. and the scores of large research laboratories in major schools of medicine that have been supporting large teams of cancer researchers for 30 years or more, we could hardly produce a reasonable estimate of the number of negative-outcome experiments that have been carried out by thousands of investigators and their aides, each adding some small increment to the abandonment of one theory or the support of another, the uncovering of a hitherto unsuspected relationship, the recognition of a fresh faint clue, a new tool, and improved microtome, the invention of a larger electron microscope. No single laboratory, no single investigator or team bears a heavy responsibility for trying to demonstrate that cancer can be cured, now. Each need only add a single useful increment to the corpus of knowledge; that is a great enough responsibility.

But scores of two- or three-year projects in delinquency prevention or substance abuse abatement every year must disband their teams and close

their doors after the first expectations fail to materialize magically within the expected time limit and there is no fresh glamour with which to attract new grants.

This is in no sense a complaint from a failing enterprise. Rather, we are drawing upon an eight-year history of work in a single area to share some of our successes, observations and experiences with the difficult task of carrying out evaluations of "prevention" programs.

3. Another main area of concern is the character of the measuring devices used in evaluation. We all know about indirect measures, and about such devices as testimonials, diaries and films as "secondary" outcome measures. But the bread and butter of evaluation in the behavioral sciences and in intervention program evaluation continues to be so-called "objective" measures, which means changes in scores on questionnaires and rating scales or similar devices.

In our case, and in many studies like ours, the principal measuring instruments are assembled first out of collections of well-standardized instruments drawn from various catalogues and -- and this is very important intended for a purpose somewhat different from the present application. We, too, use a battery of instruments, but of the seven or eight measures we adopted that were designed by other investigators for whatever purpose -- all of them of reasonable reliability -- all but one or two have been long since discarded. Each year enabled us to observe the "behavior" of the instrument with our samples of students, in our particular problem context. Over time, one by one these instruments failed to provide adequate descriptive information or failed to discriminate. New variants had to be created. Or instruments had to be devised that more sensitively reflected our own growing experience with the behaviors of our subjects.

In this respect we have found two important sources of information that contribute to our evaluation procedures in ways that borrowed standardized tests often cannot. These two are closely related, and in fact may represent terminological differences for essentially the same process. They are: Our own clinical experience and our own field observations. Like many of you, we are fortunate in having a staff of talented and sensitive people who have an

opportunity to spend hundreds of hours listening to what our subjects say, and interacting with them in a number of different social contexts. These experiences, fed back to the research group in staff meetings and other ways, continually spawn new items and new analyses. Even if nothing else changed over the years of Project Community's existence, the evaluation instrument itself would show a steady evolution toward what we think is increasing descriptive and discriminative power. The other important source of information, of course, is the continual study of item behavior or total-scale behavior in more and more different samples of subjects.

Thus, the "effectiveness" we might have discussed five years ago is quite different from the effectiveness we speak about today. And if we have not solved the problem of discovering a solution to the problem of substance abuse we remember that in the late Forties and early Fifties medical promoters launched an enormous Manhattan Project of their own -- a multi-million dollar cancer research "crash program", generously supported by Congress and the nationwide foundations and public fund-raising appeals -- the 10-year time limit on that endeavor passed nearly 20 years ago and the search goes on -- although with a much deeper understanding of the problem and with better tools and better theories and more trained researchers.

The Problem of Samples

Which groups to study, what samples are available, the characteristics of sub-groups needed for progressively finer analyses are critical aspects of prevention program evaluation.

In our own study, ever since this program began we have been criticized by some and shunned by others because of the sample of students we have studied. In the years that we were operating an after-school program in Berkeley we were repeatedly criticized, even threatened, for running an "all-white" program. You should know that Berkeley, with a population of approximately 120,000, has a high school population that is about 47% White, 47% Black and 6% Chicano, Native American and Oriental. Our program spontaneously drew about 95% White students. Physically we were located in a substantially White neighborhood -- across the street from the University campus. So we hired a bus to shuttle between the school and the program in hopes of

recruiting more Blacks. Our staff included several Black persons and a Black professional consultant. What we discovered during the pressure of those political days was that many young Black leaders in Berkeley didn't want Black high school students in our program. Understandably, they wanted their own programs. Furthermore, some severely criticized us for spending money on Whites despite our efforts to conscientious efforts to recruit Blacks -- who are justifiably sensitive to the political ramifications of sampling. Racial and minority sample problems continue to plague us. At the present time the racial and ethnic composition of our program -- both teachers and students -- is a direct reflection of the composition of school teaching staffs and the communities they serve. Yet we are obliged to state very clearly the limitations on the representativeness of our findings.

A different kind of sample problem that bears directly on the effort to evaluate prevention programs is one that is familiar to all evaluators -- viz., the size of the sample. When the size of the N depended merely on the number of individuals one can recruit to take a test, or who can be induced to undergo some procedure for one or two afternoons, this factor can often easily be overcome, but when one is dealing with a program that requires long time commitments, and when the number of subjects one can serve is dependent upon costly outlays for staff, then prevention program evaluation is often complicated by the size of the grant. In our own program in the schools the availability of teacher time and the size of our staff -- both affected by budget -- dictate how many students we can accept in any given semester, and that number is <u>small</u>. During the first year of our in-school program only about 60-70 students participated. That was all the teacher time available for this program. Of this group, not all finished the program. Of those who did, not all completed both the pre- and post-questionnaires. More than that, the available samples in the different schools that had agreed to participate were so different on a number of significant dimensions that the data could not be pooled. In fact, only after the third year has the total number of participating subjects become large enough so that we can begin to make the necessary breakdowns of data and still be assured of enough subjects in each cell to make our results statistically meaningful.

Then there is the problem of the range and variety of subjects available for study. If all your subjects represent an extreme degree of involvement in the function measured -- say school failure or delinquent acts or substance abuse -- you might never discover that the program has only a slight or negligible effect on extreme cases but a strong effect on moderate cases. Or, if through the unannounced interventions of some school counselors the majority of your program participants in a given school turn out to be those most "turned off" students for whom counselors can't find any other suitable place, a realistic measure of the utility of your program will have been thoroughly undermined.

A third closely related sampling issue warrants attention. Nearly any large metropolitan or suburban high school has at least a couple of hundred young people whose personal lives -- often for reasons beyond their control -- are so stressful that they simply manage as best they can from day to day. Having a significant impact on the lives of these students is a challenge to all the social services and helping professions. For many what often is needed is a major change in life circumstances -- removal from a bad family life, a chance to associate with different kinds of peers, different school alternatives, or an opportunity to earn some money. In most schools, where very little in the way of counselling services exist, these people quickly turn up in large numbers in any program that seems to offer some support. And so they should, but a very high concentration of such students places an extremely heavy burden a program to show clear significant differences in group mean scores over a relatively short period of time while the major environmental stress continue to operate in full force.

The Criterion Problem

Last in this presentation, we wish to address the problem of evaluation criterion. The criterion problem, has plagued all prevention program evaluations. Whereas a scientific experiment demands a clear and precise criterion against which to measure effects, we now believe it imperative that prevention programs build in at the outset a number of kinds and levels of criteria, both because of the diverse demands of different "consumer" groups and because

subsidiary outcomes may point to unanticipated promising areas of further development. This is where serendipity has its full play. As to the first of these, the diverse demands or expectations of various "consumer groups", we would like to point out some of the criteria of effectiveness proposed to us implicitly or explicitly by the expectations of these various groups.

Currently our school programs are chiefly supported by N.I.D.A. grant, so it is reasonable that we should propose as one of the chief criteria of success some notable decrement in substance abuse. Rarely has it been possible for us to show major shifts in drug abuse patterns from $Time_1$ to $Time_2$ on gross comparisons within the group of participants, or between participants and controls. What we have found as our instruments become finer and our analyses more subtle is that significant changes not evident in gross comparisons of mean scores can be shown to occur in particular sub-groups, i.e., in persons with certain characteristics. Thus, we can show that drug abuse is consistently associated with such factors as low trust of others, negative attitudes toward parents, low self-esteem, poor peer relations, etc. We are also able to show, for example, that students with a high level of trust tend to diminish their drug use in the course of a semester whereas those with low trust scores do not, and we are able to show that for some of our groups of students significant changes in trust of adults occurs over the course of a semester. In other words, by small increments our data disclose the complexity of the problem, and this, too -- as the cancer people have long insisted -- is important progress, even when the "cure" eludes you.

As further considerations with respect to criteria, the granting agency and its advisory panels of scientists or prominent practicioners can usually be counted on to entertain the broadest possible view of acceptable criteria. The groups more immediately involved as users of the service, on the other hand, have much narrower criteria.

In our own case, the students might derive their criteria from the following kinds of questions:

(1) Is the program interesting and enjoyable?

(2) Can the adults in the program understand us and are they trustable allies?

(3) Does the program make me feel better?

(4) Am I learning something that is helpful and important to me?

(5) Am I accepted in a group of respected peers?

For the program staff, being able to give evidence that students answer positively to most of these questions is basic affirmation that the intervention is perceived as important and is valued by adolescents, and in turn these are important secondary evidences of effectiveness. Surely, if the users don't like the program, it is unlikely that it will survive.

School administrators have still another set of criteria. They are likely to derive theirs from questions such as the following:

(1) Is the proposed activity one that is acceptable to parents and other school authorities?

(2) Does the program interest a significant number of those students who seem uninterested in the regular course program?

(3) Will it improve attendance?

(4) Will it improve general attitudes toward school and educational goals?

(5) Will students regard it as useful and effective in their immediate daily lives?

(6) Will a significant number of teachers accept or oppose the program?

Parents will evaluate the program in still other terms:

(1) Is it a program that increases my son's or daughter's interest in school?

(2) Does it appear to relieve some of the stresses of adolescent development?

(3) Does the program inculcate values or standards or goals consonant with my own?

(4) What potential harm can come to my son or daughter as a result of participating in the program?

For the School Board there are of course very important other considerations, like:

(1) Is this program expensive?

(2) Does it require new or additional resources?

(3) Will it improve school performance of marginal students and hence improve our ranking in State evaluations?

(4) Will it significantly reduce aberrant or anti-social behavior?

Questions similar to these are asked by the parellel agencies and service users in other kinds of prevention programs -- by City officials, hospital authorities, and agency trustees. It is certainly advisable for any preventive program to have several levels or kinds of criteria on which to base its evaluation, for the more ways in which the programs' outcomes can be interpreted to the several consumer groups, the more likely it is to receive continued support. And, in turn, the more likely it is to refine measurement techniques to address the complex questions of program evaluation.

END