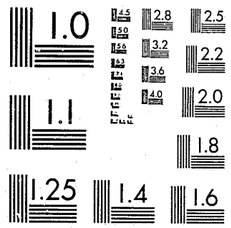


National Institute on Drug Abuse

WORKING WITH EVALUATORS

A Guide for Drug Abuse Prevention Program Managers

U.S. DEPARTMENT OF HE TH N

Working With Evaluators

A Guide for Drug Abuse Prevention Program Managers

Edited by
John F. French
Court C. Fisher
Samuel J. Costa, Jr.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Alcohol, Drug Abuse, and Mental Health Administration

National Institute on Drug Abuse
5600 Fishers Lane
Rockville, Maryland 20857

This volume, part of a series of National Prevention Evaluation Resource Network publications, was developed for the National Institute on Drug Abuse by the Single State Agency of New Jersey under Contract Number 271-81-4911. William J. Bukoski, Ph.D., served as the NIDA project officer.

John F. French, Court C. Fisher, and Samuel J. Costa, Jr., are with the New Jersey Department of Health, Alcohol, Narcotic and Drug Abuse Unit.

DHHS Publication No. (ADM) 83-1233
Printed 1983

Cover Illustration · Noël van der Veen

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402

FOREWORD

Working With Evaluators represents a significant advance in the development of scientifically tested prevention programs that meet the needs of parents, schools, youth and communities. This volume has been written to assist prevention program staff to work cooperatively and effectively with evaluators and researchers to apply their skills, knowledge and sensitivities in the design and implementation of noteworthy evaluations.

The prevention field has taken significant strides forward relevant to evaluation by breaking through the resistance and fear of evaluative findings that have proven to be so typical of social programing. In contrast the field of prevention clearly recognizes and accepts the tenet that if the field is to continue to develop and to emerge in the 1980's as a scientific discipline, this evolution will be based in part on the knowledge gained from evaluative research and program evaluation.

The development of this volume and more importantly the National Prevention Evaluation Resource Network (NPERN), cogently illustrate the many positive benefits to be derived from joint State-Federal projects. As a result of the consortium of States (Wisconsin, New Jersey, Pennsylvania) involved in that effort, a system for evaluation had been created that is sensitive and responsive to the unique evaluation needs of State and local prevention programs without imposing constraints or inapplicable standards. Just as sound evaluation results from the partnership of a well trained evaluator and a skilled program staff, so too will effective prevention programs result from the partnership of States, communities, families, parents and the Federal Government.

William J. Bukoski, Ph.D.
Research Psychologist
Prevention Research Branch
Division of Clinical Research
National Institute on Drug Abuse

PREFACE

The real pleasure in (disease) prevention
is in watching nothing happen.

Donald Millar, M.D.
Centers for Disease Control
(New York Times, Jan. 20, 1980)

The real pleasure in evaluation is in watching -
thus helping learn to make - it happen.

Anonymous

And the real pleasure in creating this monograph on prevention evaluation was in working with and through a stimulating network of people. In addition to the authors and editors, many people contributed significantly to help shape the monograph.

Early outlines of the monograph were reviewed in depth by David Twain, Rutgers University Graduate School of Criminal Justice, and Nancy Kaufmann, Wisconsin Bureau of Alcohol and Other Drug Abuse. A final outline was prepared in a two-day intensive work group attended by most of the contributing authors, and editorial staff.

Following submission of several chapter drafts by each contributor, a five-member national consumer review group of prevention and evaluation practitioners was convened, selected with assistance from the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, the National Council on Alcoholism, the Center for Multicultural Awareness, and many individual prevention specialists.

The review group included Barbara Bell of the New Jersey Division of Narcotic and Drug Abuse Control; Barbara Kline of the Rock Island (Illinois) County Council on Alcoholism; Patrick Ogawa of the Japanese-American Cultural and Community Center in Los Angeles; Carol Stein of the National Federation of Parents for Drug Free Youth; and Richard Stephens, Cleveland (Ohio) State University.

The consumer review members each independently read and critiqued the first full draft of the monograph, then met with the editors as a group to consolidate suggested changes, and reviewed a second draft incorporating their suggestions. Hugh Cline of the Educational Testing Service provided an independent technical review.

John F. French Court C. Fisher Samuel J. Costa, Jr.

New Jersey Department of Health
Alcohol Narcotic and Drug Abuse Unit

CONTRIBUTING AUTHORS

Chapter 1: **Court C. Fisher**
New Jersey Department of Health
Trenton, New Jersey

Chapter 2: **Samuel J. Costa, Jr.**
New Jersey Department of Health
Trenton, New Jersey

Chapter 3: **Richard P. Neuner**
Minnesota Institute
Anoka, Minnesota

Chapter 4: **Leona Aiken**
Temple University
Philadelphia, Pennsylvania

John F. French
New Jersey Department of Health
Trenton, New Jersey

Chapter 5: **Michael Klitzner**
Wisconsin Bureau of Alcohol and Other Drug Abuse
Madison, Wisconsin

Ileana C. Herrell
Montgomery County Health Department
Rockville, Maryland

James M. Herrell
Montgomery County Health Department
Rockville, Maryland

Chapter 6: **Royer F. Cook** (Case Study 1)
Institute for Social Analysis
Reston, Virginia

Eric Schaps (Case Study 3)
Pacific Institute for Research and Evaluation
Lafayette, California

Daniel Solomon (Case Study 2)
Developmental Studies Center
San Ramon, California

Chapter 7: **Allan Cohen**
Pacific Institute for Research and Evaluation
Lafayette, California

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

(What It's Mostly All About)

Evaluation is about

participation
empowerment
learning
survival.

These are not words we normally associate with evaluation, but they are elements of the basic purpose and message of this monograph: to foster participation, empowerment, learning, and survival in alcohol and drug abuse prevention programs.

This is a monograph about evaluation for managers (and other decisionmakers) of prevention programs. It is a product of the National Prevention Evaluation Resource Network (NPERN).

NPERN is a program of the Federal Department of Health and Human Services, National Institute on Drug Abuse (NIDA). In 1978 the Prevention Branch of NIDA started NPERN to improve the number and quality of evaluations conducted by and about drug abuse prevention programs. The National Institute on Alcohol Abuse and Alcoholism (NIAAA) later added its support to encourage greater access by alcoholism prevention programs to evaluation resources.

NPERN works primarily by bringing experienced evaluators together with alcohol and drug prevention programs to help the programs meet evaluation needs. This direct on-site technical assistance was provided first in a 1978 pilot project in six States. A larger scale national technical assistance phase operated through 1981.

As part of NPERN's program several publications were also written and published. A Handbook for Prevention Evaluation is a summary of evaluation knowledge and technique applied to the prevention field and is written primarily for evaluators. This monograph, Working with Evaluators, is a companion to the Handbook and is designed primarily for prevention program managers.

Although it is written with the assumption that you—as a program manager—will have direct access to evaluation consultants through the NPERN network, the monograph will also be useful to managers in working with evaluators generally. Indeed, it can help you to understand, design, and conduct your own program evaluations even if you have no outside assistance and expertise to help accomplish this.

As a user of the monograph, you are encouraged to read—or skim—it through at least once from beginning to end. Each prevention program manager will bring different sets of experience, interest, and need to this monograph, and you will each find different chapters or sections to meet your interest. Some redundancy is built in from chapter to chapter to maintain continuity, but the monograph as a whole is shaped by the following structure:

Chapter 2, A Model for Program Change, introduces a conceptual framework for evaluation as part of a nine-step continual process of program planning, feedback, and change. Evaluation of program process, outcome, and impact is introduced, along with ways to categorize information and target areas. Chapter 2 lays the groundwork for more detailed discussion of the process and content of evaluation in later chapters. It should be reviewed by every monograph user and is must reading for program managers with little evaluation background.

Chapter 3, Program Issues in Prevention Evaluation, shifts focus to highlight some characteristics of alcohol and drug abuse prevention and its programs in relation to the evaluation model of chapter 2. It presents four major questions that prevention program managers must ask to participate effectively in evaluation.

Chapter 4, Evaluation Issues in Prevention Programs, puts the program manager inside the evaluator's head, to understand basic design and methodology questions that must be considered in conducting any evaluation. By constructing and critiquing one case study of a poor evaluation, chapter 4 highlights technical issues that managers and evaluators must examine together to assure useful evaluation. Chapter 4 also describes and comments extensively on basic quantitative and qualitative methods and provides an introduction to cost-benefit analysis. Chapter 4's focus is on the content more than the process of evaluation and may be useful as a continual reference for prevention program managers.

Chapter 5, Preparing for the Evaluation, elaborates the 9-step model introduced in chapter 2. It takes program managers through each step in detail, emphasizing their responsibility and participation with the evaluator. Chapter 5 can be read and used as a checklist for good evaluation process.

Chapter 6, Case Studies in Prevention Evaluation, ties the earlier, more didactic, discussion of evaluation content and process into three case studies. Emphasizing real-life process, the case studies focus on communication between program decisionmakers and evaluators, and the relationships among these interpersonal communications, program realities, and evaluation needs that encourage or hinder useful evaluation.

Chapter 7, Politics and Science in Prevention Programing, also uses case material but focuses on the importance of the program's external political context for the success or failure of both the program and its evaluation.

Overall, this monograph discusses evaluation as participation, empowerment, learning, and survival. These themes flow from the experience and understanding of evaluation shared by the authors.

Participation is fundamental. Starting in chapter 2 which describes evaluation as part of a process of continual program change, the need for program managers and evaluators to collaborate is emphasized. This is not simply a matter of good personal relations but follows from the nature of evaluation itself.

Fundamentally, evaluation is a way to describe selectively and then to judge the value of something—in this case your prevention program. The political and organizational history of evaluation reinforces an ideology—and a reality—that this process of description and judgment is "scientific," carried out by experts on less expert people and programs.

This monograph affirms that science and expertise are indeed involved in the evaluation of prevention programs. But it affirms something more—that evaluation is not simply "objective" science composed of facts outside your own interest and influence. Good (and bad) evaluation, like good (and bad) science, is fundamentally a human activity shaped by the intentions, knowledge, and values of the people who do it. That includes you as a prevention program decisionmaker. As manager, your primary responsibility is continually to define and to carry out the ends and means, goals and methods, of your program. Evaluation is an extension of this same responsibility at a second level. To the extent that you contribute to defining the goals and methods of an evaluation, you will influence, if not control, its process and outcome. Participate!

This monograph is also about empowerment—yours. One intention is to provide you as a prevention program manager with enough of the "stuff" of evaluation, its values, language, and technique, that you can participate intelligently and effectively with evaluators and other decisionmakers in the conduct of evaluation. The monograph won't turn you into a full-time evaluator. It can help you become a better contributor to and user of your own program evaluation, and thereby an even better manager.

Although technical aspects of evaluation are discussed throughout the monograph, chapter 4 contains the most concentrated discussion. As you delve into this, remember another fundamental characteristic: evaluation is about the certainty and uncertainty of what people know and can know about the world, including prevention programs. Evaluation is about reducing the uncertainty of what we know. All the more technical aspects of evaluation, including the most abstract, complex, and specialized scientific or mathematical issues, are fundamentally about identifying different kinds of uncertainty and reducing it. Evaluation is also about understanding that any approach to reducing uncertainty in the real world has accompanying costs. Keep this principle in mind as you use the monograph to increase your own knowledge and power.

Empowerment comes not only from learning the content of evaluation, but from participating in the evaluation process. In chapter 5, the authors take you step-by-step through the process of preparing for an evaluation and working with an evaluator. Both this chapter and the chapter 6 case studies try to capture the feel for assertive, intelligent give-and-take between program manager and evaluator that is the hallmark of good evaluation process.

Participation and empowerment are twin aspects of your process as a manager in program evaluation. Learning and survival are likewise twins—but they are the goals. Evaluation contains a natural tension between acting in the world based on current belief and knowledge and remaining open to new experience and knowledge that may change belief and action in the future. This is the tension between growth and change, continuity and status quo. It is a tension and balance that affects each program, the field of alcohol and drug abuse prevention in general, the larger society, and the political economy.

Chapter 3 explores some of the prevention program issues, including changes in the prevention field itself, that contribute to the change/survival dynamic. Chapter 7 likewise focuses explicitly on the survival value of program evaluation, emphasizing how recent political and economic changes have shifted the human service emphasis from learning, change, and growth to a more survival orientation.

How does your prevention program fare in the midst of these changes? What criteria are your funding decisionmakers using to divide a probably shrinking pie? Assuming you're still in the pie, what criteria and information are you using for your own program and budget decisions? This, too, is the stuff of evaluation. You may even find that asking these questions—challenging your own and your program's actions and beliefs—can become as interesting as the actions and beliefs themselves. To incorporate the questioning "evaluator" perspective may contribute to your becoming a more committed doer and manager.

Read the monograph through, pick and choose what interests you most, read and use again. We hope you find the monograph as useful as we found it fun to create. Try it!

CHAPTER 2: A MODEL FOR PROGRAM CHANGE

(It Goes Round and Round and Never Stops)

impact process outcome feedback
target group goals objectives
cost effectiveness utilization
a red herring program development
indicators research design
validity reliability random sampling

The above terms, among others, appear numerous times throughout this monograph. Such is the language of the prevention evaluation field. If evaluation is used not simply as pruning shears for funding agencies, but as a means of aiding program development and renewal, then most of the terminology can become part of the everyday program vernacular. It would be best for program decisionmakers and evaluators to speak the same language.

The purpose of this chapter is to describe an evaluation model based on constant feedback about various aspects of a program to promote continual program development. But first, we must discuss the need for a model, and following the model description, how to tie the model to various phases in a program's development.

NEED FOR AN EVALUATION MODEL

Funding for health and human services is always tight but has become more so in the recent past. Drug and alcohol abuse prevention, as a new kid on the block of human services, especially needs to prove its worth to various sources of pressure and funds. Taxpayers, Government agencies, foundations, and others all seek more effective evaluation of programs in the human services field. "More effective" implies that past evaluations have been lacking in effectiveness. This is a justifiable implication, but the need today is to build from past problems rather than to tear down past evaluations. The potential for evaluation research is enormous. The field itself has contributors coming from the many scientific disciplines involved in the evaluation of human services—psychology, sociology, anthropology, political science, statistics, operations research, systems analysis, economics, and computer science. The evaluation of any one human services program (for example, a substance abuse prevention program) can draw on the growing literature from all these fields.

One of the most often criticized aspects of evaluations is the underutilization of the results. One of the reasons is that the questions of the decisionmakers who could best use the results are not always considered during the early phases of the evaluation. If decisionmakers are not asked what information they need, the evaluation may not even address the appropriate issues. The audience of decisionmakers we refer to could range from funding sources or key program administrators to program staff or community activists. For example, if a funding agency wants a strict cost-efficiency analysis, the program manager's interest in which aspects of the program actually help the participants the most—regardless of cost—can go unnoticed and unexamined. Conversely, if an evaluation is limited to an internal investigation of the success of

different approaches to prevention with no interest in economic realities, the project's administrators may have difficulty in providing the type of information (for example, bottom-line costs) that some funders demand.

Because the importance of using evaluation results cannot be overemphasized, it is repeatedly stressed throughout this monograph. If the results of an evaluation are ignored, or never reach the critical decisionmakers, the evaluation plan was not well thought out or implemented. In the case of the evaluation model presented here, utilization of results will be seen as a basis for both program survival and improvement.

A program manager is not expected to keep abreast of developments and techniques in the evaluation field, of course. This is why there are evaluation consultants. That a good evaluator will be aware of the appropriate techniques and applications for various methodologies should be taken for granted, but one of the problems in the past has been methodological deficiencies. An evaluator has to be flexible, willing, and able to divorce himself from his favorite method if it does not fit the situation at hand. But how does a manager know whether or not the evaluator is suggesting an appropriate method? By being a critical consumer of evaluation services. A good manager will demand to be informed of the potential uses and limitations of alternative designs for the program evaluation. A good manager needs to know the costs (financial and informational) of one method compared to another. Even if the manager has no control over the conduct of an evaluation—as in the case of a funding agency hiring an outside evaluator with carte blanche to find out only what the agency wants to know—the manager has the right to know what is being looked at and how it is being done. Ideally, a good evaluator-manager team will develop, pooling their knowledge of the theoretical, the applied, the ideal, and the practical aspects of both prevention programs and their evaluations. This monograph and the previously published Handbook for Prevention Evaluation (French and Kaufman 1981) encourage team effort.

To build cooperation, a consistent frame of reference and language is needed for decisionmakers and evaluators. The evaluation research model developed several years ago under the auspices of the National Institute on Drug Abuse's (NIDA's) Prevention Branch (Bukoski 1979; French and Kaufman 1981), building on work by Waller and Scanlon (1973) and others, provides the context for the evaluation issues, strategies, and methodologies presented here.

No rigid, standard form of evaluating prevention programs is proposed. Rather, a flexible model is presented to encourage the incorporation of new developments in both prevention programing and evaluation methods. This framework provides a rational approach to program evaluation and shows how evaluation methods can be incorporated into a program in a manner most helpful to the prevention program itself.

THE EVALUATION MODEL

This model can be used with any alcohol or drug abuse prevention approach. It features three levels of evaluation:

process, outcome, and impact

categorizes information into three types:

descriptive, comparative, and explanatory

and can focus on one or more of four major target areas:

individual, program, service system, and societal.

These three evaluation parameters—level, information type, and target area—are discussed below.

Level

Each level of evaluation (process, outcome, impact) has its own set of indicators and methodologies. Ways to measure what is going on—methodologies—differ among the three levels, as do the things that are measured—indicators. The three levels are discussed below, with a brief overview of all three followed by a more thorough discussion of each.

Process evaluation is a thorough description of the various aspects of a prevention program. It attempts to present a complete picture—the dynamics and characteristics of an operational, ongoing prevention program. Process evaluation examines the target population, the personnel operating the program, the services delivered, and the utilization of resources for program components. These and other aspects of the program all provide indicators at this level of evaluation.

Outcome evaluation is what most people think of when evaluation is mentioned. It is concerned with measuring the effect of a program on the people participating in it. Outcome evaluation attempts to answer the question: "Has the program had a significant effect on participants and is that effect in the desired direction?" In essence, this level of evaluation is an attempt to determine if the program has met its objectives in producing changes in perceptions, attitudes, behaviors, or other effectiveness indicators among its targeted client group.

Impact evaluation examines the total effect of prevention programs on the community as a whole. The key word here is community, which may be defined as a school, neighborhood, town, city, State, etc. Community-wide indicators such as incidence and prevalence of substance abuse, related criminal activity, and institutional/societal policy and change are measured through methods such as epidemiologic studies or community surveys. The attempt is made to gauge the impact of a program operating over an extended period of time or of several programs operating within a specified geographic area.

The three levels of evaluation are not mutually exclusive. Rather, they can be viewed as successive phases in the development of information in a comprehensive evaluation effort.

Process evaluation.—The information gathered during this evaluative phase reflects all of the inputs into a program, the patterns in which these inputs interact, and the various transactions and interactions that take place within a program. Important process information includes the theory on which the program operates, needs assessment, policy development, program design, and the characteristics of program clients, staff, physical plant, decisionmaking structure, and financial resources. These types of data can provide continuous feedback to use for internal monitoring which can help guide and direct resource allocation, organizational decisions, and ongoing program development.

Process information can also contribute to accountability and replicability outside of, or external to, the program. How can process information from different programs be compared? One cannot simply compare programs without considering their operating contexts. These contexts are, themselves, part of the process information. By categorizing this information into four general areas—human resources, physical resource variables, contextual variables, and program specific variables—it becomes easier to identify variations between or among programs.

Human resources include all client and staff variables affecting the program. The number and description of clients served, staffing patterns, qualifications of staff, and attitudes and behaviors of both clients and staff are all considered human resources of a program.

Physical resource variables include descriptions of the physical plant, equipment, and materials and the program functions and activities which utilize these resources. Financial resources and expenditures are important program inputs which also provide a basis for cost analysis.

Contextual variables describe the community and institutional environments in which a prevention program operates. These directly affect the workings and effectiveness of the program. The demographic and socioeconomic makeup of the community are important factors, as are community attitudes and rates of various social problems (e.g., arrests and substance abuse related medical episodes).

Program-specific variables can be roughly divided into organizational structure, program service delivery, and participant/staff/program interactions.

Organizational Structure.—An analysis of an organization can yield important information regarding lines of authority, communication, and decisionmaking as well as the history of the program. For instance, there may be important differences between a freestanding prevention program and one that is part of a larger organization. Over time, most facets of an organization can be expected to change, and a description of the evolution of the current structure—and plans, if any, for future change—is very important.

Program Service Delivery.—Information regarding program service delivery includes the needs being addressed, the assumptions/theories underlying the particular prevention strategy, and actual program practices. The last involves the structure of delivery as well as content. Is it a sequence of presentations or sessions or is it a one-time delivery? Are the sessions scheduled in advance or given on demand? Are the

timing and structure of delivery the same as that originally planned? The services actually delivered need to be looked at in relation to the program's theoretical base in two ways: first, does the program actually carry out the planned prevention strategy, and second, do the services respond to the assessed needs? As discussed later in this chapter, the actual program delivery may deviate from the intended delivery at several phases in a program's development.

Participant/staff/program interactions.—Participant/program interactions include referral or selection procedures, client expectations, and the time and quality of participation. Regardless of modality and program, some identification or referral of clients is needed. It could be a formal referral network or simply membership in a group identified "at risk"--for example, junior high school students in a particular school district. Similarly, all participants have expectations regarding the program and its potential effects on them. These expectations influence the degree or quality of participation in the program. Someone with less motivation would not be expected to invest as much energy as someone who wants to gain as much as possible from the program.

Participant/staff relationships involve both the frequency and duration of interactions as well as the quality of contact between clients and staff members. Counts can be obtained and examined relatively easily; qualitative assessments are more difficult. Client and staff perceptions of the "what, where, how, and why" of the interactions are important, as is the comparison between these perceptions.

Staff-staff and staff-program relationships can be examined to see how staff get along, work together, and share common goals. Absenteeism and turnover rates can highlight problems. Also of importance is the congruence between intended and actual staff roles as well as the staff's expectations for both the overall program and individual roles within it.

To summarize:

process evaluation is a fancy way of
answering the question
"What's going on?"

in a new program, process evaluation
is the **only** way to know what's going on,

and in any program, process evaluation
tells you if what's going on is
what you **wanted** to go on.

Outcome evaluation.—Information gathered during this phase usually addresses specific program objectives concerned with changing participants' behavior, attitudes, values, or knowledge. The ultimate goal of all prevention programs is the reduction of drug and/or alcohol abuse. However, depending on the theory underlying the program, a more immediate objective may be something like "increase self-value" or "improve social skills." These objectives are theorized to be associated with decreased substance abuse. In other words, the program attempts to reduce the risk inherent in some state such as low self-esteem, poor school performance, or maybe simply ignorance about drugs and alcohol, thereby decreasing future substance abuse.

To assess whether program objectives have been met, they must first be identified. This is not always as easy as it sounds. Using process evaluation, both intermediate and ultimate objectives can be identified by examining the development of the program. Even if a full-scale process evaluation is not being done, some process information must be collected to identify the program's objectives. What was the problem or need leading to the program's initiation? How does the program purport to alleviate the problem and meet the need? What effect does the program hope to have on its participants? Will it change attitudes or change behavior in a more immediate way? Does it attempt to clarify values or increase knowledge of risks? How long must clients participate in order to benefit from the program? How long are program effects expected to be sustained?

Many program managers may find such questions simple and the answers clear. These managers will also have a good understanding and clear statement of program objectives. However, some managers will not know their programs' objectives immediately. And the objectives of some programs are not easily specified. Thus one benefit of an evaluation may be the learning process undertaken to articulate the objectives of the program.

Most programs have multiple objectives, all of which need to be identified. Different interested parties, whether staff, participants, funding sources, or others, may emphasize certain objectives more than

others. All these factors need to be considered when objectives are listed. If some important objective is omitted, an outcome evaluation may fail to detect a significant contribution of the program.

Depending on the program, the intermediate objectives may be to produce changes in one or more of the following areas:

Attitudes	Intended future use
Personal adjustment	Interactions with family and peers
Knowledge about drugs/alcohol	School performance
Criminal activity	Social-recreational activities.

This list is not exhaustive, and some managers may immediately identify other areas where their program seeks change. The program manager needs to make sure that all relevant objectives are identified before an outcome evaluation is actually conducted.

To summarize:

outcome evaluation tells you whether
what's going on
changes the participants.

Impact evaluation.—Information produced at this level of evaluation is broader in scope than process or outcome information. There are, however, parallels between outcome and impact evaluation. An outcome evaluation measures changes in program participants, whereas an impact evaluation measures changes in the entire population for whom generalized effects are expected. The identification and estimation of impact are particularly important in evaluating prevention activities. For example, the results of an impact evaluation can be used in decisions about program expansion. The results of an impact study on an entire high school population where only some students participated in a prevention program could aid in expanding the program to reach even more students, perhaps in other schools.

Generalized effects of a program occur throughout the community—however defined—and across prevention programs within a community. Thus these effects are often measured in aggregate or cumulative form such as incidence/prevalence levels, rates of drug or alcohol arrests, and hospitalizations. A decrease in substance abuse in the community may have many other results. For instance, an improved school environment and lower maintenance costs may result from reduced substance abuse. Of course, one task of the impact evaluation is to determine how much of the overall improvement is attributable to the prevention activities operating within the community.

Before program impact can actually be assessed, some important barriers that limit the extension of program outcome must be carefully considered. For example, if a program is aimed at a very limited subgroup (by age, race, ethnicity, geography, etc.) of a high risk population, then the magnitude of any measured impact on the entire population might be quite small. Other factors to be considered for an impact evaluation include a definition of community related to a program's size and impact, intended and unintended effects, and delay and durability of effect.

Definition of community.—The probability of a prevention program reaching members of a target group is obviously related to the size of both the program and the group. The definition of community should relate to the scope and objectives of a program and be limited to an area in which detectable impacts may result. Take the case of a program limited to one class within one school. The impact of the program will probably be limited to families of the students involved, some of their peers, and perhaps their neighbors. The definition of community should be so limited. Compare that to the case of a television show where the potential impact, and thus the community, are limited only by the scope of the broadcast (local, regional, or national broadcast).

Intended and unintended effects.—By definition, intended effects of a program are always positive. They are, after all, based on program objectives. Unintended effects may be either positive or negative. For instance, a program aimed at decreasing one type of substance abuse—alcohol—may increase a different type—cigarette smoking. Though these effects are not expected, knowledge of them may help in modifying the program—for example, adding a lung cancer film to the film on alcohol related brain damage!

Delay and durability of effect.—If an impact evaluation is implemented too soon after a program is initiated, no impact may be found. Obviously there may be a delay before any generalized effects are measurable. To assess the durability of the impact of a program, timing is again important. If possible, a followup study would indicate the length of time that the overall impact of a program can be sustained.

The issues of intended and unintended effects, as well as of delay and durability of effect are as important for outcome evaluation as they are for impact evaluation. Or, evaluations must consider

what happened, expected or not,
how long it took to happen,
how long it did (or would) last.

All of these factors need to be taken into account by the program manager and evaluator. Developing a rational plan at the impact level may be more involved and more costly, but the knowledge gained can be significant.

To summarize:

impact evaluation shows whether
what's going on changes the larger community.

Finally, looking at evaluation as a whole:

each evaluation level can lead you
through feedback loops
to program improvement, or
to put it graphically,

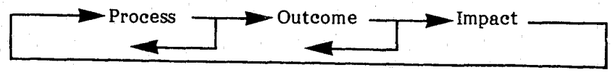


Figure 2-1 illustrates a list developed by NIDA of major indicators and approaches for the three levels of evaluation. Note that process and outcome evaluation focus on effects within the program, whereas impact evaluation focuses on effects at the community level. Relevant to this model, various methodologies are discussed in chapter 4 of this volume and in the Handbook.

Information Type

A second parameter of evaluation is the type of information that can be generated. Three types can be identified: descriptive, comparative, and explanatory. Descriptive information is the easiest and least expensive to obtain. As the name implies, this type of information describes the program, the clients, the staff, the environment, and so forth. Much of the process-level information obtained in describing a program is necessarily descriptive. Hence, it is important that the program records from which the information is drawn are adequate. A straightforward management information system for recording descriptive information can be started early in a program's development or can be the first step in an evaluation process.

Comparative information involves variables thought to significantly affect program functioning, but does not assign causality. For example, staff attitudes concerning prevention can be compared to the program participants' attitudes toward prevention. Both sets of attitudes may affect program functioning, but determining which set caused the other is the old chicken and egg problem—which did come first? The cost of comparative information will be higher than that of descriptive information in terms of time, effort, money, and design, but more complex issues can be examined.

Explanatory information is used to try to answer even more complex questions such as, why does the program work? If two groups of 12th grade students show different levels of substance abuse, can the difference be attributed to the prevention activities of one group? More importantly, what program components are responsible for the effects? Obviously, gathering and analyzing this type of information requires even more sophistication in terms of design and theory testing, as well as more financial and other resources. But if the purpose and goals of the evaluation require it, the effort expended is worthwhile.

In general, the type of information sought is a function of data availability (what data are already gathered and what can be obtained), evaluation design (within the constraints of availability, what does the manager want to know) and analytic technique (in what form does the evaluator want the data). A fuller explication of the process of choosing information type(s) is found in chapter 4.

Figure 2-1. Drug abuse prevention evaluative research model (Bukoski 1979)

LEVEL OF EVALUATION	PROCESS →	OUTCOME →	IMPACT
Focus of evaluation	Prevention program effects		Aggregate or cumulative effects at the community level
Potential indicators of effectiveness	Description of target audience/recipients of service Prevention services delivered Staff activities planned/ performed Financing resources utilized	Changes in drug-related: Perceptions Attitudes Knowledge Actions: Drug use Truancy School achievement Involvement in community activities	Changes in: Prevalence and incidence of drug use Drug-related mortality/ morbidity Institutional policy/ programs Youth/parent involvement in community Accident rates
Potential prevention evaluative approaches	Examples: The Cooper Model for Process Evaluation NIDA-CONSAD Model NIDA-Cost Accountability Model Quality assurance assessment	Examples: Experimental paradigms Quasi-experimental designs Ipsative designs e.g., Goal Attainment Scaling	Examples: Epidemiologic studies Incidence and prevalence studies Drug-related school surveys Cost-benefit analysis

Target Area

A third facet of the evaluation process is the target or focus of the program and hence the focus of the evaluation. For example, are changes in individuals over time being sought? Are community (however defined) or societal changes in attitudes/behaviors of interest? Depending on where the center of interest lies, different questions can be asked of different people. The evaluative focus is usually one of the following targets—individual, program, service system (comprising several programs), or societal. The choice will depend on the needs and resources of the decisionmakers involved in the evaluation process. For example, a school board in an urban area may want to evaluate various prevention projects throughout the school district as a whole, or one principal may want to find out if a specific group activity is succeeding in its prevention activities. These two situations will result in different types of evaluation activity, with more emphasis placed on community-wide impact evaluation in the first case than in the second. However, an evaluation focused on one target area can still have an effect on others. For instance, an evaluation concerning a group of students in one prevention project could contribute to a better understanding of the overall service system of which that program is a part.

The three parameters—level, information type, and target area—and their relationships are graphically displayed in figure 2-2.

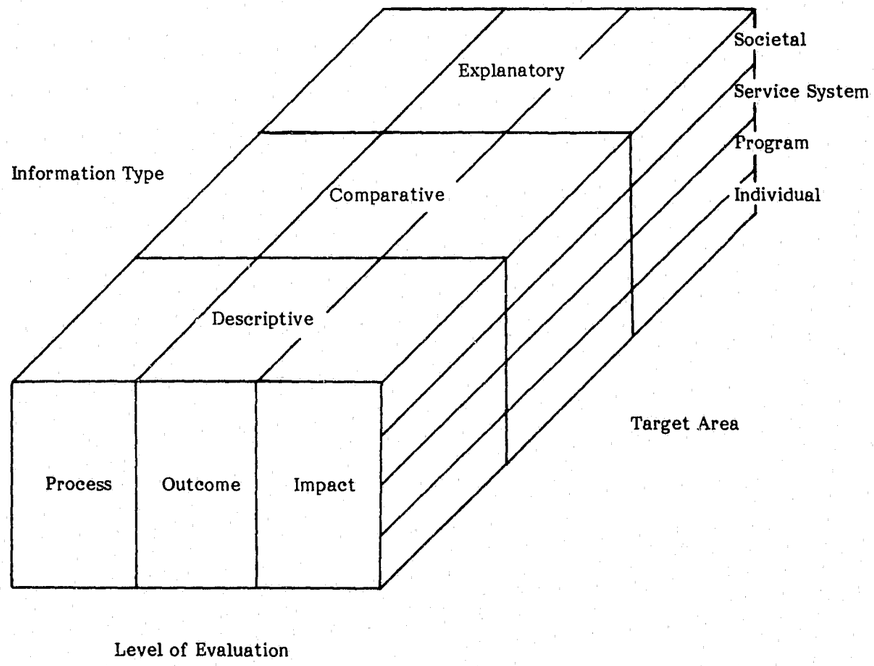


Figure 2-2. Evaluation Considerations (French and Kaufman 1981)

DEVELOPMENT OF AN EVALUABLE PROGRAM

Every program is evaluable—some information is always available to indicate what's going on. A major objective of program evaluation is to use this information base for decisionmaking. Continual program improvement is contingent on feedback to the manager and other staff regarding program development and implementation. With this in mind, evaluation should become an integral part of ongoing program development, supplying appropriate feedback to decisionmakers. Certain issues of program relevance, program quality, etc., can be examined at different phases of program development and operation. The information obtained can provide a foundation from which criteria for further development and management decisions can be established.

The greatest power of evaluation will be realized if evaluation has a role from the first stages of program development. For example, a process evaluation documenting the earliest phases of program development can provide information that would otherwise be unavailable. However, regardless of when the evaluation takes place, feedback can enhance the chances of further growth and improved program effects.

Five major phases of program development were delineated in the Handbook for Prevention Evaluation. The same distinctions are presented here, emphasizing the information needs of the manager and questions appropriate for each phase. The phases are:

- o needs assessment
- o policy development
- o program design
- o program initiation
- o program operation.

The discussion below looks at the first three stages as planning phases and the last two as implementation phases.

Planning Phases

Needs assessment.—The initial phase of program development is establishing whether and to what extent a certain problem exists within a given subgroup in the community. For example, is there a growing substance abuse problem among a high school's student body? Once this information is obtained, a specific cause of the problem is postulated leading to the definition of a need for a specific process to overcome the problem. For example, if the problem is caused by a lack of organized activities involving high school students, then an alternatives program for high school youth would be proposed as a means of ameliorating the situation. If the problem is inaccurately measured, or the causal assumption is wrong, then the program may eventually be found ineffective. The manager needs to have accurate information to confirm that the program is based upon the correct assumptions concerning the problem while the prevention program is still in the planning stage rather than when the program is in full operation.

The **ideal**—problem assessment leads to the definition of need.
The frequent **reality**—the problem assessment is used to justify what somebody already believes.

Policy development.—During the second phase, the goals and specific objectives of the program are defined, based on the theory postulated in the previous phase. Many different factors, not all of which are internal to the program, need to be taken into account at this point. Financial resources, values, attitudes, and concerns of various individuals (policymakers at the levels of program, local government, State and Federal government, program staff, and potential program participants) need to be identified and their impact on program policy assessed. Depending on the specific problem, goals and objectives may have to be limited in a realistic sense to fit the sociopolitical environment. Given the context of these variables, the manager will want an accurate translation of the theory into policy. A clear understanding of the factors involved—whether they would support or impede the program's development—is needed to ensure a rational policy development.

The **ideal**—goals and objectives flow from previously formulated theory.
The frequent **reality**—programs can operate for years without formulating anything but the most obvious goals.

Program design.—The final planning stage transforms the program policy into significant program characteristics. Specific program components and activities must be developed in relation to overall policy. This is the operationalization of the policy, where the program decisionmaker needs to know what has been done previously to meet similar objectives. How can the same thing be accomplished now, given existing resources, program capacity, staff size, facility limitations, staff background and qualifications, and community characteristics? All of these factors need to be taken into account in order to produce a fully detailed program design.

The **ideal**—program components and activities are rationally justified by goals and objectives.
The frequent **reality**—trial and error.

The introduction of an evaluation at any of these planning stages can increase the amount and quality of feedback. To bring the reality closer to the ideal, the evaluation should do more than just assess the attainment of specific objectives. If stated objectives are not reached, information concerning stages of development before program operation becomes critical. At earlier stages an evaluator can ask questions that would also be of interest to the program manager. For instance, at the needs assessment stage, the assessment of the problem can be examined. If the objectives of the program are met, but the problem does not really exist, should the program be labeled a success? Or maybe the assumptions regarding the cause of the problem or the definition of the need are erroneous. In that case, the objectives may not be met in even a smoothly operating program because the policy developed and implemented may have no bearing on the problem.

The foundations of process-level information are found in all three of these planning phases. Evaluation at this time can provide information on the flow from

problem ⇒ need ⇒ theory ⇒ policy ⇒ goals ⇒ objectives ⇒ design

Information needed for process evaluation may be available later while the program is in operation, but it would probably be of more immediate help to the manager if available during these planning stages. Information would also tend to be available more efficiently with less cost in terms of time, effort, and money before program implementation.

Implementation Phases

Program initiation.—At this stage, the program is established and implemented; translation of theory into action takes place. The manager can now see if the implementation matches the program design. That is, information on participants, resources, and constraints can be compared with those in the program design. This stage can also be viewed as a debugging phase where problems in implementation are corrected and the program is set up for smooth operations. Is the program operating as designed? Are staff assignments recognized, accepted, and carried out? Are the participants receiving the types of services planned?

The **ideal**—bugs are recognized and corrected.
The frequent **reality**—the bugs survive.

Program operations.—Once the program is fully operational, it does not simply run by itself. Good management and direction are needed to keep the program functioning and improving. In addition, a program does not operate in a vacuum. Continual upgrading and development of the program must include mechanisms for adapting to changing needs and problems in the client population and community. Some changes may be the result of the prevention program, as measured by outcome and impact evaluation. Others may be due to some external forces, such as local, State, or Federal political decisions, changing levels of community involvement, or changing supports and constraints of funding sources.

The **ideal**—operating programs continually increase their ability to meet objectives.
The frequent **reality**—maintenance of the status quo or irrational change.

None of these phases necessarily represent discreet, mutually exclusive periods of time. Program development is a dynamic process, with constant feedback and improvement. Different aspects of a program can be in different stages of development at the same time. As needs of the community change, so too must the program evolve. Evaluation is one tool that can be used to aid in that development. The model presented in this chapter is one method of ensuring a rational approach to both the evaluation and development of the program.

CHAPTER 3: PROGRAM ISSUES IN PREVENTION EVALUATION

(What Managers Need to Know or Remind Themselves About)

The scene is the director's office of a local prevention program. Scattered across the desk are all the signs of a late-evening vigil, including several books opened to dog-eared pages. The titles tell most of the story. An evaluation is being considered, and dusty volumes of college textbooks on statistics and research methods are being frantically reviewed for long-forgotten definitions: chi squares, t-tests, and Type II errors. The director appears to be wondering what possible direction she can give to the evaluation when she doesn't even remember what a quasi-experimental design looks like.

The director's predicament is not uncommon. Most conscientious program decisionmakers are aware that they have a role to play in the evaluation process. Some have watched evaluation studies take place within their own programs or have begun to explore the literature on prevention evaluation. Unfortunately, too little has been written on the specific role of the program manager.

Some program professionals, as in the example above, try to become conversant enough with research terminology to at least participate in planning at some level. Others, who have little or no background in evaluation research may fail to see the importance of their involvement and turn the entire task over to an evaluation consultant.

Undoubtedly, the manager needs to know enough about evaluation to ask critical questions concerning the methods being used. Other sections of this monograph address concerns about evaluation models and measurement. The focus of this chapter, however, is on program knowledge rather than evaluation knowledge. Amid the work and anxiety of an evaluation project, the program decisionmaker frequently loses sight of the fact that:

The most significant contribution program managers make to development of the evaluation lies in what they know about the program rather than what they know about the evaluation process.

To appreciate the significance of this statement, it is important to understand what makes an evaluation work. Weiss (1972, p. 6) makes an important distinction between research and "evaluation" research by noting that, in the latter case, the questions to be considered are those of the program rather than those of the researcher. Sooner or later the decisionmaker must consider these issues:

- o What do I need to know about the program?
- o What decisions am I prepared to make?
- o How should the evaluation results be presented to help make those decisions?

Many elaborate evaluations have failed to yield valid or useful results because the evaluator made inaccurate assumptions about the program itself or because the users of the evaluation findings had not been clearly identified.

Program information from the perspective of the decisionmaker is crucial to the evaluation process. It represents a view of the program the evaluator does not have and provides a context for evaluation activities. Program considerations affect every aspect of the evaluation process, from the selection of questions to the choice of instruments to the use of results. They influence what kind of evaluator should be consulted and what kind of staff adjustments will be necessary to accommodate the evaluation.

In some ways it might seem presumptuous to devote a chapter to explaining prevention programs to program decisionmakers. After all, aren't most managers already familiar with the resources and services of their organization? Yes and no. They usually have information about the program, but need to understand it from an evaluative point of view.

Decisionmakers usually keep at their fingertips such program facts as the annual budget, a description of services, and an organizational chart. However, at times the manager needs other kinds of information. For example, in long-range planning, questions must be asked about the program's mission, the consumers of its services, and its potential for change. In the same way, certain aspects of the program need to be considered in preparation for an evaluation. However, many program decisionmakers have not been shown these connections. Too often, evaluations are not geared to the needs and circumstances of the program, and the program staff's questions are never incorporated into the design. There is always the risk that the program will serve the evaluation rather than be serviced by it.

Asking important program questions at the beginning of the evaluation process helps to ensure that the results will be genuinely useful. False starts due to misunderstandings or confusion are eliminated, and a true partnership can develop between the evaluator and program personnel.

In this chapter, program issues relating to evaluation will be grouped into four areas and discussed from a manager's perspective:

- o What is the program and what is it meant to do?
- o What are the evaluation questions to be asked by the program?
- o What kind of evaluation will fit a particular program?
- o Will the evaluation be worthwhile for the program?

Reflecting on a prevention program from this perspective is not only helpful for the program decisionmaker, but as Patton (1978) points out, equally valuable for funding sources, line staff, and consumers. Perceptions about program goals and services are not always shared among those involved at different levels. An evaluator may receive very different impressions of the same program when it is described by an administrator, a staff member, or a client. As many program perspectives as possible should be integrated for the evaluation to be successful.

The program manager should be involved throughout the evaluation process. Programing issues concerned with interpretation and utilization of findings are equally as significant as those that take place in early phases of a study. Most importantly,

the decisionmaker's knowledge
of the needs, purposes, and goals of the program
is essential to evaluation.

WHAT IS THE PROGRAM AND WHAT IS IT MEANT TO DO?

This is the simplest of questions, and one for which every program manager has a ready response. All programs have goals and objectives, even if they are implicit and unwritten. Yet, there may not be an identifiable program to evaluate or even agreement about the program's purpose. Evaluators cannot work with this ambiguity. Many note that consultations with prevention programs frequently begin by backing up and reexamining program goals.

Evaluators encounter two common problems with program objectives. The first has to do with the relationship between objectives and the program process and outcomes. A prevention program may have a beautifully written action plan that no longer describes the services currently provided. Perhaps funding was cut. Perhaps there was staff turnover, or a particular project was changed slightly. Maybe the program never did reflect the stated objectives, which might have been written originally to satisfy an external audience. Without objectives that accurately describe the program's current intended outcomes, the evaluation may proceed on a meaningless course.

The second problem is more complex but no less common. Many programs' stated objectives describe only program effort or process. For example, a prevention program directed toward school children might include the following objective—deliver eight teacher-training sessions during the school year. This objective is clear and measurable but describes only the process, not the outcome of that activity. Such

statements of program process are necessary for the evaluator to understand the program's services, but alone do not link the activity to its outcome. The evaluator may be unsure what outcomes to examine. Even worse, the program may include an impressive array of prevention services without any clear indication of the specific results expected. Both outcome and impact evaluation rely heavily on well-defined statements of what condition(s) should exist as a result of the program. In addition to a description of program process, objectives setting forth the intended program outcomes are essential.

Leaving aside the evaluator's use of outcome objectives, their importance as a guide for the program decisionmaker is unquestioned. Stated another way, "If you don't know where you're going, you may end up someplace else." A program may show all kinds of results, but it is difficult to judge success or failure without some objectives against which to measure those results.

Goal setting is the first major task in preparing for an evaluation, and one of the manager's responsibilities. Do you have clear and concise goals and objectives relating to program effort as well as to outcome? Can your services be clearly identified and defined? Is there agreement about the program's intended results? Do you have a clear sense of what represents success or failure? How much change is satisfactory?

Programs with articulated, measurable outcome objectives make both daily management and evaluation design much easier. Valuable time and resources that would otherwise be spent on goal setting and program planning can instead be used to discuss specific evaluation methods.

Other aspects of the program may also help to identify its structure and purpose to both the manager and the evaluator. In the prevention field, for example, programs can be categorized in a number of general ways that help to describe their goals as well as their strategies of service delivery. Although these program dimensions may not be specifically written down, they are no less important to decisionmakers in describing the program.

Prevention/Health Promotion

Prevention programs employ not only widely different strategies, but try to effect different goals. The most notable distinction, perhaps, is between programs intended specifically to prevent alcohol and drug problems and those with more general goals, such as health courses with substance abuse modules. Within an evaluation, recognizing these distinctions is important; they help evaluators appreciate the kind of program results acceptable or of importance to decisionmakers.

Indirect Service/Direct Service

Many prevention programs deal with intermediary groups to promote change in a target group. In such cases, program goals may be stated in terms of the eventual change desired in the target group. For example, a school-based program may have as its goal the development of social competencies among elementary students. However, the program activities may be directed toward the training of teachers and school administrators. In this case (as in similar activities like information distribution, training, and consultation), the program manager must distinguish ultimate consumers from those directly affected by program activity.

Etiology of Abuse/Model of Prevention

Programs differ in their perspective on the causes and prevention of alcohol and drug abuse. Some base their services on models of individual attitude and behavior change. Others approach the problem from a perspective of social standards or cultural norms. Implicit in every prevention program is a set of beliefs about what causes people to develop problems and what preventive strategies are likely to be effective. Identifying these beliefs is extremely important in defining the kinds of results sought. For example, one community adopted a prevention program designed to change norms regarding public intoxication. Although the community organizers used familiar strategies of awareness and community education, evaluators would have missed some of the program's substance had they looked only for measures of individual change. A clearly articulated program philosophy is essential in creating an evaluation design, deciding what to measure, and choosing measurement tools.

The program purpose, written or unwritten, is the cornerstone on which all other evaluation questions rest. The evaluator's role is to determine actual effects of program services. However, the role of the program decisionmakers begins with a clear statement of what they intend to accomplish.

WHAT ARE THE EVALUATION QUESTIONS TO BE ASKED BY THE PROGRAM?

Once the program decisionmaker has defined goals and objectives of the program, it is time to ask similar questions about the evaluation. Evaluations must also have goals and objectives. Evaluators and program administrators alike are often dismayed by how few evaluation studies yield results useful for program direction. To be sure, part of the problem lies in conditions outside the evaluator's control. Nonetheless, more often than not the program decisionmaker finds that the study has failed to address essential questions.

Except for the fundamental questions regarding the program's intended outcomes, no aspect of the evaluation is more important than developing the questions that need to be answered. As with program goals and objectives, evaluation questions should be stated as specifically as possible. For example:

- By the end of the project year, can an increase be shown in the number of schools using the entire curriculum developed by the program?
- Can a decrease in the number of arrests for driving while intoxicated be shown in Baker County over the first 6 months of the project?
- Can test scores of program participants show an increase in knowledge regarding the risks of drug use during pregnancy?

Obviously, the type of change the evaluation questions examine depends on program outcome objectives set forth by the organization. These first two phases of preparing for the evaluation are interdependent.

Because funding sources and program managers sometimes want different things from an evaluation, the manager may want to set some priorities. Certain questions may be more important to the organization than others or may be more answerable given the time and resources of the study. For example, a program director may be interested in comparing two different prevention strategies. However, this kind of comparative study may be less pressing for the organization than having other information available to the county for the next funding cycle.

As in the goal-setting process, a number of considerations are helpful in developing evaluation questions. The manager must ask why and for whom the evaluation is needed. Program evaluations are conducted for many different reasons and audiences, for example:

- To provide **feedback** for internal management to guide development of the organization.
- To assure **accountability** to some external source. With decreasing availability of financial resources, programs are called upon to use evaluation results to justify new or continued funding. In some cases, the manager may know exactly what criteria the funding source will use to judge a program. At other times, though, the program is forced to make assumptions about what kind of evaluation results will be convincing to authorities.
- To **market** new and innovative program methods. Other services provided by an organization may be well accepted in the community, and a manager may want to use the evaluation to add credibility to more recently developed services. In particular, evaluation findings may be used to support decisions about replicating pilot programs.
- To meet **requirements** of a grant or contract. The manager should, of course, look beyond the program's mandate for evaluation to consider ways in which the research findings can be useful for both the program and the mandating agency.
- To satisfy the **curiosity** of someone in the organization (particularly in programs where innovative strategies are being used). Although such questions may have little relationship to the stated program objectives, some of the most dramatic program effects are discovered through the personal conviction and questioning approach of someone deeply involved in the delivery of services.
- To respond to the **needs of users.** Evaluation is best formulated with participation by users regarding the questions to be asked and the way findings will be used. Don't forget anybody: legislators, school board members or county commissioners, funding source representatives, boards of directors, program administrators, line staff, and consumers. The concerns and viewpoints of as many user groups as possible should be incorporated into the evaluation questions.

Program decisionmakers must also recognize when an evaluation might be inappropriate. For example:

- o When you don't know what the program is—there may be no agreement on program goals and objectives, or they may not yet be sufficiently defined.
- o When you don't have the resources to answer the questions you need to answer.
- o When the answers won't make any difference—when the potential users of the evaluation results are unable or unwilling to take action based on those results.

These factors should be considered seriously by the manager before undertaking an evaluation. Although pressure is increasing for prevention programs to become involved in evaluation efforts, decisionmakers should recognize when evaluation is incapable of yielding useful results.

WHAT KIND OF EVALUATION WILL FIT THE PROGRAM?

Even program decisionmakers who appreciate their role in developing program objectives and research questions may believe their involvement ends when evaluation methods are discussed. Managers with little or no background in research techniques may be inclined to withdraw and simply wait until results become available. In fact, the evaluation design and the selection of appropriate instruments should begin with yet another set of programmatic questions best answered by the manager. In too many cases decisions regarding evaluation design and methods are left entirely to the evaluator. This can lead to problems, including the possibility that the resulting data cannot be used. Selecting appropriate evaluation methods begins at the program level with the question

How can the information be collected and presented in a way that will be convincing and useful?

Program managers can ensure the usefulness of evaluation findings by playing an active role in determining methods. Evaluators are human too. They represent a number of disciplines giving them a variety of perspectives and experiences. The manager should choose an appropriate evaluator to help answer the program's questions. The major consideration is the consultant's willingness to work in partnership with the program. However, other factors influence an evaluator's ability to respond to program needs.

- o An evaluation may address issues ranging from changes in individuals to effects on entire communities. Inevitably, evaluators have varying levels of experience with different areas of social research. One consultant may be excellent for measuring change in individual student attitudes but have little background in evaluating a community organization project. The skills necessary to measure individual change or social change are not mutually exclusive, but the manager should look for an evaluator experienced with the kinds of questions being studied.

- o The evaluator must be sensitive to the program's cultural and ethnic factors. Ethnographic studies, for example, demand that the evaluator become intimately familiar with the cultural community being studied. Even with more traditional techniques, the importance of cultural sensitivity on the part of the evaluator cannot be overemphasized. In multicultural or ethnic communities, it cannot be assumed that standardized instruments will yield valid results. Not only do issues such as language and methods of data collection come into play, but also the community's norms for such things as drug use, social interaction, and healthy lifestyles.

- o Evaluation methods can generally be divided into two types, qualitative and quantitative. Traditionally, only quantitative methods were acceptable in sound evaluation practice. More recently, a number of noted evaluators—Campbell (1975) and Cronbach et al. (1980), for example—have moved away from insisting on quantitative methods, and opened up the possibility of qualitative approaches. These include participant observation, program journals, and unstructured interviews. Depending on the prevention program, quantitative or qualitative—or both—methods may be called for. Evaluators, however, may be more comfortable or skilled in one area, and the manager must strive to match the evaluator's style with the needs of the program.

These approaches are not mutually exclusive. Many evaluations combine qualitative and quantitative methods and attempt to measure change at both individual and group levels. Based on training and experience, evaluators may approach the project with a set of biases. Perhaps they have a favorite instrument used successfully with other programs, or a conviction about good evaluations that does not allow

for a broad range of techniques. In any case, evaluators influence the design, and it is critical that they be able to respond sensitively to the kinds of evaluation questions being addressed.

Other program issues determine the kind of evaluation to be conducted, including the needs and capabilities of the organization.

- o Money! Good evaluation need not be expensive, but certain direct cost decisions must be considered. Will clients be paid for their participation in the evaluation? Will other professionals need to be hired?
- o Program recordkeeping. Does the quality of existing records meet the information needs of the planned evaluation?
- o Data analysis resources. Do resources, including computer access, exist at the level necessary to analyze the data collected?
- o Time constraints. Will the results be available when they are needed?
- o Program staff availability and expertise. How much are program staff expected to contribute to each phase of the evaluation? Will they be able and willing?
- o Money!

The kind of evaluation that fits any single prevention program depends, in part, on all these variables: finding an evaluator with appropriate experience, matching an evaluator to the cultural dimensions of the program, deciding on the appropriateness of qualitative and quantitative measures, and looking carefully at the resources of the organization. There are also other factors outside the organization's influence, such as the mandates of funding sources. In each case, the program decisionmaker must play an integral role in designing the evaluation. The study itself involves far more than simply choosing instruments and interpreting printouts. It is a process of deciding how to ask appropriate questions and how to represent the findings in a useful and convincing way.

WILL THE EVALUATION BE WORTHWHILE FOR THE PROGRAM?

In even the best-planned evaluations, where program objectives have been articulated, questions clearly stated, and a study design developed, there is usually some sense of hesitation on the part of the program decisionmaker. Will the evaluation process end up costing the program more than it offers? For whatever reasons the evaluation is conducted, will the findings warrant the amount of time and attention it involves?

These are important questions for the manager to consider. In every case, the process can be better managed if some of the potential costs and benefits of evaluation are first analyzed.

An evaluation project can cause disruption within an organization in countless ways. Evaluation studies often bring with them additional forms to fill out, new assignments for staff, demands for clerical assistance, and increased attention to program details. An evaluation process frequently means that new and unfamiliar faces will be injected into the program's daily operation. Staff may feel the pressure of having their professional activities scrutinized, and awareness of outside accountability usually creates some degree of anxiety.

Left unattended, these dynamics can result in serious resistance to the evaluation process. All other preparatory steps are useless if the staff does not maintain program conditions necessary to complete the study. It is essential, therefore, that the manager seriously examine all possible ways in which the evaluation might negatively affect day-to-day operation of the program.

To the extent possible, persons affiliated with the program should be drawn into the evaluation planning from its inception. Evaluators should become familiar to staff, and the reasons for each component of an evaluation design should be thoroughly explained at each stage of the process.

In some cases, disruption cannot be avoided. The program might need to be modified to accommodate an evaluation design. For example, if the evaluation requires data on a program's parent-education component, more emphasis may need to be placed in this area for a period of time to develop a large enough sample for study.

Other clashes may occur between the program philosophy and aspects of the evaluation design. (Many of these issues can be avoided through the kind of design planning discussed earlier.) For example, some program professionals believe it is unethical to randomly serve some clients but not others, a feature of

some evaluation designs. Issues may also emerge regarding the use of confidential information or the presence of an evaluator as an observer in group activities.

These kinds of situations cannot help but be disruptive to a program. However, to the extent that such changes are well planned and thoroughly explained they need not have negative effects. Potential disruption to both the program and the evaluation process can be minimized if the manager anticipates and plans for such possibilities.

Other aspects of the evaluation process may have unpleasant repercussions if they have not been considered. For example, sometimes evaluation is initiated without planning for possible negative results. Particularly where an evaluation may be used to justify the program's funding or continued existence, the manager must carefully consider the potential effect of less-than-positive findings. In the same way that staff resistance or other internal effects of an evaluation process must be examined, the manager must also look at the ability of the program to accommodate indicated or recommended changes. The evaluation process can be particularly costly to a program that is prepared to receive only enthusiastic validation. Even negative evaluation findings can be used constructively if the program is resilient enough to accept criticism and consider change.

Program disruptions caused by evaluation can be offset by potential benefits. In addition to providing external accountability and support for prevention programs, evaluation can influence internal decisionmaking and provide continuous feedback to staff, helping to modify or improve program practices. For consumers of prevention services, who either participate in the program or are concerned about its effectiveness, evaluation assures some measure of quality control. Finally, whether the results are anticipated or the findings are of any significance, evaluation can prevent what Weiss (1972, pp. 110-128) refers to as "barnacle-encrusted" programs. In other words, just by incorporating the process and rigors of evaluation, prevention managers and staff can infuse their program with creativity and continue to grow and change in ways that improve their services to people.

Successful evaluations are a marriage of program knowledge, good management, and research skills. For the manager, the importance of moving through the planning stages described here cannot be overstated, each stage building on the other. Without a clear sense of what the program intends to accomplish, it is impossible to ask meaningful evaluation questions. Without specific questions, appropriate methods cannot be chosen to conduct the study. Without measures that are sensitive to the needs of the program, the evaluation threatens to harm more than help. Without adequate resources to analyze and interpret data, the best measures may come to naught. Without clear and relevant presentation of findings to evaluation users, the whole effort may be fruitless.

These are program issues. The success of any evaluation is intricately tied to the manager's active participation in reflection and planning. This chapter began with a director wondering what possible direction she could give to an evaluation when she couldn't even remember what a quasi-experimental design looked like. The answer: a considerable amount. Old college textbooks on statistics and research methods are useful, but the manager's primary contribution to the evaluation process is understanding the program and what it needs to know.

CHAPTER 4: EVALUATION ISSUES IN PREVENTION PROGRAMS

(The Heavy Stuff—What Else?)

This chapter looks at evaluation from the evaluator's perspective and is designed to provide program decisionmakers with technical information so that they can:

appreciate the difference between good and bad evaluation design,
better understand what evaluators do,
become more active participants in the evaluation process, and
become wiser consumers of evaluation.

Technical aspects of evaluation are presented throughout the chapter. Evaluation terminology is emphasized so program decisionmakers can better understand and communicate with evaluators.

Evaluators, in designing an evaluation of a program's effectiveness, have an overriding responsibility to set up the evaluation so the question of whether the program produces desired effects can be answered as accurately as possible. Accurate answers demand attention to many issues in evaluation design. Managers must understand these issues for two reasons. First, in using evaluation results to make decisions, program managers need to be wise consumers, able to judge the quality of evaluations, rather than forced to take results at face value with no understanding of how they were generated. Second, managers in the process of having evaluations designed for their programs will be better able to understand the evaluator's activities. Evaluators often do, or ask program staff to do, certain tasks that may seem a waste of time at best, or costly and disruptive of program functioning at worst. Well-informed program managers who understand what is at stake with various aspects of the evaluation can contribute to the quality of their program evaluation. A director may well ask, "What will it take to convince others that my program is valuable?" An adequately designed evaluation that documents the nature of the program and then shows its effectiveness is at the root of answers to that question.

ISSUES IN EVALUATION DESIGN

An easy way to consider design issues is to scrutinize an evaluation. First, we'll describe an evaluation design. Then we'll backtrack and examine it to show how, through faulty design, an evaluation can lead to incorrect conclusions about the program. We will then consider issues of theoretical and technical importance in the evaluation process.

Example Program and Evaluation

The following hypothetical example was created to illustrate poor evaluation and issues of evaluation design.

Prevention program.—The program was intended to improve self-concept among junior high school adolescents in seventh and eighth grades. The program's theory was that improved self-concept would cause a decreased desire to use drugs as an escape from the difficulties of adolescence, as well as an increased resistance to peer pressure to experiment with drugs. The program was designed specifically for children

who were struggling with their adolescent development reflected by academic difficulties and problems in relating to family and peers. The program consisted of groups of students meeting with program staff once a week after school over the course of a semester.

Program staff.—Two staff members each led one student group. The first was the school guidance counselor, who had substantial previous experience working with troubled adolescents and wanted to show the school system the worth of such programs. The second was a foreign language teacher who was thinking of going back to school and changing careers in the direction of working with adolescents in a counseling setting. She wanted to try leading a student group to see if she would enjoy intensive contact with adolescents. The guidance counselor had been trained in the self-concept curriculum at a special workshop and had run the program once at a local community center. She introduced the program to the school and trained the foreign language teacher just before the semester began.

Participants.—Program participation was voluntary. The program was advertised in the school through a poster campaign. Each group leader also solicited students to insure adequate participation. Finally, all teachers in the school were asked to encourage their homeroom students to participate, especially those who seemed to have problems.

Evaluation.—The guidance counselor wanted data showing that the program was effective in improving self-concept. She consulted a school psychologist who suggested that she use a self-concept scale that he was developing and had already tested on some high school freshman and sophomores. Because he was interested in data from junior high students, he agreed to analyze the data in exchange for having the use of the results for further development of the test. He suggested that the guidance counselor administer the test at the beginning of the semester as a pretest and at the end of the semester as a posttest. Since the program was ultimately supposed to prevent or delay the use of drugs, the school psychologist also recommended, and the guidance counselor adopted, a well-known scale of self-reported drug use.

At the beginning of the semester, the 2 groups contained 35 participants, 16 with the guidance counselor and 19 with the language teacher. Participation waned so that by the end of the semester only 18 participants remained, 13 with the guidance counselor and 5 with the language teacher.

Because the guidance counselor was concerned about data confidentiality, she instructed the students not to put any identifying information on their pretests or posttests. The only information she kept was which were pretests and which posttests.

The school psychologist also strongly recommended gathering self-concept and drug use information on students not participating in the program, taking these measurements at the same time as the pretests and posttests. The language teacher asked nonparticipating students in her classes to voluntarily take the test at the beginning and again at the end of the semester. She got the highest response rate from her advanced language class and ended up with 20 pretests and 17 posttests. She also did not require identifying information on the tests, but merely kept pretests and posttests separate.

The school psychologist analyzed the data using the t-test to assess whether average self-concept score was higher on the posttest than on the pretest. He applied the t-test separately to group participants and nonparticipants and found no statistically detectable self-concept change in either case. A similar analysis showed no change in the average scores on the drug use test. The guidance counselor, clearly disappointed in the results, concluded that her program had no beneficial effect on participants.

Developing the evaluation.—As stated above, the guidance counselor wanted to show the school system the worth of drug prevention programs by collecting data to confirm that this program was effective in improving self-concept. The school psychologist pointed out that it would also be helpful to obtain a measure of change in self-reported drug use.

It must be recognized that the results of evaluations are used as a form of argumentation, as a means of persuasion. Unfortunately, not only did the counselor and the psychologist neglect to consider whether they were asking the right questions, they also failed to identify the prime users of the evaluation findings and the ways the data could be used to explain the program's effects. Beyond these problems, the study did not adequately assess the theoretical bases of the program.

What is to be evaluated is at once a political and a theoretical question. Often programs mounted as drug prevention programs are not directed to drug use itself but rather to improving life skills, with the expectation that a number of self-destructive and antisocial behaviors will be affected. Thus a self-concept program, offered as a drug prevention program, might also be implemented by the juvenile justice system. The underlying assumption would be that similar connections exist between poor self-concept and criminal

behaviors such as vandalism or delinquency. In both instances self-concept would certainly be measured. But emphasis on the thorough measurement of drug use rather than criminal behavior would, in part, be occasioned by the agency funding the program, the concerns of the audience for whom the evaluation is intended, and the theory on which the program is based. Each of these factors needs to be considered carefully to sharpen the focus of the evaluation.

Suppose the guidance counselor and her friend had sought a meeting with the school principal before conducting the evaluation. They might have found that the principal:

- o didn't believe in the worth of self-reports of such behaviors as drug use,
- o didn't believe that improving self-concept had anything to do with reducing drug use, or
- o didn't have the final authority to decide whether the program should continue.

Such a meeting could have raised many issues that might have been resolved to increase the impact of the evaluation. First, the value of self-reports is a measurement issue. The worth of a measure, that is, its reliability and validity, is an empirical question—one that can be answered by collecting and analyzing data or by reference to past research.

Second, the link between changed self-concept and reduced drug use is an issue of the validation of a theory, which also can be empirically tested. The questions to be asked in an evaluation are derived from the goals of the program and the theory behind them. Third, the question of who has the power to use the information leads back to the motives for the evaluation.

Three questions which can lead a manager to a usable evaluation are worth repeating:

what do you want to know?
why do you want to know it?
how will you use the information you get?

The first, the one most often asked, depends on the goals of the program. The second depends on the goals of the individuals who seek the evaluation. And the third depends on the quality of the information as perceived by those who will use it in some decisionmaking process.

Obviously, these three questions overlap. The motives for the evaluation will dictate in major part what research questions will be asked and how the answers will be used. Suchman (1967, p.143) named several ways in which evaluations can be abused. Some of these are:

Eyewash—evaluating only those program aspects which are expected to look good.

Whitewash—covering up program failure by deliberately choosing nonobjective, or biased information, such as testimonials.

Submarine—seeking information on program weaknesses in order to destroy rather than improve the program.

Posture—seeking an evaluation only as a gesture to display scientific objectivity.

Postponement—using evaluation as an excuse to delay decisionmaking.

Such abuses are sometimes based on the desire to support unfounded beliefs about the program or on the desire to acquire or maintain power or status. These motives are not reserved for the conscious abuse of evaluation research. To some extent, they motivate all evaluation. Directors without faith in their programs are rare. The school guidance counselor wanted to show others that the program worked, and her belief in the theory was the cornerstone of her motives both for starting the program and evaluating it.

At a different conceptual level, evaluations can be motivated by the desire to improve a developing program or by the desire to demonstrate that a fully developed program is effective. Of course, nothing prevents the evaluation from serving both purposes. In our example, the guidance counselor was apparently satisfied that the program was operating according to plan. For instance, she gave no indication that she was interested in improving the program by identifying group leader characteristics that might guide the selection or training of future leaders. Relevant to the second motive, program evaluation can be motivated by a variety of reasons—to meet funding requirements, to enhance acceptance of the program, to test its theory, to support expansion, or simply to satisfy a natural curiosity.

Clearly, there is a relationship between the program's stage of development and the purposes best served by an evaluation. Even replications of well-established programs are appropriate for evaluation, if only to increase effectiveness relative to cost or to monitor activities to ensure that they accurately reflect the intended program model. In such cases the program administrator is typically the decisionmaker who will use evaluation findings. Evaluations of more mature programs are more likely to be used by several decisionmakers. In either case, there is a need to understand the motives of all key actors and information users to develop a pertinent evaluation design.

Clarifying program goals.—The theory underlying the program also determines what should be measured. A program begins with a set of goals. These goals get translated into program activities which, it is assumed, will affect the behaviors encompassed by the goals. Until the goals of a program have been clearly defined, and the link—from goals to activities to outcomes—has been made, we have no guidelines for what to measure. In our sample evaluation, the guidance counselor gave insufficient consideration to the potential effects of the program. Changing self-concept is an intermediate outcome, not an end in itself. The goal of the program apparently was, by improving self-concept, to produce a further behavioral outcome—preventing or decreasing drug use. But improved self-concept might manifest itself in other areas, such as school performance or improved relationships with family and peers. Such potential outcomes have to be specified and incorporated into clear operational goal statements. These statements guide the choice of variables to be measured in the evaluation. Good evaluation is preceded by a careful articulation of the goals of a program. In our sample evaluation no such activities apparently preceded the choice of measures, hence the paucity of dimensions of outcome considered. An evaluator can be very useful to program staff in helping them define and articulate goals and turn these into testable evaluation questions.

The importance of clarifying every step in program development can be illustrated by returning to the theory behind the sample program, which can be stated as a set of three ordered propositions, each building on the previous one:

- There is an association between self-concept and drug abuse. Those who view themselves positively tend to abuse drugs less.
- A change in self-concept will cause a change in drug abuse. As self-concept improves, drug abuse (or its potential) will decrease.
- The program, as designed and implemented, will improve self-concept.

This theory implies as its consequence that participants in the program will have reduced likelihood of drug use. A theory is affirmed by testing its consequences. If the program has no effect on the drug abuse patterns of participants, then at least part of the theory is false. The association between self-concept and drug abuse has been documented in the literature, but the evidence to support the claim that changes in self-concept cause a reduction in drug abuse potential is not clear. The falsity could lie here—in the second proposition above—or it could be found in the design and implementation of the program. Improving self-concept might reduce drug abuse, but the program as implemented might not improve self-concept. In any event, when the implied consequence is false, then at least part of the theory behind it must be false.

However, when drug use is reduced, one cannot logically conclude that the theory is true unless no other possible explanation exists for the change. In an infinite universe this is a practical impossibility. Logically, the truth of any theory cannot be proven; it can only be inferred with degrees of certainty. At some point, however, the weight of the evidence becomes great enough so that it is reasonable to act as if truth has indeed been proven. The majority of people in the world are probably not aware of Newton's Law of Gravity. Fewer are aware that this Law does not explain the phenomenon as well as Einstein's much stronger, more inclusive theory. Even fewer would be willing to test the truth of either theory by jumping out of a tenth-floor window.

The strength of a theory can be increased in two ways. First, if one tests the consequence several times and finds it true each time, the plausibility of the theory is increased. But this requires enough information on program activities to repeat them accurately. The literature in the field of substance abuse is filled with evaluations that describe programs so inadequately that their activities cannot be repeated. Although these evaluations can draw conclusions about program outcomes, they allow no opportunity to repeat the study. It is claimed (Patton 1978) that one team of evaluators paid so little heed to program activities that they actually evaluated a social program that had never been implemented! Luckily for science, the team found the nonexistent program to be ineffective. Outcome studies are incomplete unless they clearly link program activities both to program goals and their underlying rationales.

A second way to increase the plausibility of a theory is to test it against a reasonable, explicitly formulated alternative theory and its implied empirical consequence. The more competing theories discounted, the more plausible the theory being tested.

To give an idea of the complexity of testing a theory, here are some of the competing explanations that might have been considered in developing the sample evaluation design:

- Students might simply outgrow the tendency to abuse drugs.
- The charismatic influence of the group facilitator causes the change.
- Personal attention being paid to students causes the change.
- Students who choose to enter the program bring to it an intent for change that could have occurred without the program.
- The availability of drugs might have been reduced during the time the program was in operation.

To the extent that each proposition in any theory has already been demonstrated, then the focus of concern is changed. If the evidence that changes in self-concept cause changes in drug use is sufficiently strong, then emphasis should be placed on the program's translation of theory to goals, strategies, and specific program activities.

The more competing theories we discount,
the better able we are to claim that our chosen theory is plausible.

The most frequent complaint of evaluators, shortly after initial program contact, is that program objectives are not clear, specific, and measurable and sometimes are not even articulated. Often the goal statements written in funding proposals reflect the politics of obtaining funds more than actual expectations for the program. Program objectives, derived from the goals, are concrete statements of measurable actions or behaviors regarding the intended accomplishments of the program. Such statements are often referred to as **operational statements.** Because of conflicts that sometimes exist between various interest groups, as well as often unconscious resistance to evaluation, the process of both identifying program goals and translating them into operational statements can be difficult and painful.

Scrutinizing the evaluation design.—Looking back at the example's negative results, we must ask whether the program really had no effect or whether the evaluation design might have allowed a real effect to go undetected. The opposite is also true; an evaluation that yields positive results may show effects that do not exist or are attributed to the program when they are really caused by something else.

In the case of the example evaluation, there are substantial reasons to expect negative results, even if the program were effective. These reasons span issues of both process and outcome evaluation. Keep in mind that evaluation is about the identification of differences and their comparison, whether stated or implied. The evaluator's job is to locate the sources of differences, or variation. Any part of the variation that cannot be explained is called **uncontrolled variability**, and any source of uncontrolled variability in the design weakens it because it reduces the amount of variation that can be explained.

Issues of process evaluation.—Process evaluation of the sample program was nonexistent. Many process evaluation questions could have been asked that could have reduced uncontrolled variability. First, what about the service delivery aspect of the program? What did the guidance counselor and the language teacher actually do in running their groups? Perhaps the guidance counselor went beyond the curriculum, whereas the language teacher, who had no prior experience, had to struggle to present the material. Technical competence is not the only possible source of difference between the group leaders. The guidance counselor believed strongly in the program, having introduced it in the school, but the language teacher sought the position to gain counseling experience, not because of personal commitment to the program. Differences between the two group leaders were a first source of uncontrolled variability in the design.

What about the nature of the participants? We have no information about them. Note that there were a number of routes into the program. A student could volunteer without any contact from the school staff or could be drafted into the program. Possibly the students drafted by the guidance counselor were a select group with special problems, whereas those drafted by the language teacher were especially bright students because they were taking foreign languages early in their academic careers. Finally, all teachers were asked to refer students. Thus another source of uncontrolled variation was the nature of the participants, including the mechanisms by which they entered the program.

What about the extent of participation of the students? We don't know whether each participant actually experienced the program to the same extent. Maybe some students attended all sessions while others attended almost none. This expands our second source of uncontrolled variation to encompass not only who the program is reaching, but to what extent as well.

What of the quality of the relationship between the group leaders and participants? We have some inkling that the quality differed between group leaders, because the guidance counselor retained 81 percent

of the initial participants, whereas the language teacher retained only 26 percent. For a given level of technical competence, some staff will have better relationships with participants than others, producing yet a third source of uncontrolled variation.

Note that while this differential **attrition,** or dropout of participants, might reflect the quality of relationship, it might also result from the different technical competence of the two leaders. Or there might be some simpler explanation. For example, the language teacher had a number of advanced students in her group and the local high school started a special program for them that conflicted with the schedule of the self-concept program.

The evaluation design has obviously failed to give any information about the nature of the program as delivered, the nature of participants and their level of participation, or the quality of the relationship between program staff and participants. The bottom line is, we don't know whether the program as designed was ever delivered to the participants for whom it was intended. Without this information, questions of whether the program worked seem either presumptuous or preposterous.

Issues of outcome evaluation.—Let us assume that a program of known characteristics had been delivered and that participants did receive the program as planned. In that case, issues of outcome evaluation are at the heart of the judgment as to whether the program had the desired effect on participants. These issues encompass four phases of an outcome evaluation:

- o At the design phase, how participants and nonparticipants were selected.
- o At the measurement phase, how the variables were chosen, and then how they were measured.
- o At the analysis phase, whether the appropriate statistical tests were employed and whether the evaluation design was sensitive enough to detect program effects if they existed.
- o At the interpretation phase, to what extent one may give meaning to the data and generalize the findings.

Design phase.—When we ask whether a program is effective, we are really asking whether participation in the program has changed individuals from the way they would have been had the program never existed. It is not enough to simply measure changes within program participants. No matter how much change takes place, we have no foundation to argue that the change is due to the program. That argument can only be made by comparison. The ideal comparison would be created by turning time back—by repeating history with the one difference of interjecting the program during two otherwise identical passages through time.[1] In the example, we would then compare the individuals with themselves at the conclusion of the two time periods. Any differences could then irrefutably be ascribed to the presence of the program—we could then prove causality.

Since time cannot be turned back, other, less than ideal comparisons must be found by playing a scientific version of the game,

What would have happened if . . . ?

We can approximate what would have happened if the program had not existed by comparing two groups as identical as possible except that one group does not participate. The experimental or treatment group participates in the program, the comparison group does not. If the two groups are comparable at the outset and differ only on the variables of interest after program intervention, program participation probably produced the difference.

The comparability of the groups is critical. The sample evaluation included no systematic construction of comparable groups; only extraordinary luck might have produced participant and nonparticipant groups that were initially comparable. So,

the best evaluation requires
comparable treatment and comparison groups.

Another, less elegant way to approximate "what would have happened if" would be to conduct an extended series of measures over time on the participants, both before and after the program. Then, if a sharp discontinuity emerges in this **time series** once the program is introduced, the difference between expectations based on past measures and actual later findings is probably due to the program. A major problem with this approach is that we still cannot rule out the effects of **history,** of events or conditions that in addition to the program might influence the measures. It is far easier to rule out such confounding effects, if they exist, using a comparison group.

The major problem in making comparisons is in the selection of subjects. Problems relating to subject selection continue throughout the evaluation. Even when comparable groups have been constructed, attrition of treatment or comparison subjects during the evaluation weakens comparison. At the design phase, attempts should be made to estimate the amount of attrition and to devise ways to minimize it. In the sample evaluation no consideration was given to this problem.

Matters of design go beyond subject selection. All matters of process evaluation are ideally settled at the design phase, before the evaluation begins, and not as an afterthought once the evaluation is started. It is possible to have designs that are unable to detect program effects. Evaluators have the responsibility of designing evaluations that can detect effects of programs, if they exist. The number of participants is a critical part of this issue. In the sample evaluation, the number of participants was abysmally small, particularly at the posttest.

Confidentiality and informed consent are also design issues. The guidance counselor in the sample evaluation weakened the already insensitive design by not providing the information necessary to match the pretest and posttest of individuals. Confidentiality does not require a complete lack of identifying information. One can ensure confidentiality and still be able to match pretests and posttests. Finally, ethical issues of withholding potentially beneficial treatment from participants assigned to comparison groups must be thrashed out at the design phase.

Measurement phase.—Issues at the measurement phase can be classified in two categories: what should be measured (already discussed) and how program outcomes should be measured.

Measurement of outcomes is usually equated with the administration of paper-and-pencil tests, but measurement goes beyond this. Behavioral observations at the one extreme and formal records at the other extreme can be used to measure the same variables. Regardless of the approach to measurement, a number of standards must be applied. Are the measures suited to the population being measured? The guidance counselor in the example evaluation did not consider whether the students could understand the items on the self-concept test, a test that had been tried only with high school students. The content of the measure is critical: for example, items about whether individuals feel confident of being accepted by a good college are better suited to high school students than to younger students.

The **reliability** of a measure, its stability over repeated measurements, is also a critical matter. If the same test measures something twice, and the scores of individuals change unpredictably, then the measuring instrument is unreliable. We would, for example, throw out a bathroom scale that showed our weight to vary by 10 to 20 pounds each time we got on the scale. Such measures with a lot of "wobble" introduce another source of uncontrolled variability in the design. In the sample evaluation no attempt was made to establish the reliability of the measures, that is, to find out whether the measures were stable.

An equal problem is whether the measures are valid. Just because the school psychologist thought he had created a test of self-concept doesn't mean, in fact, that the test measured self-concept at all. The **validity** of a test, that it measures what it purports to measure, needs to be established. Just because a measure is reliable, does not guarantee that it is valid. However, reliability is a necessary condition for validity. It is pointless to ask what we are measuring if we are unable to measure it in a stable way.

Thus evaluations may fail to show program effects due to measurement failures in reliability and validity. The school psychologist's self-concept test was of unknown reliability and validity. It is possible that the participants' self-concept did change, but that the self-concept test, being unreliable, invalid, or not suited to participants, failed to detect the change. In the same manner, the sample evaluation's drug use measure may have been inappropriate for this particular group, for example, by emphasizing drugs that students were not trying, while failing to consider other drugs that were popular.

In a good evaluation, great effort is expended to develop sound measures. For example, the evaluator could ask to try out instruments on individuals similar to the participants, and perhaps to test them more than once. He might ask staff members to participate in the process to study the test administration procedures. In validating a self-concept instrument, the evaluator might ask staff members to identify some students with good self-concepts and some students with poor self-concepts and then see whether the test scores concur with these judgments. Where school records are used, the evaluator may want to check on their accuracy before using them in an evaluation. The sample evaluation failed to deal with the issues of measurement that are at the heart of good evaluation.

Analysis phase.—Some evaluation designs are unable to detect real effects of the program. When we say "detect real effects" we mean that a statistical test confirms a true change in some measure.

The ability of a statistical test to detect real effects is called the **power** of the test. Statistically speaking, we call the change from pretest to posttest averages the systematic effect. But in addition to systematic effects, there are other, uncontrolled sources of variability. Statistical tests work by comparing the extent of systematic effects noted in the data with the amount of uncontrolled variability in the data.

The sample evaluation reveals numerous sources of uncontrolled variability—two vastly different group leaders potentially selecting vastly different types of students into the program, with unknown levels of participation using measures of unknown reliability and validity. To use nontechnical terms, all this noise or slop in the design obscures whatever systematic change might have existed. As designed, the evaluation was almost doomed to show either no change or uninterpretable change before the data were ever collected.

Much could have been done to increase the power of the example design. Ways to increase power include increasing the number of participants, linking the pretests and posttests of individual participants, looking at the effects of each group leader, and gathering other pretest measures that are related to self-concept.

Interpretation phase.—Let us pretend for a moment that the sample evaluation had been properly designed with comparable treatment and comparison groups, and that appropriate data analysis led to the conclusion that self-concept had improved by virtue of program participation in the guidance counselor's group but not in the language teacher's group. How may we generalize the findings for future implementation of the self-concept curriculum? First, we must ask to what population of children the results apply. Second, ask to what extent the program effects would generalize to other group leaders and to other ways of measuring the same outcome variables, such as self-concept.

The answer to the first question is obvious. The results apply only to the population of individuals from whom the participants were drawn. Does this mean that if the program worked for these students, it will work for the student body at large? Not necessarily. These participants, selected through volunteering or being drafted, were not representative of the school population. With more complete information on the participants we could generalize about the type of student who might respond to the program. The findings cannot be generalized because the evaluation failed to identify a clearly defined target population and draw a sample representing this population.

Another problem appears if the program works with one group leader but not the other. We must then return to process questions about each leader and the quality of her relationship with the participants. The possibility exists that change was due to the characteristics of the group leader rather than of the curriculum. Change can come from a variety of sources. The same sort of question can be raised about the measurements: was any change or lack thereof peculiar to the particular test employed, or would the same results have been found with other measures of self-concept? In all, we ask to what extent evaluation findings are peculiar to our program and the measurement of its outcomes.

The validity of an evaluation.—Every issue discussed so far speaks to whether evaluation results give a valid picture of program effects. Four frequently discussed types of validity provide a way of thinking about the quality of an evaluation.

Construct validity.—We can scrutinize a program by asking whether we have done what was intended when we translated the original theory to program goals and then operationalized the goals to the program activities. To begin with, we have a set of abstract notions, or constructs, about what we are trying to transmit through the program. We also have what we are trying to measure as outcomes of the programs, for example, decreased drug abuse, improved self-confidence, increased acceptance of responsibility. The extent to which, first, program theory relates to program practice and then to evaluation activities is referred to as construct validity.

Internal validity.—If a change has been noted in program participants, we still need to ask whether the change is attributable to the program or to some other factor. For example, if participants' drug use decreases after a big crackdown on drug dealers in the town, we wouldn't be able to clearly attribute the decrease to the program unless we had some data from an appropriate comparison group. The ability to attribute change to the program as opposed to change from other sources is the internal validity of the evaluation. Whether an evaluation has internal validity is largely determined by the presence of comparable nonparticipant groups in the design.

External validity.—All questions of to whom, and to what situations, the results of an evaluation can be generalized are matters of external validity. A design may be internally valid but have poor external validity due to the highly restricted sampling of the participants or the unique conditions under which the program occurred.

Statistical conclusion validity or conclusion validity.—Several times we have questioned whether the design was powerful enough to detect program effects by a statistical test. In fact, any set of data may be analyzed in a number of ways, some more appropriate than others. Issues of statistical power and appropriateness of analysis can be summarized by asking whether the statistical manipulations of the data led to an accurate assessment of whether or not the scores of program participants changed. These are issues of statistical conclusion validity, or conclusion validity.

In essence, accurate evaluation findings that are scientifically sound and programmatically useful are difficult to achieve. The review of the example revealed numerous **threats to validity**, or failures of the design to permit sound conclusions about program effectiveness. For example, the small sample sizes and the unreliable measurements are threats to statistical conclusion validity; the lack of an adequate comparison group is a threat to internal validity; the lack of documentation of program activities is a threat to construct validity; the lack of documentation of the nature of participants is a threat to external validity.

To summarize, confusion can occur

at the beginning
inside
outside
and at the end.

ISSUES IN EVALUATION METHODOLOGY

The previous discussion of evaluation issues has separated them into those of process and outcome evaluation. We continue this distinction addressing first techniques and terminology in process evaluation, then some important technical issues of outcome evaluation.

Process Methodology

There is some disagreement in the field of evaluation research about the appropriateness of describing the gathering of information on program process as evaluation. Some purists would claim that since evaluation by definition makes judgments of worth, any information which simply describes an object or phenomenon is not, in the true sense of the word, evaluative. Others argue that since description is a necessary prerequisite for determining worth, it is entirely appropriate to consider it as an evaluative activity, at least by implication. We take the latter position and claim that, depending on the stage of program development, it is reasonable to develop an evaluation design that consists solely of process information. Obviously, outcome evaluation provides more information, but even the best outcome evaluation will include and build on process evaluation.

Process evaluation can be used to provide feedback for internal monitoring, to guide resource allocation, and aid ongoing program development. It can be used to provide accountability to funding sources and to illuminate the changing nature of a program as it evolves. In this sense process evaluation is no more nor less than management information and can be an end unto itself.

Process evaluation is also necessary for linking outcomes to key program components. A comprehensive evaluation tests hypotheses about the influence of specific program characteristics and activities on various outcomes. A careful process description of the program is necessary to understand the findings and to replicate both the program and its evaluation.

A basic distinction in process evaluation is between **input** and **process**. To appreciate what happens in a program, it is necessary to know what has been brought to it. These inputs include human and physical resources and the milieu in which the program operates. Each contributes directly to the actual operation of the program.

Program inputs.—Human resources include mainly staff and participant characteristics brought to the program. Important staff characteristics include qualifications as measured by educational level, training, and experience. Formal education alone is not a sufficient measure to judge abilities. Consideration must also be given to training and experience specific to the field of alcohol and drug abuse prevention. Involvement in workshops, conferences, work activities related to prevention, and community involvement all play a part, as do acquired skills in specific prevention techniques, such as values clarification or

alternatives strategies. Such skills are also important for administrative staff, along with experience in their expected roles. One basic measure of staff effort is expressed as full-time staff equivalents (FTE). These can be calculated by type of staff activity for both paid and volunteer staff.

Client characteristics include a range of demographics, dependent to some extent on the type of prevention program. Basic demographics should be collected, such as race/ethnicity, gender, age, grade level, family structure, or socioeconomic status. This information can help to determine if the program is serving the intended target population. A major issue is the extent of cultural disparity between staff and participants, and its effects, both positive and negative, on the program process. These effects are a question for both process and outcome evaluation.

Both staff and participant inputs should be measured at program inception and at key points during development and at the study's conclusion. This allows the choice of a stable period for analysis and provides information on changes over time that could have a direct bearing on program outcome.

The beliefs, values, and attitudes brought to the program by staff and clients alike will have a major impact on program effects. Staff and participant attitudes toward alcohol and drug abuse, and the extent to which these views are similar, are important input considerations. Staff attitudes toward prevention will greatly affect program activities. It is a truism that events often coincide with our expectations of what will happen. Staff attitudes toward drug abusers and beliefs about the etiology of abuse will greatly affect the approach to program tasks. Stated role expectations for both staff and participants will influence performance. Organizational as well as individual expectations, and any discrepancies that exist between them, will greatly influence program process.

Basic demographic data should be collected for all participants and staff. Personnel folders should detail past and ongoing staff education, skills, and training. Data on attitudes and expectations can be gathered from interviews (ranging from structured to open-ended) and observations by trained observers.

Physical resources include space, equipment, and supplies. Each type of resource can be disaggregated for future analysis in relation to program functions and activities. Physical resources are more amenable than human resources to easy conversion to a common measure—money. Money, in and of itself, is not viewed as a true resource. Rather, it is a means of obtaining commodities and measuring their value. If a program has a cash balance of $50,000, this means little except as it is translated into the number of counseling sessions or the equipment it will purchase. Monetary conversion of resources, process, and outcomes becomes a foundation for later standardized cost comparisons.

Environmental variables directly affect the workings of the program. Descriptions of the socio-economic structure of the community and its population are necessary to develop a needs assessment that clearly identifies the potential participant pool. The incidence and prevalence of social problems are important, particularly those directly related to alcohol and drug abuse. For school programs, measures of variables such as disciplinary actions, school grades, and vandalism are needed.

Input data provide a basis for determining if the program as implemented serves the intended target population, and if this population adequately represents those shown in need. Other relevant questions are whether the staff meet necessary standards and if resources are sufficient to accomplish program objectives. Specific questions must arise out of the particular program situation.

Program process.—As with inputs, program process can be measured using both qualitative and quantitative indicators. Three basic aspects of a program's functioning should be examined during a process evaluation:

- o organizational structure
- o patterns of interaction
- o program service delivery.

The field of organizational analysis is growing rapidly, with increasing sophistication in methods. For example, structural analysis compares formal patterns, as found in organizational charts, with actual patterns of authority, responsibility, and communications. Systems analysis is more concerned with measuring the dynamic aspects of the organization. One useful way to describe the organization is presented by Cline and Sinnot (1980), using five interdependent dimensions.

The task dimension describes the organization as a set of tasks interconnected by authority and accountability relationships. Major tasks and the activities undertaken to achieve specific objectives are identified and described. For school based prevention programs, two possible data sources are the course curriculum and job descriptions.

The function dimension describes the organization as a set of operating units interconnected by the ways in which they act and react to one another. While the task dimension focuses on activities within units, this dimension emphasizes the interrelation of units in achieving organizational goals. A common data source is the organizational chart, which is taken as a starting point for examining actual structural relationships.

The information dimension is concerned with mapping the flow of information and identifying key decision points. This dimension is closely related to the task and function dimensions, in that decision-making is part of the formal functions of various individuals and units. This dimension represents the first step in an analysis of the decisionmaking activity.

The fiscal dimension describes the organization as monetary resources connected by budgetary and accounting relationships. The major focus is on the allocation of resources, which leads to measures of cost effectiveness. Budget and expenditure statements are the basic source of information in describing this dimension.

The personnel dimension, which describes the organization as a group of persons interacting on a daily basis, is probably the most difficult to express in quantifiable terms and is more likely to be described based on observations of interactions. This is a time-consuming process, with the observer's major task being to limit observations to the most important interactions.

An alternative to Cline and Sinott's approach encompasses the three basic aspects of function already mentioned--structure, interaction, and service delivery—and develops a comprehensive description of the organization as it attempts to achieve its goals.

The major emphasis of process evaluation is the delivery of services. An evaluation of services should describe intended content, the timing of delivery, and its integrity, that is, whether what is delivered matches what is intended. Quantitative measures can include the number of meetings or sessions, the number of participants, the ratio of staff to participants, actual versus expected attendance, and the physical surroundings of the service delivery.

Qualitative and quantitative methods.—Only recently have the arguments about the relative merits of quantitative and qualitative approaches started to reach a resolution. Cronbach, et al. (1980, p.223) provides the evaluator with a cautionary note:

> The evaluator will be wise not to declare allegiance to either a quantitative-manipulative-summative methodology or a qualitative-naturalistic-descriptive methodology. He can draw on both styles at appropriate times and in appropriate amounts. Those who advocate an evaluation plan devoid of one kind of information or the other carry the burden of justifying such exclusion.

Quantitative methods leading to hypothesis testing view the program as a fixed stimulus applied to the social system. These methods employ experimental designs and statistical techniques to determine if hypothesized effects occur. It is in this sense that Cronbach uses the term, "manipulative" methodology. The program is seen as a manipulation of an existing reality.

Qualitative methods employ participant observation, open-ended interviews, and other so called subjective approaches to examine the program as a system into itself, and as a part of larger systems. The emphasis is on what the program is and does as seen by those involved. In the past, qualitative methods were viewed by quantitative researchers (number crunchers) only as a way to develop and formulate hypotheses for future examination by objective quantitative methods. Now there is a growing recognition that the information from the two paradigms complement each other and that the issue of subjectivity versus objectivity should not be drawn along methodological lines (Patton 1978).

This issue is crucial to the evaluation of prevention programs, where the cultural mix of participants, staff, and community is a major factor in determining the structure, dynamics, and outcome of the program. The evaluator who doesn't appreciate the enormity of cultural effects throughout the entire evaluation process is likely to do a disservice to the program.

The information, both quantitative and qualitative, that could be gathered in a process evaluation is practically limitless. The major problem for the evaluator and the program manager is to decide what is essential for the evaluation. Within resource constraints, limits should be set to allow enough freedom to identify key elements related to goal attainment, without taking away from the full richness of the program.

Outcome Methodology

In this section, we present major technical matters that are critical to understanding outcome evaluation design and analysis. First, we cover the construction of comparable experimental and comparison groups for an evaluation design. Here, we consider **threats to internal validity**, which are either eliminated or produced by the construction of the experimental and comparison groups. Second, we consider **concepts of statistical inference**. Finally, we review some **concepts of measurement**, expanding upon definitions of reliability and validity.

Threats to internal validity.—Attributing change in program participants to the program itself requires proof that participants are more different after experiencing a program than they would have been had they not experienced the program. The strategy used to make the participant-nonparticipant comparison is to construct comparable groups that do and do not participate and compare the groups at the same points in time. Perhaps the most critical issue in outcome evaluation is how these comparable groups are formed.

An obvious way to select comparable groups is to match two groups on important variables. However, there's a trap in this—which variables to match. In a self-concept program, for example, we would want to match on variables known to be related to self-concept. While we may not be sure what those variables are, we suspect the list is long. If we try to match but miss some critical variable related to self-concept, then we can't claim comparability; our evaluation is undermined before we begin. Our theory for prevention needs to be carefully assessed to guide the variable selection process.

True experiments.—Another approach might be to take all the individuals who could be participants at any point in time and randomly divide them into participants and nonparticipants. If this is done with reasonably sized groups, (e.g., N=30), the result will be two groups theoretically comparable on all variables. But how does sampling theory lead us to this statement?

Imagine splitting a group of 100 people randomly into two groups by flipping a coin to determine each person's group membership. These groups should be approximately equal in height, education level, need for approval, anxiety; in fact, in every characteristic one might name. Why? Because the outcome of the coin toss is in no way related to any other variable, and the laws of probability are permitted to operate fully. The coin cannot tell how tall, how well educated, or how anxious anyone is. These variables (by chance) will be distributed equally across groups.

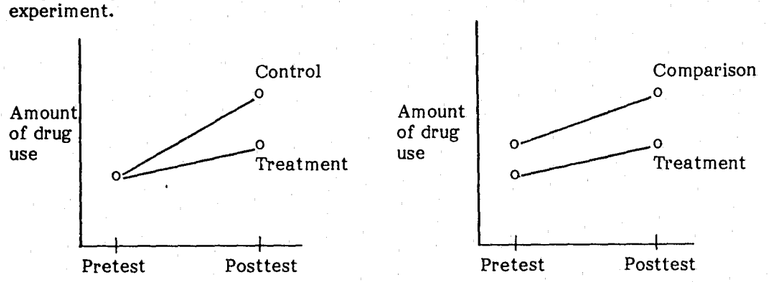
This method of constructing groups, referred to as **random assignment**, is the one method of constructing groups theoretically comparable on all variables that might influence the outcome of an evaluation. Experiments or evaluations using this method for constructing groups are called true experiments or randomized experiments.

Quasi-experiments.—Although true experiments are the most desirable, sometimes they cannot be constructed. For example, if the whole fifth grade of a school is to receive a program, no fifth graders remain to serve as controls. Ethical issues may also preclude withholding the program, even temporarily, from some potential participants. These situations call for quasi-experiments, a category in which the experimental and control groups are not constructed by random assignment. Unfortunately, in quasi-experiments, some internal validity is lost. This means that if one does find a difference between treatment and comparison groups at the end of the experiment, one cannot be certain that the difference was due to the program's effect. It could be due in part to differences that already existed between the groups.

So profound is the difference between true and quasi-experimental designs in yielding answers to evaluation questions that the groups in the two types of designs are called by different names. In a true experiment, the nonparticipant group is called a control group. In a quasi-experiment, the nonparticipant group is called a comparison group.

Internal validity in true versus quasi-experiments.—The reason for having control or comparison groups is to mitigate threats to internal validity, that is, to eliminate confounding effects that prevent attributing outcomes to the program. To illustrate, figure 1a shows one possible outcome of a true experiment involving a school prevention program. Both groups increase drug experimentation over the semester, but the group that participated in the program showed less increase. The program apparently retarded the rate

of increase in drug use. Now consider the same effect in a quasi-experiment in which volunteers were participants and nonvolunteers were controls. In figure 1b the comparison group also showed a greater increase in experimentation than the participant group. Is this difference clearly attributable to the program? No. The self-selected treatment group was less prone to use drugs than the comparison group before the program began. It is possible that the different initial levels of drug use, regardless of the program, influenced the rate of increase in drug experimentation. The main threat then to internal validity in quasi-experiments is the selection factor that brings the treatment and comparison groups into the experiment.



a. True experiment

b. Quasi-experiment

Figure 1. Some outcomes of true and quasi-experiments

Achieving random assignment through delayed treatment.—Randomly assigning individuals to receive or not receive potentially beneficial treatment is contrary to the belief that treatment should be readily available to all who wish it. A way to achieve random assignment and ultimately to have everyone participate is to delay but not to deny participation to some individuals. This useful technique for achieving random assignment is illustrated in figure 2.

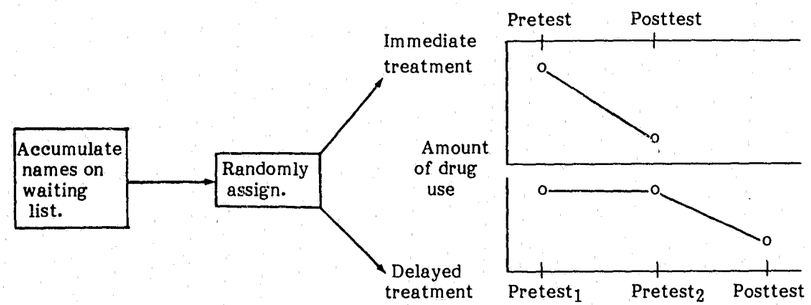


Figure 2. Waiting list technique for achieving random assignment

Suppose there were more applicants for program participation than program slots. One handles this by having people wait until slots become available. Assume there are 60 people on the waiting list and only 30 slots. The waiting list is used to construct true experimental and control groups randomly assigning the 60 individuals to one of two groups. An immediate treatment group enters the program without delay and a delayed treatment group enters the program after the immediate group leaves. The delayed treatment group serves as the control group, as shown in figure 2.

All individuals are pretested at the same time. Next, the immediate group receives treatment. When the immediate group completes the program, both groups are tested again. Finally, individuals in the delayed treatment group receive their posttest at the completion of treatment.

Attrition destroys the benefit of random assignment.—Groups constructed by random assignment at the beginning of an evaluation should be equivalent at the end of the evaluation if membership in the groups remains stable. Differential attrition, or mortality, of participants from the groups destroys the internal validity of true experiments. In evaluations, every effort should be made to keep subjects in the groups throughout the experiment.

Random assignment to program components.—There are not always waiting lists. In some circumstances, even temporary nonparticipants cannot be designated. The evaluation then may be a contrast among variations in programing, rather than between program and no program. For example, if there is a conventional program against which a novel program might be compared, then the random assignment might be to the conventional versus the novel treatment.

Regression effects.—In statistical analysis there is a tendency with repeated measures to regress toward similarity, or to the group mean. This is called a regression effect, regression artifact, or statistical regression. This problem is particularly acute when groups are selected on the basis of extreme scores, e.g., high drug use versus low drug use. True experiments control this problem to a large extent because groups are randomly selected rather than preselected. In quasi-experiments these effects can be troublesome because of the process of forming comparison groups. Comparing volunteers in a program with nonvolunteers in a comparison group is a very poor approach. The uncontrolled selection factors that determine who will volunteer undermine attempts to attribute any posttest differences to the program itself.

This is another dimension to the problem that arises when selecting subsets of individuals from two different groups so they match on specific variables. For example, suppose a prevention program is mounted in a school with substantial drug problems while the comparison group for evaluation might be drawn from another school with less drug use. An approach might be to test children of both schools on drug use and to select subsets of children from the two schools whose drug use levels matched. While this may appear to solve the problem of noncomparable groups on drug use, it does not, due to regression effects.

Regression effects occur because measures are not perfectly reliable. If the drug use scale is given twice, there will be different amounts of reported use. If the test were unreliable, a respondent with a very low drug use score on the first measurement would likely have a higher score on the second measurement. Why? Tests do not have perfect reliability because respondents change some answers between two test administrations. If a student gives a very low estimate of use the first time he took the test, the only way he can change his answers is to report higher use levels. In contrast, if a respondent reports very high-use the first time, the only way his answers can change is to lower levels.

Regression effect has nothing to do with the true level of the behavior.
It has to do with the unreliability of the measure.

For example, suppose the test asked, "How many times did you smoke marijuana last month?" and alternative choices were 1-5, 6-10, 11-15, 16-20, 21-or-more times. A frequent user may puzzle over the choices 16-20 versus 21-or-more, but can't really decide, and arbitrarily picks 21-or-more. He's got a high use score. The next time the subject encounters the item, he still can't decide and randomly chooses the category 16-20. The subject's drug use hasn't changed. What's changed is his random choice of responses in an uncertain situation. The same argument goes for the low end of the scale.

If we sort respondents into two extreme groups based on the drug use score on the first test administration and retest them, regression artifacts should cause the data to look like those in figure 3.

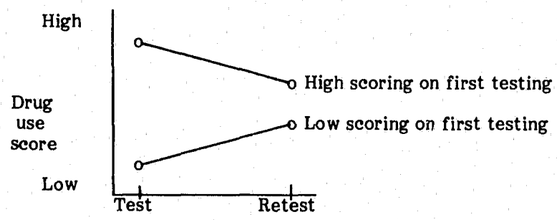
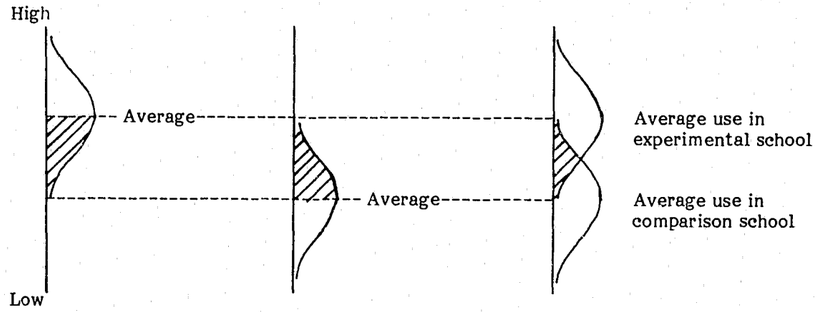


Figure 3. Regression artifacts in a single group

The extreme scorers on the first test administration have scores closer to the average (or middle) of the drug use range of their group on the second administration. The amount of change is a statistical function of the reliability coefficient of the test being used. The less reliable a test, the greater the chance of a regression artifact.[2] As shown in figure 3, one could not attribute any change in outcomes to program effects. One could not conclude that the program lowered drug use levels for high users or that the program increased drug used behavior of those in the low-use category. Obviously, the process of selecting treatment and control groups has serious implications for correct interpretation of evaluation data, given the imperfect world of measurement.

One way to attempt to achieve comparability is to match students on drug use from two schools, where average drug use levels differ. This situation is illustrated in figure 4.



a. Drug use in experimental school

b. Drug use in comparison school

c. Possibilities for matching on drug use scores

Figure 4. Pretest drug use from two schools

There are several major problems with this matching approach. The experimental school has a higher average drug use than the comparison school. Students from the two schools are matched together by use scores. Only those students in the shaded area in figure 4c can be matched because the school averages are different. The greater the difference in average scores, the fewer matches can be found. Thus the first problem with this approach is that the sample size available for analysis is smaller, reducing the power of the analysis.

Further, the students are being matched on only one factor—their drug use scores. The unstated and undoubtedly false assumption is that students in the two schools are similar in all other respects which have a bearing on drug use. However, introducing other matching variables would further reduce the number of possible matches, leading to even smaller sample sizes.

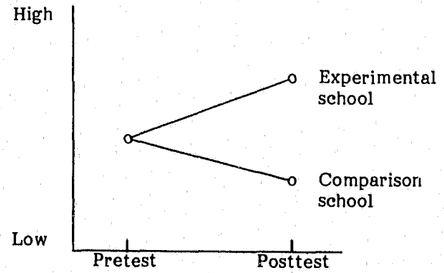


Figure 5. Results of matching from nonequivalent groups
(Each group regresses to its own average.)

Finally, the matching approach can substantially increase regression effects. The lowest scorers in the experimental school can be expected to have a higher average score on retesting. For the same reason, the highest scorers in the comparison school will have lower scores on retesting. And these are the very students we have chosen by matching scores. As figure 5 shows, it will appear that the program has caused increased drug use, while comparison subjects (with no intervention) will appear to have decreased drug use.

Statistical regression, or regression effects, operate whenever extreme groups are used in designs. They are subtle and treacherous and most likely to creep into evaluation designs when matching is used to achieve apparent pretest equivalence in quasi-experiments.

To summarize:

true experiments are more desirable because
they overcome threats to internal validity.

Concepts of statistical inference.—When we do an evaluation our interest goes far beyond the particular individuals who participated in the evaluation. We wish to generalize to other individuals who might participate in the program. Put another way, concluding from an evaluation that a program worked and ought to be continued or tried elsewhere, really predicts that the program will work in the same way for other individuals in a comparable setting.

We base conclusions from our data on the rules of statistical inference, which constitute a logical system for making such generalizations based on probability theory. We will review this logical system defining many of the terms associated with it as we go.

Populations versus samples.—The first necessary distinction is between populations and samples. A population, for our purposes, is a clearly delimited group of individuals, say, all the fifth graders in a particular school system. A sample is just a subgroup from that population. Our evaluation on randomly selected samples from a population allows generalizations about that population, and statistical inference is the basis of the generalizations. If we could study the whole population, we wouldn't need statistics.

Power and Type II error.--Although the purpose of statistical inference is to generalize from samples to populations, it's easier to understand statistical inference if we work backwards. Assume two populations of individuals who are identical. More specifically, they are identical on the variable of self-concept. Put in the usual statistical terminology, the two populations have identical self-concept arithmetic means. Arithmetic means are what we commonly describe as averages; they're usually referred to as means in the context of statistics. Now suppose we assign one population to a self-concept program and the other population serves as a control group. At the end of the program, the mean self-concept score in the population that participated in the program is five points higher than that of the control population. That is, there is a true difference between the population means. We conclude, all other things being equal, that the program produced the five-point advantage.

Given this true difference in population means, suppose we do the following exercise. Draw a random sample of 25 people from each population and note the difference in mean self-concept in the two samples. Having recorded this difference, we return the people to the population and draw another pair of samples, note the difference between their means, and return them to their populations. If we do this repeatedly, we will observe that the difference between the means will usually be around five points, in favor of the sample from the participant (treated) population. Sometimes the difference will be greater than five points, still in favor of the sample from the treatment population, and at other times, the difference will be smaller than five points. In a few cases, perhaps, the sample from the control population has a higher mean score.

That is, individual samples do not perfectly reproduce
the populations from which they were drawn.

To continue, suppose that instead of having repeated measures of the populations, we could only look at one pair of samples. On the basis of the sample self-concept means—in the treated versus untreated samples—we would have to draw a conclusion as to whether the program worked. What sort of rule might be used to reach a conclusion? We could use a rule that says, "if the treated sample is above the untreated sample by any amount, decide that the program worked." Now, for most pairs of samples we drew, there would be a difference in favor of the treated group, and we would correctly conclude that the treatment caused a gain in self-concept. In statistics, a correct conclusion is one that reflects what is actually true of the populations from which the samples were drawn.

In most instances, we would correctly conclude that the program had caused an increase in self-concept. But for some pairs of samples, those in which there was no difference or perhaps a reversal, we would incorrectly conclude that there was no effect. This sort of error is called a Type II error—more about this later. Note that this problem of failing to detect a difference that really exists in the population is precisely what we were concerned with when we discussed the statistical power of an evaluation design. Power and Type II errors are opposite sides of the same coin, that is, detecting versus failing to detect a true effect of a program.

Or, in other words,

when you improve the power of a design,
you reduce the chances for a Type II error.

Type I error.—Now consider another situation. Once again, begin with two identical populations, and treat one population with the self-concept program. This time, however, assume that the program has no effect; that is, the two population self-concept means are identical. Again, imagine taking pairs of samples from these populations and calculating the difference between their means over repeated samplings. Most differences will be about zero. But, from time to time, the mean of the sample from the treated population will be somewhat higher than that of the control sample. In those instances we can make the error of concluding that the program worked, when, in fact, it did not really work in the population. This sort of error is called a Type I error.

Keeping in mind equal population means (the program had no effect) versus unequal population means (the program had an effect), we can differentiate the two situations in the form of a pair of hypotheses. One hypothesis, the **null hypothesis**, says that the group means are equal; the program had no effect. The other hypothesis, the **alternate hypothesis** or research hypothesis, says that the program worked, that is, the group means are unequal. Note that these two hypotheses exhaust the possibilities for the outcome of an evaluation. If we can amass evidence that one hypothesis, the null hypothesis—of no effect—is false, then we are simultaneously amassing evidence that the alternate hypothesis—there is an effect—is true.

Now, in the real world we have no knowledge of the population; we are trying to infer what exists in our population from looking at sample data.

Based on probability theory
we make conclusions about the population(s)
and then qualify those conclusions
by stating the odds that they are wrong.

Again, let's say we observe a five-point advantage in self-concept in our treated over our control sample. We make the statistical decision to reject the null hypothesis, that is, we conclude that the populations must be different because the samples are different, as in the first situation discussed. But there's another possibility; the population means might really be the same, as in the second situation, but by chance we've drawn samples that make it seem that the populations are different. Through probability theory we are able to determine the chance that we will have made an error in rejecting the null hypothesis, that is, a Type I error.

The probability of a Type I error is called the **level of significance** of the statistical decision to reject the null hypothesis. In evaluation reports, you will see sentences like, "The treatment group had a significantly higher self-concept mean than did the control group ($p<.05$)." "Significantly higher" says that the null hypothesis--the group means are equal—is being rejected. The ($p<.05$) in parentheses gives the probability that this conclusion is wrong. This is another way of saying there is less than a 5 percent chance ($p<.05$) that the decision to reject the null hypothesis is wrong. Note that we are worried only about Type I error when we are rejecting the null hypothesis, that is, concluding that the groups are different, or that the program worked. A final point, the lower case Greek letter alpha (α) is sometimes used to indicate the probability of Type I error. When people ask what alpha level you're using, they're asking how much Type I error is associated with your statistical decisions. It is only by convention that no more than 5 percent Type I error is acceptable to reject the null hypothesis.

Power analysis.—Historically, science is conservative. Hence the emphasis has traditionally been placed on Type I errors. Nobody wants to conclude that some intervention or treatment has an effect when it doesn't. In the context of program evaluation, however, there also should be enormous concern for Type II errors—of failing to conclude that an effective program is effective because the power of the design is very low. The lower case Greek letter beta (β) is used to note the probability of a Type II error.

The power of a design depends on a number of factors, such as the magnitude of the program's effect. In the previous situation in which the treated population was five points higher than the control population, samples from the two populations could sometimes be expected to have the same means and to lead to Type II errors. If the difference between means in the populations had been larger, say a 20-point difference, the chance of drawing samples that showed no difference would have been much smaller, thereby decreasing the probability of a Type II error, or conversely increasing the power of the design.

Uncontrolled sources of variability in an evaluation design decrease the power of the design. To determine the power of a design, consider the amount of difference between the populations relative to uncontrolled variability. The term effect size is used to mean the amount of difference, or the effect of the program, relative to a measure of uncontrolled variability. The amount of uncontrolled variability is always considered relative to the number of subjects in the design. Increasing the sample size increases the power of the design.

An analysis of the power of a design is best performed while the evaluation is being planned. To accomplish this, an estimate of the effect size (difference relative to uncontrolled variability) is required. Evaluators will often ask if any pilot data for a program already exist or can be collected before a full-scale evaluation is mounted, to make an estimate of effect size. With such an estimate, the number of subjects required to detect those effects in a evaluation design can be determined. Sometimes the effects are so small that enormous numbers of subjects would be required to detect them. In such instances, using large numbers of subjects, the execution of a labor-intensive and costly evaluation may not be warranted.

Power analysis may also be performed after an evaluation. This is particularly critical when the evaluation has detected no effect of the program (the null hypothesis was not rejected). In this case the concern is whether the design was so weak in terms of statistical power that an effect that really existed could not have been detected in the design.

Some common statistical tests.—Statistical tests are calculations to determine what the probability of a Type I error (false rejection) would be if the null hypothesis were rejected. If the probability of Type I error is low based on a statistical test, say less than 5 percent, then we would typically reject the null hypothesis.

Many tests can be used, and the choice depends on the nature of the data. Here we mention only some very common tests. The simplest is the t-test, which tests whether two groups are different or not on some measure, using the mean. If there are more than two groups in the design, Analysis of Variance (ANOVA) is used for the same purpose, to test whether the several groups in the design differ.

Analysis of Covariance (ANCOVA) is a statistical procedure that does what ANOVA does but also adjusts for initial uncontrolled sources of variability, increasing the power of the statistical design. For example, if participants vary widely among themselves in self-concept before the program begins, then it will be difficult to detect later changes. ANCOVA reduces this uncontrolled variability by linking various pretest and posttest measurements on each individual.

In quasi-experiments, where the treatment and comparison groups are not equivalent, such statistical procedures must be employed to tease apart two potential sources of difference between groups at the end of the experiment: the effect of the treatment, and the initial differences between the groups. Any statistical adjustments are approximate at best—they do not guarantee accurate estimates of the effect of the treatment.

Concepts of measurement.—When we scrutinized the sample evaluation, we identified two important properties of measures. First was reliability, or the stability of a measure. Second was validity, or the extent to which a test measures what it purports to measure.

Reliability.—The definition of reliability really encompasses two aspects of measurement, stability and internal consistency. Stability means that if one takes a test twice and doesn't change on the trait being measured, then the test score also should not vary much over repeated testing. The usual way in which this type of reliability is established is by administering the same test twice to a group of people and computing a measure of the extent of agreement between the two test results. The basic measure used is called a **correlation coefficient.** The coefficient will equal 1.0 if there is perfect agreement between the two measurement points. It will equal zero if there is no relationship between the scores at the two measurement points. It will be negative, somewhere between 0.0 and -1.0, if scores get reversed over the two measurement points; that is, if the high scorers at the first measurement point become the low scorers at the second measurement point and vice versa. The correlation coefficient is referred to as a reliability coefficient in this context.

The second aspect of reliability, internal consistency, is a measure of the extent that all the items or questions on a test agree with one another, or measure the same thing. If we have a self-concept scale, a person with a poor self-concept overall should respond in the same way across all items on the self-concept scale.[3] A common measure of such reliability is Cronbach's Coefficient Alpha, another index that equals zero if there is no consistency among items, and approaches 1.0 as internal consistency increases. The Kuder-Richardson formula is another common measure of internal consistency.

These measures are appropriate only with homogeneous tests, those where all the individual items are measuring one thing. It is possible to increase reliability of the score on the whole test by increasing the number of items on the scale. Statistical estimates have been created of the extent of increase in reliability to be expected by increasing the number of items. The classical estimate is the Spearman-Brown prophecy formula.

Validity.—Validity of a measure is a broad concept. There are a number of ways to establish validity. At the lowest level is face validity; that is, the content of the items seems to agree with what the test is supposed to measure. For example, in a test of depression, if it appears that people who are depressed will respond in one way, while nondepressed people will respond in another way, then the test has face validity. Concurrent validity means the agreement of the test with other measures of the same trait taken at the same time. If psychotherapists identify a group of clients who are depressed and a group who aren't depressed, and test scores agree with these judgments, then the test has concurrent validity. Predictive validity means that the test is able to predict accurately what will happen in the future. If we construct a scale of Propensity to Experiment with Drugs, and scores on this test taken at the beginning of the school year are related to the amount of drug experimentation that occurs throughout the following school year, then the test has predictive validity.

Construct validity is the most complex and abstract of the validity notions. It considers how the measure of a variable relates to other variables, on some theoretical basis. For example, depression might be closely related to poor self-concept and lack of hope for the future. We might not expect depression to be related to intelligence. Assessing how well a measure's association or lack of association with measures of other constructs adheres to our theoretical notions is at the heart of establishing construct validity.

The assessment of the validity and reliability of tests and other measures is an arduous process. Often evaluators will suggest that existing tests on which validity studies have been performed be used, in order to avoid having to study the validity of a test created especially for a particular evaluation.

The Worth of the Program

So far, a major thrust of this chapter has been on the ways in which outcomes can be specified and measured. Effective outcome evaluation design and analysis can provide an answer to the question of whether a program causes effects that differ significantly (in a statistical sense) from no program or in comparison to other types of programs. But the question remains—is the program worth the effort?

Worth, or value, is defined at a number of levels and along many dimensions. In a most general way, all of human culture, all the social and political forms we participate in, are concerned with the continual redefinition of worth.

Much of our social and political life concerns the valuing of material things, even as we link these to more symbolic, ideal, or spiritual concerns. The material resources available to maintain and enhance human life come in limited quantity. In most circumstances, therefore, we must make continual choices to use material resources for some purposes, leaving fewer for other purposes. All such choices involve both material resources and the purposes we want them to fulfill.

In the last quarter century much work has been directed at developing methods for valuing the material worth of social programs, under the general categories of cost-benefit and cost-effectiveness analysis. Much of the following discussion about the worth of the program focuses on basic concepts from these analytic approaches. Remember, however, that any such economic analysis applied to alcohol and drug abuse prevention is itself worthwhile only in conjunction with other social and political approaches to valuing. Economic analysis is an extremely fruitful way to look at a prevention program but is not a substitute for continuing concern—and conflict—about the human values programs are intended to enhance.

When we ask, "Is the program worth the effort?" in economic terms, we are really asking about the relationship between the value of resources consumed and the value of outcomes produced. When resources are invested in an activity, we expect that the activity will be effective in producing benefits and that the

benefits will outweigh the costs. The greater the benefits in relation to the costs, then the more economic worth there is to the activity. In order to measure worth, we must examine the:

- o **consumption of resources,**
- o **effectiveness** of the activity, and
- o the **relationship** between the two.

Consumption of resources.—The costs of a drug program are the values of the resources used for its activity. Costs are most often expressed in units of money. Money, then, is a measure of cost; it is not in itself a resource. Thus we can talk about the cost in dollars per mile of a vehicle operated for a program, or dollars per hour of a group facilitator's time as measures of the value of these resources.

To the economist, the cost of a resource is the value of its next best alternative use. If we have $100 and only two choices for disposal—put it in a savings bank at 5 percent interest or buy a one-year membership in a health club—the economist would claim, and rightfully so, that the true cost of the membership in the health club over a year's time is $105, the value at the end of the year if we had invested the $100 in the savings account. In the same way, the cost of a facility for a prevention program equals the value of what might have been produced by using the same facility for other purposes. This is a fine distinction but an important concept. In using resources for prevention programs, we deny their use for other activities. The cost of the resource is, then, its foregone opportunity—what we lost by not using it for other purposes—not what we paid for it. However, in a competitive, economically motivated market, the market value is the true measure of cost. A facility that is rented to the program at the going rate has a cost equal to the rental fee.

But where a market is not perfectly competitive or does not exist, cost estimation becomes more complicated. For example, the use of a facility may be donated to a program. The foregone opportunity cost for the facility might be assigned based on current rental rates. But what if the facility has been vacant and no one else was interested in using it? Although there are several ways of imputing costs in such situations, one common approach is to ignore the costs of otherwise unusable resources, on the theory that "the only free lunch is the one nobody else will eat" (Yates 1980, p.47).

Costs include more than physical resources and salaried staff. Resources such as volunteers, student interns, or evaluation consultants contribute to program operation. The values of these resources are in the worth of their time. Participants' time also has value. For example, a summer alternatives program might prevent participants from getting a job. Thus the opportunity costs of human as well as physical resources must be considered in calculating total program costs.

Direct and indirect costs.—Another dimension of costs is the distinction between direct and indirect costs. Direct costs are represented by the use of limited resources for producing services that would not be produced if the problem did not exist. If alcohol and drug abuse were not problems, we would not need prevention or treatment programs or law enforcement and criminal justice activities directed at the problems. Instead, these resources could be used for other activities that would enhance the social welfare.

Indirect costs represent the loss to society of what could have been produced if drug abuse did not exist. Rufener et al. (NIDA 1975) base their estimation of the indirect costs of drug abuse on the foregone earnings of abusers. This requires the assumption that increased unavailability for employment is causally related to drug usage. This unavailability can range from unemployment to work time lost for treatment, incarceration, or to the ultimate loss, premature death. Society must forego the goods and services that could otherwise have been produced, had the problem not existed.

Community and operations costs.—The above view of costing is referred to as the community or social perspective. It includes costs to the program and to various components of the community. While this perspective is comprehensive, the estimation of many social costs is difficult. This difficulty can be avoided by taking an operations rather than community perspective. The operations perspective merely looks at accounting entries in the program's books. This approach does not provide a complete listing of resources necessary to operate the program in the future and does not consider the foregone opportunity costs of resources. The operations approach also tends to bias costs in favor of programs that are socially appealing or that are located in communities that can afford donations of time and other resources.

Let us assume that the operations approach is taken to estimate the costs of a drug prevention program. The program is located in space donated by a local community organization with about one-third of full-time equivalent staff time consisting of volunteers. The resulting cost estimate could not be used as a gauge to predict what a similar program would cost in another community, where donated space might not be available or where volunteers might not be forthcoming. Also, costs could not be compared with other, less socially acceptable programs in the same community that might not attract as much donated community involvement.

There is not a clear dichotomy between community and operations costs. At a program level, one could decide to include costs and benefits that are not reflected in accounting ledgers.

The key is to keep costs and benefits
at the same level of generalization.

Present valuing.—Resources are consumed over time. If comparisons are to be made of the values of two resources, one of which will be consumed immediately and the other at some point in the future, then there must be some way to standardize the values to take the time difference into account. The economist's approach to this problem is to convert all resources into their present value.

Resources that must be spent immediately have greater value than those resource expenditures that can be delayed for spending at a future date. Resources that are not spent until a later date can be put to alternative uses until that time, producing a return. A penny saved is indeed a penny earned. The economist takes this into account through present valuing. Present valuing allows standardized comparisons between alternative choices for investments.

Suppose we intend to spend $10,000 each year for three years for a drug prevention program. Assume that the next best alternative use of this money would produce a 10 percent return. We will use this as the **discount rate.** Since the value of a resource is equal to what would be produced by the next best alternative, the present value of the $10,000 to be spent during the first year is only $9,091, because $9,091 invested today at 10 percent would produce $10,000 at the end of a year. Using the same procedure, the second year's expenditures have a present value of $8,264 (which would produce $10,000 at the end of two years if invested at 10 percent interest), and the third year's expenditures have a present value of only $7,513. Thus we intend to spend $30,000, but the present value of our future resources is only $24,868.

When we discuss the development of cost effectiveness and benefit analysis, it will become evident that the choice of discount rate plays a major role in comparing programs. Different rates produce conflicting results depending on the time frames of expenditures and benefits. For this reason, many analysts will report results using two or more discount rates in order to determine the effect of the rates on the findings.

In summary, several major issues must be carefully considered when developing a cost assessment. Not all costs are easily expressed in monetary terms. The level of detail in collecting data and reporting costs must be based on a consideration of how much accuracy is added to the final cost figures. Data must be available in sufficient detail to allow accuracy in reporting costs for variables that represent the greatest use of resources, without being unnecessarily specific. Certainly office supplies represent an important cost, for example, but one would not count the number of ball point pens used per month. But knowing the major costs of a program is only the first step in assigning worth. The second is in knowing the benefits, or positive effects, produced by the program.

Effectiveness.—When an analysis of costs and outcomes is conducted, the importance of identifying and testing for all relevant outcomes is brought home forcefully. For instance, although reduced delinquency might not be an intended program goal, if it occurs as a result of the program it should be considered as part of the worth of the program.

We have already discussed the major methods for determining if program outcomes are statistically significant compared to control groups or to other programs with similar goals. In a cost analysis, we must be able to specify the amount of change due to the program. It is not enough to say, for instance, that the experimental group had a significantly greater improvement in self-concept than a control group. We must know how much change can be directly attributed to the program.

Very often, we can obtain outcome data on a level of greater specificity than we can cost data. Most evaluation designs not only allow, but also require information regarding change at an individual level. Repeated tests of self-concept give the amount of change for each participant. However, it requires more effort to gather cost data specific enough to give the exact cost of the changes produced in individual participants. Therefore, for most analyses, the average change is used. However, this depends on the type of decision to be made, as we shall discuss later.

Effectiveness can be stated in three major ways. First, one could measure marginal variables, which compare differences over time for the individual participant, or between prevention approaches. A usual

example would be the change in self-concept score before and after program participation. Typical evaluation designs use these kinds of comparisons.

Another way is the goal referenced comparison, where effectiveness is measured in terms of how close the program comes to achieving its stated objectives. The catch in this approach is that quantification of goal statements is often done intuitively, and only after the evaluation effort can program administrators adequately state expectations for program performance. To satisfy the needs of funding sources, the manager may write an objective which says something like "At the conclusion of program activities, illicit drug use among participants will have decreased by 40 percent." But where did the 40 percent come from? Is it a reasonable expectation based on prior experience, or is it a number concocted to satisfy the needs of others to know what they should expect from the program?

If a goal is reasonable given past experience, then comparing performance to the goal is a good way to assess effectiveness. But if the goal statement is either overstated or understated, then any comparison of actual performance to the goal statement has no meaning. This illustrates yet another aspect of the program manager's quandary in developing statements of objectives. Very often the information needed to state the objective arises only from the evaluation that is supposed to be in part based on the statement.

The final major reference for effectiveness variables is the aggregate level of performance, or the norm. A program could be judged on the strength of its ability to reach the population norms for its objectives. A problem arises, however, when the norm is not a measure of what is desired. If a prevention program is directed at a group of adolescents whose drug use is higher than the norm (as determined, say, by national surveys), then how satisfied should we be to find out that the program has reduced drug use to the level of the general adolescent population—a drug use level that we are all concerned about.

The relationship between cost and effectiveness.—Having discussed how to assess costs and effectiveness, we can now move closer to the issue of worth—the relationship between the two. Cost-effectiveness is the general expression of the relationship between the values of resources consumed and outcomes produced. If cost and effectiveness are expressed in the same terms, usually dollars, then the relationship is referred to as "cost-benefit."

The outcomes of social programs are not simply expressed. The problem is in assigning monetary value to enhancements in the quality or length of life. What dollar value do we place on an improvement in self-concept? How do we express in monetary terms the benefits accruing from preventing one person from becoming a drug abuser? One measure of benefit is earnings—the value of goods that could be produced by those prevented from becoming abusers. But how, then, can we justify prevention activities directed toward the elderly, who have no future earning potential? How can this human benefit be expressed in monetary terms?

One solution to the problem of valuing outcomes that have no market value (or an equivalent) was developed when economists attempted to evaluate the effectiveness of alternative military weapons systems. Given two systems with the same objective, it was not necessary to convert benefits to monetary terms. Instead, the one that achieved the desired objective at the lower cost was chosen. The major weakness of this approach is that comparisons can be made only between programs where the effects can be expressed in the same exact terms, such as increase in self-concept as measured by the same test.

In prevention we can express in monetary terms such outcomes as reduced treatment or incarceration costs, increased earnings, and the like. The same issue of present valuing that was discussed relative to costs applies to benefits. To determine the net value of a program, we must first discount benefits, or convert monetarily expressed outcomes to present value. Having done so, it is simple to subtract the present value of costs from the present value of benefits. The result is the present value of net benefits—a monetary measure of worth. Of course, a negative value indicates that costs exceed benefits.

Another way to express the relationship is by using the ratio of benefits to costs (or vice-versa). The larger the benefit-to-cost ratio, the greater the worth of the program. A ratio of less than one indicates that the present value of costs exceeds that of prevention benefits.

A third expression for measuring worth is the internal rate of return, which is equivalent to the interest the program makes on its investment. This rate is the one that, when applied to the costs, will equalize the present value of costs and benefits. If the internal rate of return for the program is higher than the accepted interest rate for social or private investment, then the program is worthwhile.

Here is an example of the three methods. Say we have estimated the present value of costs for a drug program to be $100,000, with a present value of benefits of $110,000. The difference is $10,000—the

present value of net benefits. This tells something about the program's value, but another program will achieve the same difference with a cost of $20,000 and a benefit of $30,000. The second program has invested fewer resources to obtain the same net benefits, and a simple comparison of the present value of net benefits does not reveal that fact. More information results from calculating the benefit to cost ratios. The first program has a ratio of 1.1 ($110,000/$100,000). The second program has a ratio of 1.5—surely a significant difference.

Finally, calculating the internal rate of return provides even more information. With a 10 percent internal rate of return for the first program and for the second program a more sizeable 50 percent, not only can the two programs be compared to each other, but also each can be compared to the accepted investment interest rate. The results of this third criterion, the internal rate of return, might not be congruent with the other criteria because of differences in the timing of expenditures of resources and the accrual of benefits. Results are also completely dependent on the choice of discount rate and on the time periods over which we discount costs and benefits. As the discount rate increases, the present value of future benefits declines sharply.

When making a choice between program approaches which achieve the same objectives, you need not be concerned with expressing benefits in monetary terms. To compare two approaches for improving self-concept, only accept a common measure and compare program outcomes and costs. In this case the measure might be increases in scores on the Piers-Harris or some other well known scale. Such a measure is accepted in the same spirit as money is accepted as a common measure in cost-benefit analysis.

Of course, there are complications. In cost-benefit analysis, assumptions are made about money that might not apply to scores on a self-concept test. Certainly we would be quick to say that a 10-dollar bill is worth 10 one dollar bills. But is a 10-point increase in self-concept by one person worth the same as 1-point increases by 10 people? Are we willing to accept these two changes in self-concept as equal in value and deserving of the same? No economic market establishes the two values as equal or unequal.

Average and marginal costs.—Costs can be looked at in two ways in cost-effectiveness comparisons, based on the question to be answered. If we can continue to support only one of two existing programs, then the average cost per unit effectiveness is the first choice for a measure. If we wish to increase the capacity of one program or the other, then the first choice is marginal (additional) costs.

Assume that a program's effectiveness is measured by reduction of marijuana users. Without calculating the exact cost per participant, we can obtain an average cost per unit of outcome by dividing total costs by units of outcome. Say that in a given time period the number of users is reduced by 2 percent for a total program cost of $10,000. Then the average cost for each percent reduction is $5,000. Compare this to another, similar program which is able to achieve a 3 percent reduction for $12,000—an average cost of $4,000 for each percent reduction. If forced to choose between programs, we would choose the latter, which achieves the same effect for $1,000 less per unit.

If, instead of choosing between programs, the question involves increasing or reducing allocations to competing programs, then an analysis of marginal costs is called for. Marginal costs are those that are necessary to increase or reduce the effect by one unit. Of two programs, say one involves awareness groups, the major cost being personnel, and the other is a fine arts club, with a major expense in art supplies. Assume that the programs have equal total costs and effectiveness. Unless the first program were filled to capacity and had to hire a new staff person just for the sake of one additional participant, it would probably be more effective to give the additional funds to this program. Increasing allocations to this program would give a better return for an equal added investment.

Cost-effectiveness analysis using average costs requires only aggregated data at the program level. Marginal cost analysis requires some data at the individual level. But these techniques only inform decisions to support effectiveness, not to improve it.

The Resource-Component Model

Everything discussed so far is defined by Yates (1980) as assessment. He considers analysis as the process that develops information after considering cost constraints, process characteristics, and effectiveness criteria. Program decisions are constantly made to shift resources to reduce cost and improve effectiveness. Yates' component-resource model nicely portrays the issues considered by any good administrator. It starts simply, with the path of resources supplying a process that produces an outcome.

Resources.—The resources of a drug prevention program are the facilities, equipment, materials, personnel competencies, and participant dysfunctions and competencies. As Yates (p. 94) notes for mental health, "dysfunctions or at least their potential in the community, are a necessary resource because without dysfunction the existence of mental health services cannot be justified."

Resource constraints limit every resource available to the program. There is an interaction between the available resources in the sense that a change in the limits of any is likely to affect the others. A program in a small facility cannot expand its staff or clientele beyond the limits of the facility. Competence of personnel will affect client entry into the system.

Process.—The process components are the technology available to the program and its delivery system, and there are constraints in both. Staff can always be better trained and better able to apply that training. The constraints on technology are measured by the best outcome possible. If use of a certain technology under ideal conditions prevents "only" 95 percent of all drug abuse, then there is a constraint on that technology. We cannot stop the other five percent from using drugs. The constraints on the delivery system are measured by the difference between the constraints on the technology and the actual outcomes that the program is able to achieve.

Outcomes.—The major outcome in prevention is self-evident. In decisionmaking, the administrator considers other possible outcomes as well, both positive and negative. It may be, for instance, that a small proportion of youth who are taught decisionmaking may use these skills to reinforce values considered deviant by society. The possibility should not be ignored, but rather should be investigated, for certainly knowledge of who might have negative outcomes and under what conditions can be helpful both to avoid the negative outcomes and improve the technology.

Application of the model.—The competent manager considers all aspects of the system for decisionmaking. These considerations may be qualitative, or what many would call subjective, because they are not easily amenable to measurement and have not been externally validated by scientific methods. Careful cost-effectiveness analysis can help validate decisionmaking as well as improve it through new and relevant information. At the level of a single program, analysis of the cost and outcomes of specific components in the context of restraints can provide information to improve program performance by altering activities to:

- o produce a specified level of effectiveness with minimal costs,
- o maximize effectiveness with a specified level of costs, or
- o develop an optimal mix of costs and effects.

In the example program developed in this chapter, the language teacher had a much higher attrition rate than the guidance counselor. If we knew the success rates, as measured by the self-concept and drug use scales of each group leader, we might identify differential outcomes related either to (a) different participant types, or (b) different levels of competency of the group leaders. This could lead to decisions regarding training or participant assignment.

Client routes of entry into the program might be related to differences in outcome. Of the various types of selection (self, other students, the two group leaders, or other teachers) it might be found that some types have better outcomes than others. Further, some activities within the overall program might be more effective relative to their cost than others. As the number of variables to be considered increases, the complexity of decisionmaking increases, and the cost-effectiveness of the analysis itself becomes an issue. The program decisionmaker must decide how much of existing resources should be directed toward evaluation based on the expected return for the investment.

Careful cost-effectiveness analyses based on accurate evaluations of outcomes can justify the continued operation of a good prevention program. But remember that all such analyses are based on the assumption of scarce resources. If resources were unlimited, costs would not have to be justified. In theory, at least, unlimited resources imply unlimited technologies. All problems could be solved. But in the real world, many resources are getting scarce, and the need becomes greater to justify the use of resources by improving the social welfare. It is at this point that the goals of the action researcher and the program decisionmaker fully merge.

NOTES

[1]This Merlin-like approach is ascribed to Reichardt (1981) who purports to have taken it from Rubin (1974).

[2]There is a result in the drug prevention literature which is chillingly like the present example. It has been suggested that drug information programs, while perhaps decreasing use in frequent drug users, may well lead abstainers or infrequent users to use drugs. If you imagine a drug prevention program intervention between the two test administrations in Figure 3, you will realize how regression artifacts may confound our interpretation of evaluations. A randomly assigned control group for the high and low users would have clarified the meaning of the data, as in Figure 1a, in which there was <u>less</u> gain in drug use in program participants than in randomly assigned controls.

[3]This sort of consistency is more clearly grasped in terms of achievement and ability tests. On a test of mathematical ability, there should not be some items which are easier if your math ability is low. We've probably all had the experience of bad multiple-choice items in which the more you know, the more difficult the question becomes because more than one alternative can be plausible.

CHAPTER 5: PREPARING FOR THE EVALUATION

(They Say It's the Light Stuff . . . But)

It's **your** program that's being evaluated.

A program decisionmaker should be involved in every stage of the program evaluation—planning, implementation, and utilization—just as in any other major program activity. Each stage requires a particular set of skills, a particular orientation, and a particular involvement. The manager's role in each stage is described in this chapter, with emphasis on what to expect, what to do, and what pitfalls to avoid.

The role of the evaluator will also be discussed as it parallels and intersects that of the program manager. Additionally, the critical roles of staff, boards, and concerned community members or service recipients will be emphasized.

First, the requisites for planning will be discussed, covering such issues as:

- o selecting the evaluator
- o manager/evaluator relationships
- o preparation of self, staff, and community
- o contracting with the evaluator.

The second section will consist of a detailed discussion of the evaluation process (French and Kaufman 1981):

step 1—analysis of decisionmaking activities
step 2—analysis of program activity
step 3—development of alternative evaluation designs — **planning**
step 4—initial selection of a design
step 5—operationalization of the design
step 6—field test of the plan
step 7—revisions resulting from field test — **implementation**
step 8—collection and analysis of data
step 9—utilization of results. — **utilization**

The chapter will emphasize how success at each step depends on satisfactory resolution of previous steps. Discussion of the ninth step, in particular, will demonstrate the dependence of utilization on all that has gone before and will also discuss the impact of the politics of an evaluation (internal and external to the program) on the utilization of the evaluation.

REQUISITES FOR PLANNING

The following is a true story. A State agency informed a local program director that his program was scheduled for evaluation during the year. The director was pleased, saying that there were many questions he would like answered. The State evaluator told the director that the agency wanted its questions answered, not his--questions pertaining to the success of the local system in adhering to certain statewide standards.

The director said he was aware of no such standards. The evaluator replied that his team had only recently written them, and they were still in draft form. The director objected to being held accountable for draft standards. Not to worry, he was assured; they would become official standards once the legal office's review was complete and would then be implemented statewide.

When the director objected to being held accountable for draft standards of questionable legality, the evaluator reminded him that the state contributed two-thirds of his funding. With that, the director relented and asked for a copy of the standards. The evaluator then told him that, unfortunately, the State agency director had prohibited distribution pending legal clearance.

Thus, the local director found his program being evaluated by his funding source, on its terms, with its evaluator, according to unofficial, legally questionable, and secret standards.

The local program staff resented the evaluation, finding the evaluation team obtrusive and incompetent. Hostile letters were exchanged. The evaluation report, after a delay of several months, was distributed simultaneously to the local director, several funding agency staff, community representatives, and elected officials. The report, which the director was denied permission to review before distribution, contained several factual errors and many interpretations subject to dispute.

This anecdote is a textbook case of how not to do an evaluation. The pitfalls evident in the example can easily be avoided by adhering to the planning requisites and the nine-step process presented below.

Selecting the Evaluator

The motives for an evaluation will have major impact on what is evaluated, ultimate use of the results, the manager and program's participation, and selection of an evaluator.

Generally, evaluators may come from three sources--the funding agency, the program itself, or an organization independent from both. The funding source may not only insist upon an evaluation, but may also provide an evaluator. The program may have an in-house evaluator or it may hire an external evaluator. Selection of the evaluator may be the prerogative of the program manager, funding agency, board of directors, or the like, depending on the impetus for the evaluation and who is paying for it.

An important issue is to whom the evaluator is responsible, since the evaluator will give primary allegiance to that person. Allegiance is a major concern because all steps of the program evaluation will be influenced by the relationship between evaluator and employer. Everything will be affected including what is done, what is inferred, what is said (and not said), and who hears it.

In general, most program managers will prefer not having an evaluator selected for (or forced on) them, and will prefer to have the evaluator accountable to them. A manager who recruits, selects, and pays the evaluator will be in a stronger position to monitor the aims, process, interpretation, dissemination, and use of the evaluation. No matter what direct authority the manager has over the evaluator, several factors should be considered to assess the evaluator's appropriateness.

Technical competence.—By education and experience, does the evaluator have knowledge and competence to do the job? Can the candidate establish evaluation goals? Develop sound designs? Select suitable measurement techniques? Analyze and interpret data? Write a coherent sentence that is also appropriate to the audience?

Versatility.—An evaluator with a repertoire of techniques will be more likely to meet the program's needs. As Patton (1978, p. 31) says, "The burden rests with the evaluator to understand what kind of evaluation is appropriate for different types of programs rather than forcing all programs into a single evaluation model."

The obligation of the evaluator is to evaluate a program as it is, unless the manager agrees to program changes. An evaluator must have the flexibility to conduct credible evaluations without modifying the program ahead of time simply to meet evaluation needs. Put differently, effective evaluation is partly an art, and there's no reason to believe that Rembrandt painted by the numbers.

Cultural sensitivity.—If a program serves a community with a significant number of language or ethnic minority members, the evaluation will need to address issues relevant to those groups. Different goals and different assessment techniques may be needed. In addition to technical knowledge of measurement issues

involving those groups, the evaluator must be familiar with the values of the groups involved, exhibit respect for such cultural diversity as it exists, and be acceptable to the community.

Manager/Evaluator Relationships

The relationship between managers and evaluators is often strained. Weiss (1977) suggests four sources of conflict.

Personality differences.—The manager and the evaluator are usually different types of people whose very differences drew them to separate fields where their differences were reinforced by experience. Evaluators see themselves as scientists contributing to the knowledge base of society; program managers see themselves as helpers who contribute through service provision. The former are data oriented, the latter are people oriented. Such differences provide potential for conflict.

Role differences.—Evaluation implies judgment; the evaluator carries the aura of the judge, the manager the judged. This role hierarchy is heightened when the evaluator is the agent of the funding source or some other outside, controlling group, and complicated when he is hired by the manager.

Lack of boundary clarity.—An evaluator's role can be as limited as the analysis of existing data, or as broad as helping a program identify its goals, conducting a full-scale outcome evaluation, and then helping the manager make changes indicated by evaluation results. Because the evaluator's role boundaries are often left undefined, tensions are probable.

Resentments over differential rewards.—Evaluators may receive more pay than program staff and may be perceived as less hard working—"We do the work; evaluators read charts." Even the appearance of the evaluator's name on a final report can be a source of friction.

That program managers and evaluators have differing and occasionally incompatible world views is nowhere better illustrated than in an article by Weiss (1977), who conducted a survey of participants in 10 evaluations of human service programs. Two major differences in perspectives were found, one in the way participants view evaluations, another in the way the parties view each other.

Both evaluators and managers expressed general uncertainty about the purposes of the evaluations they participated in—whether the studies were to serve the program, its funders, or knowledge in the field. Managers saw evaluations--practically if not ideally—as serving three functions:

- a ritual to secure funding
- an opportunity to vindicate the program
- a guide to change and improvement.

In contrast, evaluators had somewhat more idealized views about evaluation functions:

- assessment of program effectiveness to enable decisions to be made
- an opportunity to contribute to basic knowledge.

Further, managers generally preferred evaluations focusing on process and development (to guide future program development) whereas evaluators preferred those emphasizing outcome and effectiveness to facilitate judgment of programs. When evaluations conformed more to the wishes and beliefs of evaluators, managers tended to lose interest in the evaluations and to withdraw support.

Weiss (1977, pp. 33-34) also suggests a fundamental mistrust of motive and viewpoint between managers and evaluators. Evaluators are credited with fighting "for the integrity of their data" in the face of attempts by managers to impose positive interpretations on equivocal findings. Managers are alleged to grant autonomy to evaluators "less from respect for the integrity of research than from unsophistication about possible effects of evaluation." Then, as sophistication increases, "there may be more interference with the planning and conduct of evaluation research." Evaluators see managers as hampering evaluation, "often out of ignorance."

In another article, Weiss (1975, p. 15) writes that managers "are not irrational; they have a different model of rationality in mind. They are concerned not just with today's progress in achieving program goals, but with building long-term support for the program. Accomplishing the goals for which the program was set up is not unimportant, but it is not the only, the largest, or usually the most immediate of the concerns on the administrator's docket."

As these quotes suggest, a common view aligns

evaluators with knowledge and integrity,
managers with ignorance and resistance,

suggesting not a little condescension.

Trying to compare managers and evaluators along a good-bad or positive-negative dimension is inappropriate. Far more productive is viewing each party as possessing integrity, ability, and devotion to certain kinds of truth. Both are dedicated to doing the best possible job, but they have different jobs with different success criteria. Evaluators believe in and fight for the integrity of their data; managers equally believe in the integrity of their programs. As a respondent in the Weiss (1977, p. 34) survey said, "Practitioners have to believe in what they're doing, evaluators have to doubt."

Career-oriented managers and evaluators also share a need to be successful in their work, but success is differently defined, and for neither is career success dependent primarily on the effectiveness of programs. Evaluators develop careers by conducting methodologically competent evaluations useful in guiding social or program policy and contributing to general knowledge through publication. Whether the programs evaluated are successful is not their primary concern. For program managers in human service programs, success is usually defined in terms of longevity, growth, size of staff and budget, and number of people served. Because many human service programs are never adequately evaluated, and because evaluation reports are filed and forgotten more often than not, the actual effectiveness of a program may have little impact on a manager's career and reputation.

Attending to some of these differences and similarities should help managers and evaluators see themselves not as antagonists, but as complementary and even synergistic partners in the enterprise of program evaluation.

Preparation of Self, Staff, and Community

Preparation for an evaluation requires focusing on both technical and context issues. The former involves analyzing the stage of program development, assessing information needs, and determining readiness for evaluation, issues which will be developed later as part of the nine-step process. The context refers to the psychological and political readiness of the program--attitudes, beliefs, and interrelationships of managers, staff, service recipients, and advisory or governing boards.

Typically, evaluations are perceived by staff as threatening. At the least, evaluations will cause some disruption—there will be interviews, record reviews, and more forms to complete. At the worst, evaluations cast doubt on program effectiveness and staff competence, threatening the esteem and job security of program staff. The lives of staff are inevitably affected by an evaluation, to degrees ranging from mild disruption to distinct threat.

Service recipients, too, may be directly affected by an evaluation process. They may find themselves being interviewed by strangers, having questionnaires or psychological tests thrust upon them, and signing release forms. Further, any disquiet felt by the staff may be passed along to recipients of service.

Finally, parent organizations, such as local health departments, community mental health centers, or boards of directors, may also be interested in the evaluation and should be involved in the preparation process.

To create the best possible context for an evaluation, two actions should be taken by the manager. First, analyze the relative importance of the motives for the evaluation and its potential effect on the program. It is easy to focus too much on an imposed evaluation or on the temporary disruption of the program, but the real significance of an evaluation lies with its potential impact. An evaluation report based on a month of frenzied activity may be filed unread at the State agency; alternatively, an unobtrusive analysis of file data--conducted with little or no immediate effect on staff or clients—could have a major effect on the program's future.

As a general rule, the greater the evaluation's potential effect on the program--positive or negative—the more important it is for the manager to involve staff, consumers, and superordinate organizations in the evaluation process. Effective involvement of these parties, although no panacea, will improve the evaluation process, create a broader sense of ownership, and make program changes easier to put into effect.

Others might be also involved, but there's no simple guideline for determining who. It will depend not only on the evaluation circumstances, but also on a program's organizational size and structure, relationships with consumer groups, and place (if any) within a larger organization. In addition to the director and the evaluator, three groups should be considered for involvement in the evaluation planning process: staff, recipients, and advisory or governing boards.

For any program, key staff must be involved, however that term is defined. It is usually helpful to include at least one person who has a clinical or provider (as opposed to administrative) orientation.

Consumer or recipient involvement may prove more difficult to obtain. If you have an active consumer group in your community, the program is a step ahead. A citizens' advisory group may provide an appropriate representative. The population at which the program is aimed has a legitimate investment not only in the program but in its evaluation.

Significant cultural or linguistic diversity within the target population will complicate consumer input, but make it more critical. Differences may exist between cultural groups as to program goals and criteria for success. For instance, a program aimed at adolescents may have as stated goals reduction of alcohol use, increase in participation in school activities, and enhanced self concept. While one cultural group may want no alcohol use by children under 16, another group may tolerate alcohol use at home, and a third may be more concerned with alcohol-related arrests than with drinking per se. One group may be more interested in their adolescents having after-school jobs than in whether they write for the school paper or play in the band. And, certainly, the definition of self-esteem varies among cultures and economic classes. Accordingly, evaluation goals must reflect the diversity within the target community.

Measurement issues are also affected in pluralistic communities. While the controversy over the applicability of standardized intellectual measures to minorities is well publicized, measures of personality and attitude should also be culturally relevant. The number of culturally tested measures is small, and managers may legitimately expect evaluators to be aware of those that do exist. As a rule of thumb, translations from English into, say, Spanish or Vietnamese, will not yield measures of comparable meaning or validity. Review by representatives of the cultures concerned can help ensure not only adequate goals and measures, but also acceptance of results.

Finally, depending on organizational circumstances, the evaluation should involve advisory or governing boards and concerned managers of the larger organizations within which the program may be placed. Before the evaluation officially begins, three basic questions should be answered: What is being evaluated, how, and what will be done with the results? Involvement of key staff, consumers, and concerned community or governing agencies in answering these questions is fundamental to prepare for an evaluation.

Contracting with the Evaluator

The contract with the evaluator need not be a binding legal document, but should express a clear understanding (preferably written or part of a legal contract) of the responsibilities of the evaluator and the program, and the boundaries between them. Seven critical and potentially troublesome issues must be resolved prior to formal implementation of the evaluation.

Division of labor.—Who will collect the data, who will distribute forms, who will conduct interviews, and who will provide necessary training? The answer to any of these questions could be the evaluator, the program staff, students, or volunteers, etc. The worst answer is no answer; these are questions to be considered in advance.

Division of resources.—A related issue has to do with access to resources. Who provides typing, photocopying, envelopes and stamps, computer time, paper, and the like?

Timetable.—Specifying well in advance when steps in the process are to occur, or to be completed, will help all parties budget their time. Particular attention should be paid to time of delivery of the final product. Few things can dilute the usefulness of an evaluation more than results delivered too long after data were gathered. Program people lose interest, funding cycles may be missed, or circumstances may have changed. It will be more helpful to have a finished evaluation 2 months before rather than 2 weeks after a budget is due.

Deliverables.—What you expect from the evaluator should be stated at the onset. Make it clear if you want a preliminary report. What kind of final report do you want? How many copies? Will you want some public presentation or presentation to the staff?

Distribution of results.—You'd probably rather learn of the results directly from the evaluator than from the local newspaper. The final report belongs to the individual or group that provided the impetus for the evaluation and paid for it. Generally, program managers will want to receive and control access to the report to whatever extent possible.

Right of preview.—Related to the issue of control of the report's distribution is control of its content. Without invoking debate about the integrity of data, the issue here involves interpretation and emphasis. Managers will usually wish to see a preliminary or draft report and have the opportunity to recommend changes, make corrections, and discuss interpretation. The self-protective stance behind this wish is obvious enough; at the same time, an evaluator hoping to make a contribution to a program beyond the simple analysis of data will recognize the risk of Pyrrhic victories inherent in surprise attacks.

Authority to renegotiate.—Chances are that things won't go exactly according to plan. Staff won't cooperate, clients won't show up, computers will malfunction, evaluators will decide to get married, or mail will get lost. Changes in agreements will be made, and the original negotiation should make specific who has the authority to approve or to insist upon changes.

THE EVALUATION PROCESS

Planning the Evaluation

Each of the following five planning steps is a prerequisite to conducting an evaluation. The activities comprising some of these planning steps may be familiar to program managers, and most will have highly developed skills in these areas. Nevertheless, even familiar activities are worth describing in some detail, especially highlighting the ways they fit into the overall evaluation process.

Step 1—Analysis of decisionmaking activities.—An evaluation is useful to the manager because it produces information for decisionmaking. The evaluator will suggest methods for gathering valid information, but the program manager is responsible for ensuring that information gathering is guided by the correct questions--questions whose answers may be used to improve program efficiency, decrease program costs, increase program effectiveness, or plan for the program's future. These questions will provide the overall conceptual framework of the evaluation, and their content, scope, and focus will influence each step in the evaluation planning process. As Patton (1978) has noted, evaluation reports placed on the manager's bookshelf and never used are almost invariably based on questions not relevant to the manager's decisionmaking activities. From this perspective, it is difficult to spend too much time in the analysis of program decisionmaking and the development of evaluation questions.

To develop questions that provide a useful framework for the evaluation, the manager must consider both short-term and long-term decisions and the information needed to make them. Put another way, the manager and other relevant decisionmakers (funders, staff) should develop a list of statements which follow the form:

WE NEED TO KNOW _____ BECAUSE WE NEED TO DECIDE _____.

For example, the manager of a program emphasizing community planning groups might make the statement:

> WE NEED TO KNOW which alternatives programs are most appealing to area youth BECAUSE WE NEED TO DECIDE directions the planning groups should take.

Similarly, the manager in a multiprogram agency may make the statement:

> WE NEED TO KNOW which of our programs are most cost effective BECAUSE WE NEED TO DECIDE where to plan expansion.

The development of the we-need-to-know-because-we-need-to-decide list (which is, in fact, the first draft set of evaluation questions) involves three separate activities:

- o analysis of the stage of program development
- o assessment of information needs and development of evaluation questions
- o assessment of the program's readiness for evaluation and change.

Analysis of the stage of program development.—Program development is a dynamic process which can be roughly divided as follows:

- o needs assessment
- o policy development
- o program design
- o program initiation
- o program operation.

It is incorrect to view program development as a linear process, with each phase completed before the next is begun. Rather, a program may be in different phases simultaneously, and all program elements may not develop at similar rates or at the same time. The manager will ask different questions depending on the stage of development of the program (or of its various elements). Accordingly, the first step in an analysis of decisionmaking activity is to determine the stage of program development of those program elements the evaluation will address.

A major task for the manager in analyzing program development stages is to divide elements of the program into those relatively stable, and those that are evolving. All too often evaluations address outcome-type questions (is this program element changing drug use?) about program elements that are not stable in either concept or implementation. An evolving program element is much more likely to fail the test of outcome evaluation, and a potentially potent program element may thus be unnecessarily eliminated from further consideration. Because the evaluator will generally view the program at only one cross section in time, he will have difficulty assessing the relative stability of various program elements. The manager, with in-depth knowledge of the program's history, is in the best position to determine which program elements are stable and which are not.

Tharp and Gallimore (1979) describe the conditions necessary for a social program to reach stability. Their discussion suggests three criteria of stability. The first is longevity. The history of prevention programing reveals numerous false starts and blind alleys. As a rule of thumb, a program element requires at least 6 months to a year before it can begin to stabilize, and some program strategies (community organization and social policy change) may require several years before stability is reached.

The second criterion is stability of values and goals. Prevention programs and program elements seek to remediate specific drug and alcohol abuse problems or their precursors. Accordingly, program elements will be stable only to the extent that they address stable problems in ways consistent with stable community values. The manager's needs assessment data and feel for the climate of values in the community will prove particularly useful in applying the criterion of goal and value stability.

The third criterion is stability of funding. When different program elements are funded by different sources or on different funding cycles (often the case for prevention programs), a review by the manager of funding stability will be most useful in developing questions and focus for the evaluation.

Once the manager has considered the relative stability of the program or program elements, it will be important to examine the stage of development of staff responsible for program implementation. Because of high staff turnover rates in many prevention programs, a well-established program element (such as a drug curriculum module) is often implemented by a new or relatively new staff member. When this is the case, the manager may wish to postpone outcome-oriented evaluation until the staff member has had time to fully learn the new role. Sometimes the competency with which staff implement various program elements is itself a focus of the evaluation. Even when this is the case, a review of which staff members are doing what tasks will help the manager develop questions for the evaluation.

A final major issue for the manager to consider in analyzing the stage of program development is the extent to which various program elements have linkages to, and support from, the community. In their Design for Youth Development Policy, Bird et al. (1978, p. 142) note that a given program "...acts simultaneously as a subsystem charged with handling one or more of the problems on a broader scale for the community and the societal system of which it is part." Prevention professionals recognize this issue, and program managers have actively sought to use their community linkages to improve the quality and impact of their program elements. However, the development of such sharing of resources may take considerable time and effort, especially in larger communities where numerous agencies compete for the resources needed for prevention. To the extent that the program is a credible member of a community network, the manager can expect more stability in, and effect from, a given program element. Moreover, if a program is particularly well linked to other community agencies, the potential for studying community-wide impact is enhanced.

Having completed an assessment of program linkages, the manager will have a good feeling for the stage of development of the various program elements. The analysis of the stage of program development will prove particularly useful in choosing the appropriate level of evaluation (process, outcome, or impact) and in choosing among various methodologies (qualitative and quantitative). This analysis provides a background against which the manager may begin to consider information needs and to develop evaluation questions.

Assessment of information needs and development of evaluation questions.—Starting from analysis of the program's stage of development and using the guidelines set forth in chapter 2, the manager may now begin to assess the program's unique information needs, guided by the short and long term decisions faced. This will ensure that the evaluation addresses issues relevant to the manager's role as a decisionmaker. However, the manager is not the only decisionmaker needing information from the evaluation. Funders, staff, community members, and even program participants and their families have valid needs for program information. The wise manager identifies individuals who face decisions or need questions answered about the program.

Patton (1978, p. 284) suggests that people whose information needs should be considered include people:

- o who can **use** information
- o to whom information makes a **difference**
- o with questions they **want** to have answered
- o who **care** about and are willing to share responsibility for the evaluation and its utilization.

As Patton notes, this list boils down to those who come to mind when thoughtfully considering Marvin Alkin's (1975) question:

"Evaluation—Who needs it? Who cares?"

Once the manager has developed a list of relevant decisionmakers and information users, a set of evaluation questions should be solicited from them. This may not be an easy task, especially if program staff or participants, for example, are not used to having input into the evaluation planning process. One useful technique for soliciting evaluation questions is to ask these individuals to develop a list of we-need-to-know-because-we-need-to-decide statements like the ones described earlier.

Such statements can be obtained in a number of ways, ranging from formal focus groups to informal meetings and telephone calls or mailed questionnaires. The method will depend in part on the personal style of the manager and in part on situational constraints. For example, individuals may be geographically scattered or simply too busy to attend a formal session. The manager may also wish to alter the we-need-to-know-because-we-need-to-decide format. Patton's (1978) original example used an I-would-like-to-know _____ about-this-program format, and the manager will surely think of other useful formats as well. The particular format is not nearly as important as its ability to elicit important evaluation questions.

Usually, the information users and decisionmakers (including the manager) will identify a number of similar issues of program effectiveness, efficiency, and cost. As a side benefit, the manager often gains new insights into the concerns of staff, board, funders, participants, or community. For many managers, these insights alone are worth the effort to gather these statements. The program manager should combine the suggested evaluation questions into a single, unduplicated list. If these individuals are brought together in a formal meeting, a number of techniques exist for developing a group consensus, for example, the Nominal Groups Techniques (Delbecq et al. 1975). However, consensus concerning the list of evaluation questions is not necessary or even always desirable. The finished product forms a first draft of the evaluation questions for which the evaluator will later devise methods and measures to answer.

Assessment of the program's readiness for evaluation and change.—Once a first draft of evaluation questions has been developed, the manager's analysis of decisionmaking activities is almost complete. However, before proceeding to the practical issues involved in analyzing program activities (the next major step in evaluation planning), the manager should pause to consider the climate for evaluation and change within the organization, and especially among program staff.

It will not surprise anyone that a large literature (Delbecq 1974; Lippitt et al. 1958; Hage and Aiken 1970) suggests that individuals and organizations resist change. As the program managers are well aware (Kiresuk et al. 1981, p. 221),

"one of the most pervasive barriers to change is
a generic fear of change in general,
a desire to maintain the status quo."

By its very nature, evaluation portends change and becomes a threat to the status quo. But there are other reasons why program staff and others within the organization may resist evaluation. As most managers know, from a purely practical perspective, the evaluation means more work. Prevention programs are often understaffed and underfunded. It is the rare program that has staff with time reserved for evaluation activities. The evaluation may be viewed as an added burden with no apparent benefit to those taking on the additional work.

Staff may also feel that the evaluator's tools are incapable of measuring what staff are really doing. This concern may be general, such as program activities cannot be adequately portrayed through scientific inquiry. Or, it may be quite specific, e.g., the appropriateness of a given set of measures for the program's participants. Staff who have had bad past experiences with evaluators will have little inclination to repeat the experience. Finally, staff may feel that they, rather than the program, are being evaluated.

Overall, the manager may be faced with a staff who would just as soon forget the whole idea of evaluation, and who may even attempt to undermine one that is forced on them. Within such a climate, an evaluation effort will be at best difficult and at the worst a waste of everyone's time and effort. Fortunately, the manager can use two strategies to encourage acceptance of, and even enthusiasm for, the evaluation.

The first, already suggested, is involving staff in the development of the evaluation questions. This strategy helps build ownership of the evaluation and provides tangible benefits from cooperating: the staff's information needs will be addressed, they will be working for their own benefit. Moreover, involving program staff in the development of questions and other decisions gives the evaluation a level of credibility well above those evaluations seen as belonging to someone else and addressing someone else's concerns.

The second strategy to decrease resistance is to show staff ways in which evaluation can facilitate, rather than impede, their daily activities. Evaluations, especially those related to process, can provide program staff with much needed monitoring information and short-term feedback. For example, one staff member of an alternatives program confessed that he was often at a loss to remember important specifics of planning meetings with program participants. A semi-structured log for these meetings both met the staff member's immediate need and formed an important part of the program's process evaluation. As part of the design of a process or outcome evaluation, the evaluator can also help staff to redesign, streamline, routinize, and even computerize recordkeeping to decrease the amount of time these activities take. Once staff become aware of the ways in which evaluation can aid them in improving the day-to-day operation of the program, they can become avid supporters of the evaluation.

With the completion of step 1 (analysis of decisionmaking activities), the program manager will have developed the conceptual framework for the evaluation, including a fair idea of the questions to be addressed. There will be some notion of the appropriate levels of evaluation for each program element and the beginning of an organizational climate to foster implementation of the evaluation.

Step 2—Analysis of program activity—Before beginning to design the actual evaluation with the assistance of an evaluator, the manager must examine certain aspects of the program to determine their adequacy for the requirements of the evaluation. Specifically, the manager will need to:

assess the adequacy of program objectives,
review and catalog current data collection methods, and
review staff and other resources for evaluation.

Depending upon level of skill and experience with evaluation, the manager may wish to enlist the help of an evaluator in completing some or all of these activities.

Assess the adequacy of program objectives.—In almost all cases, the manager and others will want the evaluation to examine program effectiveness. From the evaluator's perspective, this question is always asked in terms of the program's outcome objectives. While most program managers have extensive experience in writing objectives that are useful for planning and management, a significant number seem to have difficulty writing objectives useful for evaluation.

Cantor et al. (1981) propose four useful steps that program managers can use to develop evaluable outcome objectives. The first step calls for listing program goals. Program objectives are often developed that are only tangentially related to program goals. Specifying goals will help in developing the objectives. Well-stated goals are outcome oriented. They specify the condition(s) the program hopes to address and the target population the program is expected to affect. Because goals are so broad in scope (e.g., reduction of

marijuana use among middle-school students in Lake City), most prevention programs will have only one or two goals.

The second step requires the development of indicators of goal attainment. Cantor et al. (1981, p. 4) define indicators as "specific, observable changes in attitudes, knowledge, or behavior which are linked either by theory or logic to goal attainment." Examples of goal attainment indicators for reduction in marijuana smoking might include improved ability to resist peer pressure, increased knowledge of alternative highs, or increased ability to cope with stress. Program staff and even program participants (or potential participants) may be involved in brainstorming indicators of goal attainment.

The third step is the selection of the three or four best indicators of goal attainment. Cantor's four criteria to select indicators include the significance and relevance of the indicator for the program's target population, the importance of the indicator to program decisionmakers, the ease with which the indicator can be measured, and the ability of the program to have an impact on the indicator.

The final step in Cantor's process is the translation of indicators into measurable objectives. Measurable objectives include a statement of the indicator, the identification of a target population, a time frame, and tne amount of change expected. Thus, measurable objectives take the form,

> "By April 8, 1982, students at Grant Middle School will report a 20 percent increase in their participation in alternatives activities," or
>
> "By January 11, 1982, 70 percent of the seventh graders will report an increased ability to cope without drugs."

Note that these objectives are stated as program outcomes or performance, not as program effort. There is a temptation to write program objectives which relate to activities rather than outcomes. For example, "teacher training will be given in five schools during the spring semester." Such process objectives are useful for program management, but they are of limited value for evaluating program effectiveness.

Review and catalog current data collection methods.—Prevention programs vary widely in the amount and quality of the records they keep. In some cases, all the data collection necessary for the evaluation will already be in place. In general, however, new data collection methods will need to be developed. In any event, the evaluator will wish to know exactly what records are currently kept, and he will want an assessment of the quality of these records.

Basically, four categories of data are regularly required for prevention program evaluation: participant, staff, program activity, and program cost. Not all these categories will be required for any given prevention evaluation. The manager can begin to get a good idea of which data will be required by referring to the analysis of decisionmaking from step 1. Working from the draft list of evaluation questions, a Data Needs Checklist can be developed. For example, if one evaluation question refers to community reaction to the program, the Data Needs Checklist will indicate a need for some kind of community attitude survey. Even the skilled evaluator sometimes finds that not all the necessary data has been gathered to answer the complete list of evaluation questions.

With the Data Needs Checklist in hand, a manager may begin to consider the data and records currently available. Client intake and exit interviews, school records, needs assessments, client records, and telephone logs are obvious sources. However, the manager may find that staff and even clients are keeping records such as logs and diaries that may be useful for the evaluation. Even if many of these records need reformating for the purposes of the evaluation, data collection currently going on will facilitate the integration of the evaluation into the day-to-day operation of the program.

The evaluator will want to know about the quality of these data. Simply speaking, the quality of records depends on three characteristics: regularity, consistency or reliability, and validity.

Regularity refers to the extent that the records are kept up-to-date. While busy staff may sometimes neglect paperwork without many negative programmatic consequences, missing data can be a disaster for the evaluation. Accordingly, quality records are kept religiously.

Consistency or reliability refers to the extent to which the same event is recorded in the same way time after time. If, for example, classroom acting-out is recorded, each similar instance of acting-out should be recorded in the same way. This requires good definitions of the events to be recorded, and it requires that all recordkeepers work from the same set of definitions. Even such simple definitions as what

constitutes a program session may vary widely from individual to individual. Consistency of definitions cannot be assumed.

Finally, validity refers to the extent that the descriptions in the records accurately reflect what actually happens in the world. For any number of good reasons, responsible individuals put things into records that simply are not true. Often people do not die from the causes listed on their death certificates or are not charged with the crimes they actually commit; participant drug use may be overreported or underreported. The manager must be concerned that those records used for evaluation purposes are valid reflections of the program.

The manager will more than likely discover that other data collection devices will be needed for the evaluation. Although the evaluator will be able to suggest a number of instruments, observational checklists, and so on, the program manager may also wish to begin searching for additional data collection devices. Readily available sources of instrument descriptions include:

- The appendix to the Handbook of Prevention Evaluation (French and Kaufman 1981)
- The Prevention Evaluation Research Monographs, Outcome Volume (Aiken 1981)
- The Drug Abuse Instrument Handbook (NIDA 1977).

Review staff and other resources for evaluation.—The availability of persons with various skills (and with free time) will probably be the single greatest constraint on the extensiveness of the evaluation. A discussion of available resources with the evaluator will be an important first step in developing evaluation design options.

Basically, all evaluations require individuals to collect, code, and analyze data. All these individuals (with the possible exception of data analysts) can probably be found within the ranks of program staff. A brief description of the tasks that must be performed follows and will allow the manager to begin considering which staff might do what.

Data collectors fall into three basic categories: interviewers, questionnaire administrators, and trained observers. Of these, questionnaire administrators require the least training, while interviewers and observers will generally need a formal introduction to their roles. In no case, however, is academic preparation directly relevant. It is more important that these individuals be comfortable around and enjoy people. Usually interviewers and observers can be trained in a 1-day session, although a complex interview or observational protocol may require a somewhat longer session. Questionnaire administrators may also require a small amount of training to insure consistency of instruction giving and interpretation of items, but this training should rarely take more than a few hours. In general, the qualities found in most prevention program staff (concern for and interest in others, some clinical insight, good communication skills) will make them excellent data collectors once properly trained.

Data coders are responsible for data storage and for the coding of questionnaires, interviews, and observational protocols. Their task may be as simple as transferring numbered responses to code sheets or as difficult as deciding whether an interview response fits into one or another category. In general, the work of the data coder is not difficult and almost everyone can help out in this role. Data coders must, however, be able to do detailed work accurately. The quality of data coding will have a direct impact on the overall quality of the evaluation.

Data analysts take the raw data and prepare summary statistics, charts, tables, and graphs. Depending on the evaluation design, they may also perform statistical tests of evaluation hypotheses that range from relatively simple to highly complex. Ordinarily, graduate training in the social sciences or statistics is necessary for any but the most rudimentary statistical analysis. How-to books on the statistical analysis of data do exist (Fitz-Gibbon and Morris' How to Calculate Statistics is one good example), but these are of limited use. Unless the manager or staff have training in data analysis, other resources for this aspect of the evaluation should be sought.

Besides person power, the manager will need to find some resource for computing. Unless the evaluation is completely qualitative (which is rare), or only a small quantity of data is collected, even the simplest data analyses become overwhelming without the aid of a computer. Some agencies will have access to computers through a school system or local government, and a lucky few may even have their own computing resources. However, the manager will often have to look elsewhere for a computer.

Happily, most prevention programs are close enough to a college or university to share in the wealth of knowledge and resources these institutions offer. Most universities offer computing facilities equipped with packages of programs for statistical analysis. Moreover, many professors are more than happy to have

"real" data for students to analyze. A call to the chair of psychology, sociology, health education, industrial engineering, social work, or statistics can sometimes lead to an arrangement for analyzing data. But be sure your data needs get met—not just theirs.

The university as a resource is by no means limited to data analysis. University students can also serve as interviewers, interviewer trainers, data coders, observers, and data analysts, sometimes free of charge. Most social work programs and many social science programs encourage or require their students to gain field experience. An offer by the manager of an opportunity for such experience may be welcomed by the dean or other faculty, but persuasion and negotiation will be necessary.

Step 3—Development of alternative evaluation designs.—The manager is now well prepared to develop evaluation design options. Here the services of a skilled evaluator will probably be necessary. Before arriving on site, the evaluator will want to review as much material concerning the program as possible. The analysis of decisionmaking activities and of program activities will have generated a number of documents: draft evaluation questions, revised program outcome objectives, a Data Needs Checklist, and copies of current data collection devices. Copies of these documents along with relevant funding proposals, brochures, program work plans, and the like should be forwarded to the evaluator well in advance of the consultation visit.

The development of evaluation design options involves two activities:

deciding on the scope of the evaluation
and
developing the design options themselves.

In general, the evaluator will take the lead role in both of these activities. However, the manager will have to remain an active participant to provide the evaluator with the information and data needed, as well as to make necessary decisions.

Deciding on the scope of the evaluation.—The scope of the evaluation will be expressed in terms of the amount of data collected and the elaborateness of the evaluation design. From the program manager's perspective, scope will translate roughly into the number of evaluation questions that can be addressed and the certainty of the answers produced. There is a tradeoff between the number of questions and the certainty of the answers. The manager will need to consider the uses of the evaluation information to balance these two factors.

The evaluator will take several factors into account in helping the manager determine the scope of the evaluation. These factors include the program's readiness for evaluation, its current data collection methods, and its resources for evaluation. After reviewing the program's materials, the evaluator will be able to give a rough assessment, such as, "We should be able to do a thorough job on the process questions, but we'll be somewhat limited in our ability to measure effectiveness for all program components." Taking off from this rough assessment, the evaluator will then specify exactly which evaluation questions on the draft list are to be included, and which postponed or dropped.

Almost invariably, the draft list of evaluation questions developed by the manager will exceed the scope possible for the agency. Accordingly, the manager and the evaluator need to prune the list. As Patton (1978, p. 137) notes, the usual solution to this problem is to rank the goals of the evaluation in terms of their importance. Patton further notes, however, that priorities set in terms of importance may not result in the most efficient use of limited evaluation resources (emphasis in original):

> The fact that a goal is ranked first in importance does not necessarily mean that decisionmakers and information users need information about attainment of that goal more than they need information about a less important goal. In a utilization-focused approach to evaluation, program goals are also prioritized by applying the criterion of usefulness of evaluative information. . . The ranking of goals by the importance criterion is often quite different from the ranking of goals by the usefulness of evaluative information criterion.

A key reason that importance and usefulness yield different priorities is that the most important prevention program outcomes are often the most distant and difficult to measure. So, for example, the most important outcome of a smoking prevention program may be a decrease in the prevalence of chronic disease. However, this outcome may be impossible to measure. Measuring a less important, intermediate outcome (e.g., being able to refuse a cigarette in a socially acceptable manner) may be more useful to evaluate and improve the program.

Another reason is that the manager may be able to obtain high-quality information without using expensive evaluation resources (Patton 1978). For example, a sophisticated sociological study of classroom climate is unnecessary if the manager can get all the needed information by visiting classrooms and speaking with teachers. This is not to suggest, of course, that such a study may not be necessary under other circumstances for other programs.

Working together, the evaluator and the manager will refine the draft list of evaluation questions to bring the most useful areas of evaluative inquiry into focus. Several different lists may be developed and measured against the scope of the evaluation that the evaluator deems feasible. In the ideal case, the information users and decisionmakers who helped develop the draft list will be involved to some degree in this process as well. Minimally, however, the final list of evaluation questions should be reviewed by these individuals before the actual implementation of the evaluation.

Development of design options.—When it is time to develop evaluation design options, the evaluator may wish to work offsite, closer to resources such as a personal library and colleagues. While the manager may view this as a loss of control over the evaluation planning process, it can reasonably be assumed that input to this point and the refined list of questions will guide the evaluator in appropriate directions. In any event, the manager will have an opportunity to review the evaluator's design recommendations and assess their adequacy in meeting information needs.

Chapter 4 has described in detail the issues the evaluator faces in designing an evaluation. Here let us briefly review these issues in the context of developing evaluation design options. Basically, the evaluator will proceed by resolving three issues for each of the evaluation questions on the refined list.

Type of information.--The first, and in many ways most basic, issue is the type of information each evaluation question requires—description, comparison, or explanation (cause and effect). Each of these areas requires different evaluation strategies.

Descriptive questions ask such things as who, what, where, when, and how, and are most often associated with process evaluation. An example of a descriptive question is, "How many boys versus girls attended the alternatives fair?" While descriptive questions can and should be answered with great rigor, they do not require elaborate research designs or sophisticated statistical analyses.

Comparative questions ask about the relations among variables without assigning causality. Such questions often concern the relationships between characteristics of the participants (age, sex, risk status) or characteristics of staff (expertise, training, enthusiasm) and program outcomes. An example of a comparative question is, "Is rock climbing a more effective prevention alternative for boys than for girls?" The evaluator may choose to incorporate such questions as formal features of an outcome evaluation design, or may choose to study them more naturalistically, capitalizing on naturally occurring variations in the factors of interest.

Explanatory questions concern the extent to which the program is causing changes in the attitudes, knowledge, and/or behavior of the program participants and others. Questions of this type are almost always addressed by evaluations designed to rule out alternative explanations for the changes observed. As explained in chapter 4, a number of design options exist which vary in the ability to rule out alternatives, thus supporting the claim that the program is responsible for observed outcomes. Often there is a tradeoff between the extent that a given design option can rule out alternative explanations, and the cost and difficulty of that option.

Type of measures.—For any given evaluation question and for any of the three information types (descriptive, comparative, and explanatory), the evaluator can choose from a wide variety of measurement techniques. These include observation, various types of interviews (structured and unstructured), questionnaires, psychological tests and measures, and reviews of archival records.

In making initial choices from among these options, the evaluator will be guided first by the specific question to be answered. But considerable weight must be given to the appropriateness of the measure for the specific target population, the expertise necessary to use the measure, and the cost of the measure. Wherever feasible, the evaluator will wish to gather data concerning a given question in more than one way. Overall, the evaluator will attempt to maximize the quality of the data while minimizing cost and disruption of the program's day-to-day activities.

Who will be measured.—It is almost a truism that the larger the sample obtained in the evaluation, the more accurate the results will be. However, the law of diminishing returns (see, for example, Hays and Winkler 1971) applies especially when resources for evaluation are limited. In many ways, the creative use

of various sampling techniques is the evaluator's most powerful tool for maximizing the resources available. The evaluator may also need to overcome such obstacles as school-imposed restrictions on who can be measured, and issues of informed consent.

The tradeoff in this case is between the numbers of individuals who can be measured and the scope, flexibility, and sensitivity of measurement. For example, a mailed questionnaire can reach large numbers of individuals, but an exploration of nuances in meaning is lost. Alternatively, small numbers of individuals may be measured in great depth and with great elaboration, but the cost of such an option may preclude measuring a sample large enough to be representative.

Overall, the evaluator will develop various combinations of measures, samples, and evaluation strategies. Now the manager and the evaluator face the difficult task of choosing among these various design options.

Step 4—Initial selection of a design.—In choosing among various design options, the manager will perhaps confront the major tradeoff in the entire evaluation planning process: striking a balance between the usefulness of the entire evaluation and the amount of dollars, staff, and other resources that can be committed to it. Unfortunately, resources spent on evaluation are often resources taken away from the services being evaluated.

Happily, much of the evaluative information that is most useful is also the least expensive to gather. Often, the refined list of evaluation questions will be somewhat weighted toward process evaluation, and the manager may wish to choose a design option emphasizing the process level.

Of course, all prevention program managers must concern themselves with outcomes, but the kinds of data derived from a sophisticated randomized experiment may well be unnecessary for decisionmaking. In some cases, qualitative outcome data may be sufficient, and in many cases, a relatively unsophisticated outcome design will be all that the manager requires.

In any event, the manager should quiz the evaluator extensively about the strengths and weaknesses of various design options, and the strength of a given option should be measured against the importance of the decisions to be made based on the data. Certainly the manager will not want to base major decisions on weak data, but neither should precious resources be expended on a rigorous study relating to a relatively trivial decision. The prioritization of evaluation questions can be used to guide the differential allocation of resources in choosing among design options.

One final consideration in choosing among design options is the ease with which important constituencies such as funders and legislators can understand the design. Designs vary in their intuitive appeal and the simplicity of their logic. Instead of a tempting flashy new technique with an air of scientism and high technology, choose the simplest design possible that will meet information needs. When the time comes to disseminate the evaluation findings, the flashy design with its complex logic and statistical analysis may be a deterrent to clear communication. All else being equal, the easier an evaluation design is to describe and understand, the greater an asset it will be.

Step 5—Operationalization of the design.—To this point, the manager and the evaluator will have been dealing essentially in abstractions. However, an evaluation becomes a specific set of activities, performed by a group of individuals, according to a detailed workplan. In operationalizing the design, pragmatic considerations are primary. The myriad practical constraints associated with implementation of the evaluation must now be considered. The evaluation design may have to be altered to fit the operating context, but generally this task is one of working out the details.

Program staff are particularly important actors in this phase of evaluation planning. They are the ones most likely to know whether this or that evaluation activity can be comfortably incorporated into the program's operation. They may also be the best resources in terms of the ability of the program participants to respond to various measurement devices. For example, an evaluator may plan to use a particular measure of drug knowledge that the program person can see is above the reading level of the program participants. Because program staff will be partly responsible for various aspects of implementing the evaluation, their involvement in the design will help build ownership and enthusiasm.

Two of the most important tasks at this step of the evaluation are:

- o selection and development of evaluation instruments, and
- o development of detailed timelines and workplans.

In general, the evaluator will take the lead in operationalizing the evaluation plan. However, the involvement of the program manager and staff in this phase of evaluation planning is crucial. Unless the evaluator is very familiar with the program and the community (and most will not be), the evaluation plan may lack sensitivity to prevailing community values and may require activities difficult or impossible in light of the program's day-to-day operation.

Selection and development of evaluation instruments.—Almost all evaluations require some measurement instruments. Reports of behavior, behavioral intentions, knowledge, attitudes, and psychological variables are all regularly assessed in prevention evaluations. In some rare instances, the selection of instruments will be a happy task of wading through several dozen choices (as is the case for self-esteem measures for white, middle-class youth). Often, however, few if any published instruments exist that are appropriate for the target population.

Though difficult, the process of instrument development need not present insurmountable problems. As noted earlier, several compendiums of instrument items for prevention evaluation currently exist and most evaluators have had some experience in the development of instruments. The use of newly developed or revised instruments will, of course, require additional time for pretesting and revision (see step 6 below). Suffice it to say, this time will be well repaid in the quality of the evaluation data.

Ultimately, the manager, program staff, and even program participants are in the best position to judge the appropriateness of a given instrument for their community. If the instruments suggested by the evaluator seem inappropriate, the manager must consider revising them or developing entirely new measurement techniques. Failing to do so risks the quality of the entire evaluation effort; doing so increases costs.

Development of detailed timelines and workplans.—Often the role of managing the evaluation will fall to the program manager or a staff member. Logically then, the manager or designee should take primary responsibility for mapping out an evaluation workplan. Ideally, the evaluation will be managed using the same techniques as other agency business. If formal techniques are employed for program management, such as Management by Objectives or Gantt charts, these should also be employed to develop the evaluation workplan. In general, however, the key issue is to determine in advance the various evaluation tasks, the necessary person power, the work assignments, and some method for ensuring the timely completion of the evaluation. In developing a workplan for the evaluation, be sure to allow enough time for each evaluation task. To paraphrase an old saying,

> the first three-quarters of the evaluation will take three-quarters of the time.
> The remaining quarter will take the other three-quarters.

The manager unfamiliar with evaluation activities may tend to underestimate the time that tasks require. An evaluator can provide useful guidance here, but a conservative timeline, that allots too much time for various evaluation tasks, will never be regretted.

A second major issue in developing the evaluation workplan is to ensure that major activities, such as testing of participants, occur at times that are convenient, feasible, and consistent with the design. All too often evaluation plans schedule pretests during summer vacation, posttests during the manager's vacation, and data analysis while the computer is tied up with other business. Here, as elsewhere, the active participation of program staff in development of the evaluation workplan can avoid problems and greatly facilitate implementation.

Implementing the Evaluation

The implementation stage of evaluation incorporates the next three steps in the evaluation process:

step 6--field test of the plan
step 7--revisions resulting from field test
step 8--collection and analysis of data.

Step 6—Field test of the plan.—At this point, the purposes of the evaluation have been established, program goals articulated, the evaluation design developed, and measurement instruments selected. Temptation (and fiscal or temporal pressures) may lead to immediate implementation of the plan. However, it is desirable, and for large-scale or complicated evaluations essential, to field test the evaluation components before plunging full speed into the process.

A field test is a practice evaluation. A small sample of service recipients will be involved in trying out the questionnaires and interviews. Data will be analyzed, and presentation formats examined. The purpose is to determine whether the plan works. The Handbook for Prevention Evaluation (French and Kaufman 1981, p. 19) says this about field testing:

> All aspects of the evaluation plan should be pilot tested, including sampling, measures, data collection plans and analytic procedures, and utilization activities. The pilot test determine (sic) whether the data collection schedule is feasible, if the collection can be carried out with minimal disruption to program activities, if the data being collected are valid, whether the variables are reliably measured, if the costs of data collection and analysis are on target, and whether the resulting information is used as intended by the decisionmaker.

This comprehensive order can be broken into three basic components: testing the design, testing the process, and testing usability of the data. The design may call for providing certain services to some people and something different to others. Certain types of data will be collected. The field test shows if the design works. Can the procedures be applied as planned? Will respondents be available and cooperative? Is the data analyzable if collected in that manner?

Pretesting the planned process may prove that questionnaires are too lengthy or ambiguous, psychological measures invalid, or anticipated file data too sketchy. More extensive training of interviewers may be required. Pockets of resistance among the staff may surface, and everything may take longer than anticipated.

Finally, a field test should help clarify whether evaluation data will be useful. Will the types of results answer the questions the manager wants answered? If not, the evaluation will not serve its full purpose.

The manager may reasonably expect that the evaluator will be expert in determining how extensive a field test is needed and designing an appropriate one. The role of the manager in the field test includes:

- o assessing the value of field testing
- o participating in planning a useful test
- o conveying to the staff and relevant others the need for a field test
- o ensuring resources and cooperation necessary to complete the test
- o helping review test results with an eye toward those aspects of the evaluation over which the program manager has control
- o working to effect any changes needed in the evaluation design.

The manager's most difficult role may be enlisting the cooperation of the staff, who may consider the evaluation itself sufficient nuisance without needing practice first. The manager's attitude and appropriate involvement of staff in previous phases of the evaluation will be the best levers in obtaining staff cooperation.

Step 7—Revisions resulting from field test.—The intent of the field test is to perfect the evaluation plan, eliminating such bugs as may be found. For example, service recipients in one program were asked by staff to submit voluntarily to interviews. As a result, the volunteer rate was quite low. Staff resistance proved to be the problem, and efforts were increased to bring staff into the evaluation process. Another evaluation required correlation of pretreatment demographic variables with posttreatment behavior. Field testing revealed deficiencies in pretreatment data gathering, which were corrected.

In a third case, field test results included an unexpected negative correlation between treatment conditions and posttreatment attitudes of Hispanic clients. The problem was found to lie in the translation and interpretation of the Spanish-language questionnaire.

These examples indicate the types of problems which can be spotted through field testing and that require the active involvement of the program manager. Each example involved a condition the manager would like to avoid, such as antagonizing clients; a problem that could reasonably be handled, such as poor records and staff resistance; or a problem that lessened the value or usability of results.

Other problems of evaluation design, technical aspects of data analysis, or problems in instrumentation are legitimately within the domain of the evaluator.

Step 8—Collection and analysis of data.—This stage has three substages: implementation, analysis, and interpretation.

Implementation.—At this point the evaluation is in progress. The bugs have been worked out, and the procedures smoothed. The manager's role now is to monitor the process, to ensure that the evaluation is being conducted as planned, and that program's services continue to be delivered without significant alteration or disruption. Clearly, not only those evaluation activities under direct program control, such as interviewing clients or differential client treatment, but all evaluation activities should be monitored.

Analysis.--This is a fairly mechanical stage in which the gathered data are analyzed. The analysis may be as elementary as frequency counts or as sophisticated as multivariate statistics, and the responsibility for conducting the analysis will be the evaluator's. Remember, though, the type of analysis and the format in which results are ultimately presented should have been decided upon much earlier in the process, tried out during the field test, and should have the manager's concurrence.

Interpretation.—Each of the nine steps being discussed is dependent on the success of the preceding steps. However, this substage has a high degree of independence. Even the most clearly phrased question may yield murky answers. The clearest of answers may contain not a clue as to explanation. The presentation or wording of results can affect how results are interpreted.

In one instance, a school-based decision-skills program for preadolescents was found to have no measured impact on later drug use. This failure may have been due to improper program implementation by the staff, poorly trained or inexperienced personnel, or application of the program to the wrong population. Or perhaps it was just a bad idea. Which of these possibilities should be discussed and/or emphasized in the report? How should the results be presented? Who gets to make the decision? These questions will be of definite consequence to the manager.

Further, suppose the program was shown to have led to a 15.5 percent reduction in later drug use. Consider the different interpretations that would attend the following statements:

The program yielded only a 15.5 percent reduction.
The program yielded a 15.5 percent reduction.
The program yielded a reduction of over 15 percent.

Or, perhaps the program was shown to lessen drug use, but program recipients rated the program negatively. Consider the difference in emphasis between these statements:

Although program recipients tended not to rate the program favorably, they did show a significantly lower rate of subsequent drug use.

Although a significant reduction in subsequent drug use was demonstrated, program recipients rated the program negatively.

The consequences of interpretation will generally be felt in one of two ways: decisions internal and decisions external to the program. In the first case, decisions to change or not change programs will be based on interpretations of results with emphasis given to some results more than others. Interpretation and emphasis may stem entirely from the evaluator, be left to the manager, or jointly derived. The manager's goal is to make or receive as accurate as possible an interpretation to make the best possible decisions.

It may be that the locus of decision lies outside the program, perhaps with the funding agency. Funding sources, of course, deserve accurate interpretations. Program managers will be legitimately concerned not only with accuracy but with the political and economic context within which decisions will be made. When the context places the program in a vulnerable status, managers will prefer some statements to others. "Only 15.5 percent" and "15.5 percent" are equally accurate information but differ in connotation and may lead to different decisions. The argument here is not for skillful deception but for decisionmaker involvement in the form of data presentation and in the interpretation of results.

Step 9—Utilization of results.—Sometimes evaluations have to be done pro forma; the fact that they are done is sufficient, with no requirement, expectation, or hope of their use. Ideally, however, evaluations will be used, and from the outset conducted with ultimate use in mind. Chapter 10 of the Handbook for Prevention Evaluation contains a discussion of factors important to the uses of evaluation. The core of its message is to

build utilization into your design from the beginning.

Davis and Salasin (1975) cite a collection of articles on critical evaluations of Federal programs. In each case, the evaluation was forced on the recipient agency by a superordinate agency and was designed to meet the latter's needs. And in each case, the managers of the evaluated programs spent their energies

criticizing instead of using the evaluation. "Utilization," Davis and Salasin (p. 623) note, "may be more apparent than real when mandated by authority. . . without collaborative involvement of the people representing the program being evaluated."

Patton (1978, p. 63) makes the point that "People, not organizations, use evaluation information," and reemphasizes that the intended users of an evaluation should help plan it. Patton's survey of Federal decisionmakers indicated that two characteristics influenced the use of evaluations: political and personal.

Political considerations are essentially external to the program, involving social issues, budget cuts or growth, or large-scale social program success or failure. These issues are discussed further in chapter 7. For now, it is useful to recall that a program is often the result of a political process and its evaluation may be part of the same or a new political movement (Weiss 1975). Although evaluation is a scientific process in search of truth, it does not always avoid fighting and is often also a method of fighting within the political arena (Lindblom 1968).

Thus, a community concerned about drug use may value the existence of a program more than a scientific demonstration of its success. Elected officials who helped initiate the program thus might pore through an evaluation looking for words of praise and ignore pages of criticism. Or, in times of decreasing public budgets and general disenchantment with human service programs, an evaluation finding only moderate success may be read as a condemnation of the program for not being perfect. However, an unevaluated program may be able to prove nothing about itself except its existence, and thus is vulnerable to any attack weighed against it.

Whatever the political climate, a program manager has to work within it and may have little or no impact on it. Thus, the second of Patton's two critical factors, personal, will usually be a more appropriate focus for the manager. By personal, Patton (1978, p. 64) means "the presence of an identifiable individual or group of people who personally cared about the evaluation and the information it generated." When this factor is present, the evaluation is more likely to be used. Consider this statement, made by an evaluator surveyed by Patton (1978, p. 66):

> Where there were aggressive program people, they used evaluations whether they understood them or not--used it as leverage to change . . . his program.

Another (p. 67) said an evaluation was used "because the decisionmaker was the guy who requested the evaluation and used the results. It was the fact that the guy who was asking the questions was the guy who was going to make use of the answers." Use of the evaluation will emphatically depend on this personal factor, most often that of the manager, whose involvement from day one in all steps will set the stage for ultimate use. As Weiss (1975, p. 19) said, an evaluation "is most likely to affect decisions when it accepts the values, assumptions, and objectives of the decisionmaker."

While the primacy of political and personal interest is acknowledged, other factors do contribute to the usability of evaluation. Glaser and Taylor (1969) compared unsuccessful with successful evaluations and found the following contributed to success:

- o from the beginning, high involvement of relevant groups inside/outside the organization
- o study designed by a full-time principal investigator
- o commitment of the host agency
- o evaluation aimed at a felt need of the organization
- o involvement of potential consumers of results
- o readily disseminated findings.

Patton (1978) reviewed the literature and listed other factors contributing to evaluation use:

- o methodological quality
- o methodological appropriateness
- o timeliness of evaluation
- o timeliness of the final report
- o whether findings were positive or negative
- o "surprisingness" of findings--were results expected?
- o whether central or peripheral program goals were evaluated
- o existence of related findings elsewhere
- o resources available to implement changes
- o evaluator-manager interactions.

Weiss and Weiss (1981) surveyed social scientists and decisionmakers to determine their views on what impeded and promoted effective utilization. They found appreciable agreement between evaluators and decisionmakers. Some major impediments over which managers have a high degree of control were tendencies for:

- o decisionmakers to ignore information contrary to their own ideas
- o policies to be arrived at by politics, not research
- o agencies to ignore findings contrary to their policies
- o decisionmakers to have difficulty defining research needs
- o lack of communication between decisionmakers and evaluators.

There were also factors that both groups agreed contributed to evaluation usefulness:

- o topic of study is of particular interest or relevance
- o study looks at variables that decisionmakers can do something about
- o report is understandable, not overly technical.

Decisionmakers placed more emphasis than did evaluators on timeliness of the reports and on the interest of the user in the population studied. Evaluators were more likely to be concerned with studies of great social concern and with dissemination of information. The number of factors is partly arbitrary and semantic. What is important is the relative value of each in a given situation. Note that none of these factors arises at the end of the evaluation. Each may be anticipated from the outset, and failure to anticipate them virtually guarantees failure of the evaluation.

However, the converse is not necessarily true. Anticipating the future does not guarantee that the future will arrive as anticipated. Davis and Salasin (1975) advise on tactics for effectively presenting evaluation results, their meanings, and changes which may result from them. They cite several important considerations in presenting results and recommendations.

- o The presenter is able to identify with the audience.
- o Essential information is repeated and restated often.
- o A combination of logical and emotional appeals is made, without exaggerating the latter.
- o The benefits and risks of change are made clear.
- o Recommendations are consistent with the values of recipients of the presentation.
- o Objections are anticipated and dealt with.
- o Free expression of resistance is encouraged.

Management of change is a topic outside the scope of this volume. However, the principles of involving key personnel from the outset and of intelligent preparation of results and recommendations will lay an effective groundwork for making needed change.

A final issue concerning use of evaluations is how to deal with negative results. There are many potential reasons for negative results: improper concept, improper implementation, improper evaluation, or external factors beyond the program's control. Some evaluation designs may help identify the causes of failure, others may not. Occasionally, failure is built into the program. For example, to secure funding, planners may promise more than may be deliverable or promise to deliver results more rapidly than is possible. In such cases, the evaluation will find that goals have not been completely met. Independent of such contrived dilemmas, however, newer programs often fail to meet even rational expectations. The recommended rule of thumb for such cases is this:

Programs must be allowed to fail.

The appropriate response to negative results from evaluations of new programs is often not radical program change, wholesale firings, or funding cuts. Rather, unfrenzied program introspection, heightened attention to implementation procedures, and renewed coordination with the community may enable programs to overcome failure. Programs not allowed to fail are not allowed to grow, change, or adapt; to take risks and be creative; or to meet intended needs.

In sum, utilization is the raison d'etre of evaluations. Planning for utilization should be an integral part of planning all components of the evaluation, from the initial stages of identifying questions to the end stage of presenting the answers.

CHAPTER 6: CASE STUDIES IN PREVENTION EVALUATION

(What Really Goes On . . . . Inside
a Triple Feature)

AN OVERVIEW

The three hypothetical case studies in this chapter are intended to emphasize the realities of the evaluation process as experienced by prevention program managers, staff, and evaluators.

The case studies present prevention programs at different stages of development and reflect various prevention modalities. Each case study emphasizes different steps of the evaluation process described in previous chapters and has its own unique motives and primary audiences for the results of the evaluation. In these case studies the interactions between the program managers and the evaluators are the most significant aspect of the narratives.

Although these case studies present a slice of evaluation life, the reader should understand that a much broader range of designs, measures, analytic strategies, and issues occur in an actual evaluation. However, the material presented does capture the essence of the evaluative experience. The stories are entitled: Double Trouble, Four Thrilling Discussions, and One Suspenseful Melodrama. The dialog at times is lighthearted; however, the message in each case study is essential to the theme of this volume—good evaluations occur when program managers and evaluators work cooperatively on an evaluation.

DOUBLE TROUBLE

Alternative Designs for Alternatives Programs

The Brightside Youth Center, located in a major midwestern city, was established 7 years ago to provide prevention and intervention services to troubled youth. It is housed in a community center and currently delivers services in two broad areas: drug and alcohol prevention services in the public schools, and a program of social and recreational activities for youths from 6 to 18 years of age. The Brightside staff consists of 12 people, most of whom are counselors and social workers. Their funding comes from a mixture of State and local drug and alcohol prevention grants and United Way support, supplemented by small amounts of private donations.

Donna Campbell is the director of the Brightside Youth Center, a position she has held for the past 3 years. Two other staff members, Joanne Martinez and Jim Cook, are assistant directors in charge of the drug and alcohol prevention component and the social-recreational activities component, respectively.

During the past several months, Donna, Joanne, and Jim have discussed their needs for evaluation of the Brightside programs. Although none has a background in evaluation (in fact, they have always been pretty resistant to the whole notion), they recognize that their funding agencies are increasingly asking for evaluation information of a fairly sophisticated nature. Moreover, Donna and her staff have recently begun to believe that perhaps some evaluation might help to identify more effectively the strengths and weaknesses of the Brightside programs. So a few weeks ago, Donna called the National Prevention Evaluation Resource Network (NPERN) to ask for some technical assistance to help them develop an evaluation strategy. NPERN responded to her request by arranging for a consultant skilled in program

evaluation to work cooperatively with the program. The consultant, Ron Fisher, is a research psychologist who specializes in the evaluation of drug and alcohol prevention programs. In preparation for his 2-day visit, Ron and Donna talked briefly on the telephone about the purposes and functions of the consultation visit.

During Ron and Donna's initial meeting in her office, they discussed basic matters relating to the Center's organization and history (objectives, staffing patterns, and the like). She also shared her motives for the evaluation with Ron, at which point he expressed pleasant surprise.

"You mean you're not under heavy outside pressure? That's as rare as someone going to an alcohol counselor on their own initiative."

Joanne joined the meeting as they began analyzing the functions and activities of the drug and alcohol prevention program. Joanne described the program's major activity as the provision of broad prevention services to two large high schools and three junior high schools on the south side of the city. (The south side population is 24 percent Hispanic, 28 percent black, and 48 percent white, mostly second and third generation Polish and Italian.) The Brightside staff conducts semester-long classes at these schools called Positive Directions for Youth, which include sessions on interpersonal communications, stress management, self-concept, family dynamics, and drug and alcohol use. Teacher-facilitators assist the Brightside staff in the conduct of the classes. Approximately 20 percent of the student population is assigned to the classes; plans call for a gradual expansion of coverage to include the entire student body eventually.

As we look in on the meeting, Ron is about to discuss potential evaluation designs with Donna and Joanne.

"I think now I've got a pretty good idea of how your drug and alcohol prevention program runs, its goals, general strategies, and so forth. So I think we're ready to start talking about some possible evaluation designs you might want to implement. How's that sound?" Donna and Joanne look at each other, then at Ron, nodding affirmatively.

"Before we go on," Ron continues, "I hope you had the chance to read NPERN's Working With Evaluators. Not only can it save time in defining terms and the evaluation process, but one of the case studies in that monograph bears a striking resemblance to your program and, in fact, with our discussion so far." Everybody nods vigorously.

"OK, very good," Ron goes on. "Now, as you might know, there are two basic kinds of evaluation—process and outcome. With process evaluation our first interest is an accurate documentation of what kinds of services and activities your program actually engages in—the exercises you use in the class sessions, what the kids actually do, etc., and second, who receives the program services—the types of kids who are in the program. With good documentation you can go on to more sophisticated process analysis. On the other hand, outcome evaluation is used to——"

"Hold it please, Ron," Donna says, smiling, but with an upraised hand as though stopping traffic. "This is all pretty new to us, so let's take it one step at a time. How is 'process evaluation' useful to us?"

"I'm sorry," Ron grins sheepishly. "Please feel free to stop me and ask questions whenever you're not sure of something. Well, process evaluation can help you in a couple of ways. It can be a management tool to help you keep track of what is actually happening in your program and what your client population looks like at any point in time. This kind of information can also be used for annual reports, reports to funders, in grant applications, and so forth, to show external funders and agencies what you are doing—and that you have solid information about what you're doing. It's pretty basic stuff we're talking about here, the kind of documentation that, to some degree, every program should have. And, of course, that lays the groundwork for cost-efficiency and other more complex analyses."

"I see," Donna nods. "And outcome evaluation?"

"Outcome evaluation is designed basically to assess the extent to which your program is achieving its major goals. In your case, Joanne, outcome evaluation would attempt to determine how well your program actually prevents the use and abuse of drugs and alcohol among the kids in the program."

"But we address more basic issues of adolescent adjustment in our program, not just drug and alcohol use," Joanne asks, "Shouldn't we assess program effects on such dynamics as self-esteem, communications skills, and so forth?"

"Most definitely," Ron replies. "Outcome evaluation should address those objectives that are usually considered intermediate objectives or correlates of drug and alcohol abuse, including attitudes toward drug and alcohol abuse. However, it's important to keep in mind that for a drug and alcohol abuse prevention program, the focus of outcome evaluation should remain on drug and alcohol use."

"I understand that," says Joanne, "but I also know it's difficult for a prevention program to show evidence of effect on drug and alcohol use in a rather brief time period. I don't want to pin the entire assessment of our program's effectiveness on behavior that even we feel won't show effects for some time."

"I agree completely, so we'll probably build several levels of measures into our outcome evaluation. But we're getting a little ahead of ourselves. Let's first talk about the general design, and then we can get into the specific aspects of the outcome criteria. Shall we talk about the process evaluation first?"

"No, I'd prefer to talk about the outcome evaluation design possibilities first," Donna suggests, "if that's OK, Ron—that's the one that scares me!"

"That's fine. Now, as I understand it, the students who attend the Positive Directions for Youth (PDY) classes are a cross section of kids selected from a larger pool. So you are taking only a fraction of those students who are 'eligible,' right?"

"Yes, that's right," Joanne agrees.

"Can we identify a pool of eligible kids approximately twice the size of the pool that you will select for the classes?" Ron asks.

"You mean at each school?"

"Yes."

"I don't see why not," says Joanne.

"In that case, we might have an opportunity for a true experiment—which is a very powerful outcome evaluation design," Ron points out.

"Sounds pretty ambitious . . . an 'experiment,' " Donna interjects. "How does that work?"

"Well, let's say that at a given school we identify maybe 100 kids who are eligible for the program. We then randomly assign them to either the PDY classes or to a control group—whatever class or condition they would otherwise be assigned to."

"What's the advantage of random assignment?" Donna looks a bit skeptical.

"Well, it's just the best way to insure that we come as close as possible to having equivalent groups to compare, that the kids in the control group will be as much like those in the PDY classes as possible, in terms of background, motivation, and so forth."

"And . . . " Donna prompts.

"And so when we compare them on outcome measures—their attitudes toward drug use, communications skills, etc.—whatever differences we find can be attributed to the program. People can't say, well, the reason for the differences is that the PDY group was smarter, or better motivated, or whatever."

"Do outcome evaluations always use random assignment?" Joanne asks.

"No, not at all," Ron explains. "In some instances, program staff may provide services to virtually all eligible clients, leaving no clients to assign to a control group. Or the program staff may have strong feelings about 'denying' services to anyone—although that kind of stance occurs less often with prevention programs than with intervention or treatment programs, since prevention services typically are not aimed at particular individuals who are clearly in need of some immediate assistance."

"I see," says Joanne. "But what would we do if we could not randomly assign students to PDY or a control group?"

"Then we would probably try to identify a group—a class in this instance—that is as similar as possible to the PDY group and use it as a comparison group."

"And collect outcome information on them at the same time as the PDY group?" Joanne asks.

"Yes, that's right," Ron replies. "Another option would be to collect the outcome information on both groups at several points before, during, and after the PDY services are delivered. That's called a 'time series design,' by the way."

"But these strategies aren't as good as the random assignment approach?" asks Donna.

"No, they aren't, but they're definitely better than no evaluation at all!"

"What kind of outcome measures should we use?" Donna queries.

"Well, the particular outcome measures we use will depend on several considerations, including the objectives of your program, the characteristics of your clients, and how much time and resources you have to devote to outcome data collection."

"All that, huh?" Joanne smiles, looking over at Donna.

"I'm afraid so!" Ron answers. "Aside from the selection of the design, there's no more critical step in the development of your evaluation than choosing your outcome measures. Remember, they're the yardsticks by which your program's impact will be measured. You want to make sure that they really reflect what you think your program will achieve. And of course we want to be sure that they are valid and reliable—accurate measures of outcome."

"Shall we start by looking at our program's objectives?" asks Joanne.

"Yes. Fortunately, you folks have done a fine job of developing realistic, measurable objectives." Ron pulls out the list of PDY objectives from the materials Donna had sent to him, developed as a result of her prior conversations with NPERN. "It seems to me that they reflect six general types of outcomes: substance use, including alcohol, drugs, and tobacco; attitudes toward substance use; self-concept; stress management; interpersonal skills; and family dynamics. Is that accurate?"

"Pretty much so," nods Joanne. "But the interpersonal area should also include things like communication skills and reactions to peer pressure."

"I see. Well, some fairly good instruments are available for the measurement of these outcomes, although measuring stress management skills may present problems. These instruments are designed for use with client populations of the same age and grade level that PDY serves. However, we're sure to encounter some reading problems, don't you think?"

"Yes, we will," Donna answers. "Perhaps 15 percent of the students at the junior high schools will have very low reading skills. Somewhat fewer at the high schools. How do we handle that?"

"Usually we administer the instruments verbally. It would help a lot if these students were previously identified. Can we do that?"

"Probably," says Joanne. "Let me check on that with school staff."

"What about other outcomes like grades, disciplinary records, and so forth?" asks Donna. "We already tried to go through school records for our kids, but the way they keep their files, it's practically impossible to hunt down data for individual students in our PDY program."

"That's a shame," Ron says. "The more important question is whether there's reason to believe that the program will influence those indices, but that becomes academic since you can't get the data anyway."

"OK, now what about consent from the parents for the data we'll be collecting?" Donna continues.

"Well, both the parents and students will sign a form that describes the reasons for the data collection and the type of topics covered in the instruments--what we call 'informed consent.' And of course you'll need to get agreement from the school authorities to conduct the study."

"So, we're basically talking about a set of paper-and-pencil instruments—attitude scales, checklists, that sort of thing—as our measures for the outcome evaluation?" asks Donna.

"That's right."

"Well, I have a couple of concerns about that approach." Donna looks troubled. "First, how can we be sure that those instruments will really measure the kind of impact we think our program has on the kids?"

"There are no guarantees," Ron admits. "The best way to help insure that we're accurately measuring program impact is to use instruments that have a good track record--that is, psychometric data on their reliability and validity—and for us to examine carefully the items on the instruments to satisfy ourselves that they tap the kinds of attitudes and behavior that the PDY program is designed to affect. One of the things I can do for you is explain why some items that don't appear to directly address the issues might be useful. Those items, in our jargon, don't have 'face validity.' Some of us call this the 'interocular test'—if the reason for its being there doesn't hit you right between the eyes, it doesn't have face validity. But there are lots of good measures that don't."

"I see." Donna nods. "My other concern is that we might be relying too heavily on paper-and-pencil types of measures. Shouldn't we do some observing or interviewing—or something other than just the instruments?"

"Yes, we could," Ron agrees. "In fact, it is best to use more than one method to measure anything. Observations, for instance, may be the best way of looking at the whole dynamic of your program without limiting yourself to the preconceived notions that tests require. But that depends on your resources; interviews and observations are very consuming of staff time, as you've already found with the school records."

"Well, let's at least consider those possibilities after we see what kind of resources the whole evaluation process will require—OK?" asks Donna.

"Of course."

"OK, Ron, what are we going to do with all these 'data' after they're collected?" Joanne wants to know.

"Well, with the kind of data we'll be collecting and the design we're using, the only real limitations on the analysis will be the amount of resources you can devote to it—particularly the availability of computer facilities. And I should be able to assist you at that point."

"We've used the computer facilities at the local university in the past, but only for some very routine tabulation activities," puts in Donna. "Maybe we could arrange something there."

"Check into that in some detail, Donna. All these data won't be much good if we can't analyze them."

"Could you give us an example of what kind of statistical analysis might be used?" she asks.

"We'll probably use Analysis of Covariance on most of the outcome data."

"Explain that, will you Ron—in simple terms, OK?"

"Sure. Basically this analysis will compare the scores of PDY kids on the outcome measures at the end of the PDY sessions with those scores of the kids who do not participate in the PDY program—statistically adjusting the scores for any differences that exist between the groups on the pretests."

"So, we're essentially comparing the amount of change in the two groups, rather than the absolute level of their scores, right?" asks Donna.

"Yes, basically that's correct."

"Will we be able to measure the combined effects of the program across all the outcome measures—sort of the overall effects?" Joanne queries.

"Yes, we can, but that will require the use of Multivariate Analysis of Covariance. There are tradeoffs here. On one side, it will cost more in computer time and require substantially more analytic

effort by a well-qualified statistician, and interpretation by us. On the other side, the additional information that could be developed may tell you more about the interplay of the different components of the program."

(Donna, Joanne, and Ron then discuss how the outcome evaluation will be implemented, including specific roles and responsibilities. Ron emphasizes the need for a pilot test of the instrument package on a small but representative sample of students. They discuss in great detail the resources required to prepare for, collect, analyze, and interpret the outcome data. Donna is especially concerned about this, since she was "burned" in her previous experience with an evaluator who drew up an elaborate design and dropped it in their laps. Only later did she realize that they did not have anything near the resources needed to carry out this grand evaluation.

Their final decision is not to include Multivariate Analysis of Covariance at this time, given resource constraints. They then move into a discussion of the process evaluation. As we rejoin the group, they are summing up the plans for the process evaluation.)

"OK," Donna says. "Let me make sure we understand what this 'process' evaluation is about—and why we're doing it!" she laughs.

"Fair enough. Go to it!"

"We'll have observers in the PDY classes recording the session events on a form that you'll help us develop. These observations will produce narrative descriptions of session events. This narrative could serve as a foundation for the future development of a formal, quantitative rating scale of both student and teacher behaviors during the sessions. Am I right so far?"

"Right. And the number of times you do the observations—the schedule for sampling the sessions—will depend upon whether your own staff does the observations or whether you can enlist some volunteers. Also, remember our discussion about the importance of the observers gaining the trust of the students and the facilitators, and remaining detached from the conduct of the sessions."

"Right—yes, we can't forget that," agrees Donna. "And this information will help to tell us whether our services—the PDY sessions—are actually being presented in the way we intend—correct?"

"Right again."

(After a break, the group reconvenes to discuss a second evaluation design for their alternatives program. At this point Joanne Martinez leaves and Jim Cook, director of the alternatives program joins Donna and Ron. Jim begins the discussion with a description of the program, called Brightside Alternatives for Youth (BAY). BAY is housed in the Brightside Youth Center and utilizes its extensive recreational facilities, which include a basketball court, a room containing a boxing ring and weight-training equipment, and a game room with ping-pong and pool tables. The organized sports activities include baseball, basketball, boxing, volleyball, and weightlifting. The social activities consist mainly of teen dances held every Saturday night at the center. Jim has three staff members who double as counselors and coaches. Counseling is done on an informal basis: as staff identify needs or problems in a youth visiting the center, the youth is asked to step into the counselor's office to "talk for a while." Ron is now asking Jim about the youths who are in the BAY program.)

"So the kids who are in the program are of all ages, and mostly Hispanic?"

"Yes. Their ages range from 6 to 19 or 20. Most of them are Hispanic; the rest are a mix of blacks and whites, from mostly working class families."

"How many kids are in the BAY program?"

"That's hard to say," Jim replies. "It depends on whether you count the after-school dropins, the kids who come to the dances, or just the kids on the teams. I could tell you who's on the teams, but we don't keep track of the dropins or the kids who come to the dances."

"Are any of the kids referrals from the courts or troubled youth programs, etc?"

"A few," Jim replies, "but nearly all of them are just kids from the neighborhood."

"I see," says Ron, looking a bit perplexed.

"I guess it sounds kind of disorganized, huh?" Jim laughs.

"Well, it's pretty loose and free flowing, but that's how these programs go. Now your formal statement of objectives says that your program is intended to 'provide a wide range of healthful activities to neighborhood youth . . . activities that can serve as alternatives to drug and alcohol abuse and delinquency'—is that right?"

"That's it."

"OK." Ron pauses, seemingly pondering the situation and what evaluation designs might be used with the BAY program. After a long silence, he continues:

"Clearly, we can't employ any rigorous experimental design here. You can't deny your services—the activities—to anyone or place a kid arbitrarily in one activity or another, so any notions of randomization are out. We could possibly identify a comparison group in the community, but that would be time-consuming and would probably result in a very nonequivalent comparison group. I think the best we can hope for here is to implement a process-oriented evaluation, perhaps combined with a longitudinal outcome evaluation."

"A what . . .?" Jim looks puzzled.

"I'm sorry. What I mean is that first we should concentrate on getting some information on the numbers and the characteristics of the kids who are in the BAY program. That kind of documentation is often meaningful to funding agencies, and it will help you to determine whether you're serving the kinds of kids —ages and ethnic mix—that you want to."

"And how do we do that?" Jim asks.

"Do you have a membership list?"

"Yes, but it's not really very accurate right now. I suppose we could update it."

"That would be helpful. Also, can we get some basic background information on the kids for your membership files—age, ethnicity, reason for coming to the center, etc.?"

"Probably." Jim looks toward Donna. "Do you think Carlos could get that information for us?"

"Yes, I think so," she answers, "although it will take at least several weeks."

"That's fine. Now, is there any sign-in procedure when the kids come into the center?" Ron asks.

"Yes. But I'm not sure how well it's followed. I could check that out, too."

"Good. An accurate membership list with some background information will tell us—and others—who's in the BAY program, and an accurate sign-in procedure will show how frequently they use the facilities and for what purpose."

"I like that," Donna approves. "It's something that I've been wanting to do for some time anyway. But what about outcome evaluation, Ron? Are there any possibilities here?"

"Yes, there are. . . 'possibilities,' but they're limited, as I indicated before. I suggest that we use a longitudinal approach, selecting a small, fairly representative sample of kids as they enter the program and following them over an extended period of time."

"Oh, that's what you meant by a 'longitudinal outcome evaluation,' " says Jim. "How long would it be?"

"At least several months. Perhaps as long as 3 to 4 years, if that is possible."

"Four years! You gotta be kidding! We might not even be here then," Jim explodes.

"That's true. But you have to remember that prevention programs may take that long to demonstrate that they actually help prevent future substance abuse. You have to decide the tradeoffs between how important this information could be and the cost to get it. You might get enough information to guide you in a shorter period of time, say 1 or 2 years."

"And how would we collect information from them . . . of what type, etc.?" asks Donna.

"One way to go would be to select kids aged 10 to 14, since the main goal of the program is the prevention of alcohol and drug abuse and delinquency, and the age of onset for these forms of deviance is roughly in that range. As they enter the program, one of your counselors could conduct a fairly extensive interview with them." Ron says.

"How extensive?" asks Jim. "Covering what topics?"

"The interview should cover current and past behavior related to drug and alcohol use and deviance—for example, the past 30 days, the past year, and initial experiences. It should also include some assessment of attitudes and intentions as well. Family environment and peer relations might also be tapped, since these may act as moderator variables."

"What are moderator variables?" asks Donna.

"Things which may influence, or moderate, the impact of the program on the individual. For example, we might find that the BAY experience is helpful to kids from a supportive family environment, but not for others."

"I see," Jim nods, "but shouldn't we also gather some information on their activities—how they view sports, what they like to play, how often, and so forth?"

"Good idea, Jim. The impact of the BAY program and its activities will probably be influenced by the stance the kids have already taken toward these activities when they enter the program."

"Then we would conduct the interviews again later?"

"Yes. I would suggest at points 6 months and 1 year after joining the program."

"Now," Donna asks briskly, "how will we get this interview developed?"

"It's not a difficult task for me to assemble a draft interview instrument, but you'll have to train your interviewers and conduct a careful pilot test of the instrument. A pilot test on three or four kids, coupled with an examination of the results, would give you a better notion of the resources that will be required for the fullblown evaluation. Can you do that?"

"What do you think, Jim?" asks Donna.

"We can handle that. It's the actual interviewing I'm worried about. How many kids are we talking about here?"

"A small group," Ron replies. "Probably no more than 30 kids over a 4 to 6-month period—assuming you get that many of the right age group entering the program over that period."

"No problem. We probably have at least twice that in the 10 to 14 age group entering the BAY program over a 6-month period. And if those are the numbers that we're talking about—30 or so—my staff can handle it."

"Are we going to need the computer to analyze these data too, Ron?" asks Donna.

"No, I don't think so, Donna. Our sample size will be quite small, and the analyses will be mainly descriptive and qualitative, not the kind of complex analyses you'll be doing with the PDY data. Still, just the manual tabulation of data and qualitative analysis will require time from your staff—perhaps as much as several weeks of time."

"Hmm," Donna looks concerned. "This evaluation work sure can devour resources. What if we can't spare several weeks of staff time?"

"Well, you've got a couple of options as I see it. One: you can drop the outcome evaluation for the BAY program and just concentrate on the process evaluation. Two: you can cut back on the length of the interview and on the amount of analysis. But you can't reduce it too much or you'll have very little of value. Remember, your 'return on your evaluation dollar', as it were, is fairly meager with this type of outcome evaluation—in contrast to the PDY outcome evaluation," Ron points out.

"Would it help to cut down the number of interview sessions?" asks Jim.

"Somewhat, but only with respect to the total person-hours over the entire course of the evaluation. For any given period, you would still have to devote the time to interviews, analysis, and writing."

(The group then launches into a discussion of specific roles and responsibilities for the BAY program evaluation. Ron's visit is coming to an end, so they conclude with a summary of the overall design and how it will be carried out over the next several months. Within 1 month, Donna will send Ron an outline of the plans they have formulated for both evaluations. Besides helping to prepare the instruments, Ron will also be available to review the pilot test data and to provide assistance with the analysis.

Several months pass. The evaluations have been implemented, and Ron has returned to the Brightside Youth Center to discuss the evaluation—results to date, interpretation of the findings, and utilization of the results. We look in on the group as Ron strides into Donna's office to meet with Donna, Joanne, and Jim.)

"So—I hear you folks have been conducting an evaluation!" Ron grins mischievously.

"More or less, Ron." Donna smiles, too. "We certainly have put a lot of work into it! Maybe you can tell us whether it's been worth it."

"You mean it's not evident by now?"

"Well, actually, we're already more aware of our strengths and problems," Donna admits, "but we do need a little help in deciphering these results. You did get the drafts describing the results of the analysis, Ron?"

"Yes, I did. Shall we start by looking over the results of the PDY outcome evaluation?"

"Fine," agrees Donna.

"Well, the results reflect an interesting mix of outcomes. You show some impact—significant differences between PDY kids and the control kids—on self-concept, attitudes toward substance use, one of the stress management subscales, and one of the interpersonal skills subscales. But no effects on family dynamics or on self-report of substance use.

I brought along a couple of illustrations of the data in order to explain a 'significant difference.' First, if you look at the top of figure 1, you'll see a portion of an Analysis of Covariance Summary Table. This was extracted directly from the computer output and shows the results of the 'F-test' for significance between

A. Portion of Analysis of Covariance summary table for self-concept

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F	Probability of F
Group	74	1	74	3.7	.05
Error	12524	620	20.2		

B. Graph of pre- and posttest self-concept scores

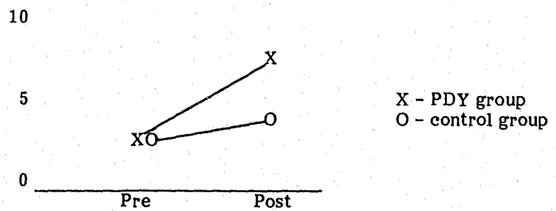


Figure 1

groups--that is, between PDY and control students--for self-concept. This tells us that, if we repeated this study 100 times, in only 5 cases would the difference between the 2 groups' scores be this large if there was no real difference. The bottom part of figure 1, which I sketched out for you, illustrates this difference graphically. Both groups have essentially the same self-concept as measured by the pretest, but the PDY group has improved considerably at the posttest. This difference—which looks substantial even to the naked eye--is what was found to be significant in the data analysis."

"Analyses of these outcome measures by school and ethnicity," continues Ron," show no significant differences or interactions—"

"What do you mean by that, Ron?" asks Joanne.

"The school and ethnicity analysis?"

"Yes."

"It means that the effects of the PDY program are the same for each school and ethnic group. However, there are some interesting differences by sex."

"How so?" asks Donna.

"For some reason, the PDY program has a greater impact on the interpersonal skills of the boys than of the girls."

"I think the boys appear to learn more of the social skills than the girls," explains Joanne, "because of the sessions where we focus on ways of relating and communicating. We emphasize to the boys that it's not effeminate to be social and express your feelings. I think most of the girls already had fairly well-developed interpersonal skills before they joined PDY."

"Certainly a plausible interpretation," Ron says. "In fact, that's what the data show. If you look at figure 2, which I also sketched out, you can see how Joanne's explanation is reinforced. As the first graph indicates, the interpersonal skills of the PDY group are much higher than those of the control group at the

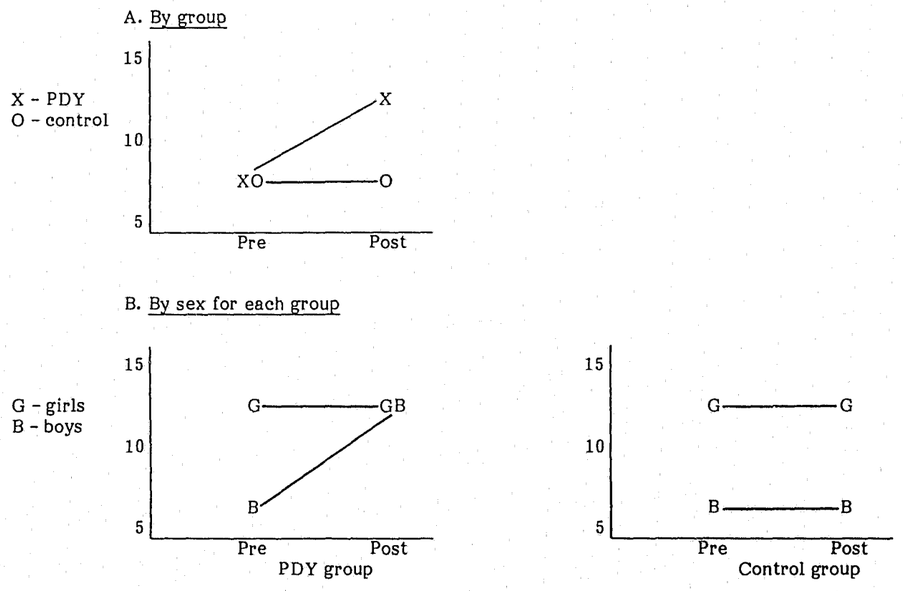


Figure 2--Graphs of Interpersonal Skill Levels

posttest. But the 'interaction' between sex and group is illustrated in the bottom two graphs which show the scores for boys and girls in both groups. The girls in both groups scored higher than the boys in the pretest, but in the PDY group, the boys 'caught up' to the girls at the posttest. So that's the key to the overall difference between the two groups. The significant difference between the groups at the posttest is due to the improvement of the boys in the PDY classes. Now, why no effects on family dynamics or substance use? These are pretty central criteria for your program."

"Well, I don't think we should have expected to influence family dynamics through PDY," says Joanne. "It's too powerful a force for us to influence in a couple of PDY sessions."

"I would agree, and, as we discussed before, I don't think you should be disappointed by the lack of impact on actual substance use in this short a time period. To really assess the effects of the program on substance use, you should follow these kids for another year or 2 when they are in the high-risk age range—15 to 18."

"Oh boy, more work down the road." Joanne casts a bemused look at Donna.

"Just trying to keep you busy, Joanne," Ron laughs. "I was happy to see that you could use a standard package like SPSS for all of the analyses. By the way, who did the computer analysis for you?"

"Hal Kleinfeldt at the university," Donna answers. "He was super. I don't know how we could have done it without him."

"How have you paid for all this?"

"A combination of great student volunteers and a small grant from the University Computer Center, through Hal's good auspices," Donna replies.

"Well, you've got some results that should be of interest to a number of people, but let's get to that later. How are you planning to utilize these findings internally?"

"We've already used them to alter the PDY sessions for the coming year," answers Joanne. "We're taking out the family dynamics sessions, and expanding the stress management component to try to show more impact in that area. Also, our process observations show that neither the stress management nor the interpersonal skills sessions are implemented in the way we intended."

"How so?" asks Ron.

"Well, both components are supposed to be built around behavioral exercises. For example, the stress management sessions were to include the actual practice of relaxation techniques by the students, and the interpersonal skills sessions were to be based on several role-playing exercises. In fact, we found that most of the sessions were of the lecture-discussion variety--the students often appeared bored and distracted. I think that's one of the reasons we didn't have as much impact on outcome measures as we'd hoped."

"That's interesting. Your findings weren't really measuring the program as it was intended to be. Instead, you were measuring, as always, what actually happened. But in this case, you found a major difference between program design and implementation. That alone is sufficient argument to justify looking carefully at the process."

Then, turning toward Jim, Ron comments, "Well, I guess BAY's outcome evaluation hit some resource problems, eh, Jim?"

"Yes, I guess you could say that. After we pilot tested your interview instrument—which worked well for the most part, by the way--it became clear that, at least at this point, our staff just didn't have the time to devote to the interviews--at least to do them with the depth and breadth we thought was required."

"Threw in the towel, huh?" Ron laughs.

"Not really! We're not quite ready to abandon the outcome evaluation. In fact, if we get the grant we've applied for, we'll add a counselor, then we should have the time to do the outcome evaluation this coming year. So we're hangin' in there!"

"Well, I think you made the wise decision. It's interesting that the same thing happened in the NPERN case study, but probably just to cut down space. Our motive is different. We've found out, through pilot

testing, that we just don't have the resources right now. Were you able to put some time in on the process evaluation?"

"Yes. That's gone pretty well," Jim replies. "We've been able to update the membership information and document the daily flow of kids into the different activities."

"Has that information been helpful to you in any way? Have you made any changes in your program?"

"Yes, it has," Jim affirms. "We plotted the membership lists on a map and found out that we haven't been attracting kids from the Ventnor district—right near here—so we've started a recruiting drive in that area. We've picked up some kids from there in the past couple of months. Also, an analysis of our daily sign-in sheets showed a huge number of kids who were dropping by the center, but who weren't members of any organized activities. That's OK, of course, but we're trying now to persuade some of them to get more involved. We've plotted that on a map, too, to help us focus our energies."

"That's great. I like the idea of the map—graphical displays aren't used frequently enough. They can also help you to get into the data and explore its meaning, often producing information you weren't looking for when you started out.

"Was there any resistance to the data collection on the part of the kids or the staff?"

"Both," Jim answers Ron with a laugh, " but only in the beginning. After the first couple of weeks or so, they all settled into the routine pretty well."

"Well, I'm glad to hear it. Good luck on the grant."

"Thanks. I'm sure we'll need it!"

"Now," Ron asks Donna. "How are you planning to utilize this information externally?"

"Well, as you know, we're putting together a comprehensive report on both evaluations. This will be sent to the State and the United Way—our major funding agencies."

"Anything else?"

"At the moment that's all we've planned."

"Let me suggest a few things. First, I think you should develop a condensed 'executive summary' of your findings, suitable for sending to the schools and others who may be interested in your findings—such as the mayor's office—but who don't want to pour through a mammoth report. Second, I think you should send these findings to other relevant local organizations and potential funders. Don't forget that each organization may be interested in a different aspect of the evaluation. Schools, for example, might want to hear more about the classroom process and its evaluation; the mayor's office may be more concerned with your impact so far and what you've done to improve the program. In addition to sending them reports, arrange to make some personal presentations as well. I think they have more impact."

"Those are all good ideas, Ron. Now if we can find the time . . ." Donna adds ruefully.

"I understand. Look, you folks have done a great job—you're to be congratulated. Doing program evaluation is no easy task."

"No, it sure isn't," Donna says emphatically, as Joanne and Jim nod in agreement. "But it's already been more useful to us than I thought it would be."

"It's strange, but that's the typical response I hear from programs after they've completed an evaluation. Invariably it has more utility than they thought it would."

"Thank you for your assistance, Ron," says Donna. "You've really been a help to us."

"Glad to do it."

FOUR THRILLING DISCUSSIONS

Planning an Evaluation of a Teacher Training Prevention Program

Characters:

- Pamela Raven. A program developer in the curriculum department of the Cinnamon Bend Unified School District.
- Lacey Strait. Cinnamon Bend School District deputy superintendent for Curriculum.
- Conrad Sizer. A local evaluation consultant referred to the District by NPERN in response to a request for assistance.
- Allen Compass. A second evaluation consultant referred by NPERN, and a colleague of Sizer.

First Discussion. A bright, crisp day in late fall. Conrad Sizer and Allen Compass have just entered the office of Lacey Strait. Strait and Pamela Raven are seated around a circular table. They rise and shake the evaluators' hands, offer coffee (accepted by Sizer, declined by Compass), trade a few comments about politics and life in a bureaucracy, extract pencils and pads. A tense silence threatens to settle. Some throats clear. Then, as if it were expected, the discussion begins:

Strait: Well, I asked you folks to come to this meeting, so I guess I'll start it off.

Compass: Sounds reasonable.

Strait: As I think I told you on the phone, Dr. Sizer, Pam Raven here has developed and launched what we think is a magnificent little program. It's intended to be a sort of indirect way of preventing drug abuse by adolescents, but it seems to have a lot of other things going for it, too. It's been operating for about a year now, run by several teachers in two schools. The program is really great. We call it the Cooperate and Progress Project, or CAPP, by the way. In fact, just about everybody loves it. Teachers love it, the kids love it, and parents love it.

Sizer: Wait a minute. Do you mean to say that there's nobody who doesn't love it? If that's the case, this must really be a first in education!

Strait: Oh, of course there were a few parents who didn't want their kids to be in the thing, some people in fringe groups who have complained, a few letters to editors in community newspapers, and the like—but compared to most of the new programs we've tried, there hasn't really been much criticism. Despite the fact that everyone likes the program, our school district is in a funding crunch. I went to a Board of Education meeting last month with Pam, expecting to request more money for expansion, and they told us out of the blue that all special teacher training funds would be cut next fiscal year. On top of that, they announced their intent to go back to "basics." So Pam and I had to do a quick turnaround to convince them to just consider maintaining it.

Sizer: Hmmm.

Raven: Yes, we got a reprieve. Rather than cutting us off immediately, Dr. Strait convinced them to consider continuation only if we can show them that the program works.

Sizer: OK, it looks like we know who we are evaluating for. Now, the question is what do they mean by "works?"

Raven: They are concerned about showing that it prevents substance abuse and other deviant behaviors—but they made it very clear that if we can't demonstrate that the program teaches the basics at least as well as traditional methods, it's out.

Strait: I got angry myself, since we know the program works.

Sizer: But how can you be so sure the program is working if you haven't evaluated it?

Raven: All you have to do is look at the classes, look at how the kids are getting along in those classes, look at their faces, talk to them a little . . .

Sizer: Have you done the same things with kids and classes that aren't in the program?

Raven: Well, not as much, I suppose. But I still know.

Strait: This is turning into a debate about the need for outcome evaluation, which is all very interesting, but is not what we're here for today. If we want to continue, we've got to evaluate the program, so we may as well start with the assumption that that's what we're going to do.

Sizer: That's a perfectly good reason to have an evaluation. In fact, most evaluations have survival as at least a partial motive. I think there's a positive value in doing a careful and controlled evaluation—sometimes in conjunction with a simultaneous collection of subjective impressions of sensitive observers, participants, and so on.

Compass: Well, that's part of the process evaluation that should naturally accompany the outcome evaluation.

Sizer: Yes, it is, but that's closer to the kind of subjective evaluation they've already done, and I was trying to draw a distinction.

(A brief silence ensues. Those with coffee sip.)

Compass: You know, I just realized something. I don't really know what we're talking about! We're supposed to be discussing the evaluation of a program, but the only thing I know about it is what we discussed on the phone. Do you think we could hear a description of it?

Strait: Yes, that's how I intended to start, but we seem to have gotten sidetracked. Pamela, since you developed the program and know the most about it, why don't you give a brief description of it?

Raven: I'd be glad to. (Looks at visitors.) Please interrupt me whenever you have a question. Well, the project got started out of dissatisfaction with some of the other approaches to drug prevention with adolescents and pre-adolescents. So many programs have tried to approach the problem head-on, with horror stories, rewards—the kids often see them as bribes—or large doses of information. It seemed to me, from watching some of the programs in operation and from talking to some of the kids, that these direct approaches made the kids resistant, suspicious, and negative. They saw it as propaganda being forced on them by narrowminded adults. So, I thought a more indirect approach might work better. In thinking about an indirect approach, it seemed to me that instead of focusing on drug use per se, or even on attitudes specifically about drug use, it might be better to focus on some of the psychological factors which seem to predispose kids toward using drugs—if my reading of the research literature is correct—things like low self-esteem, low feelings of personal control over the environment, low self-control, and the like. The idea, then, was to develop a school program which would have meaningful effects on these kinds of things fairly directly, and would then influence drug use and drug attitudes only through its influence on these psychological factors.

Sizer: I like your thinking, but that doesn't seem like a very easy task you set for yourself. These psychological factors, as you call them, sound like things that are fairly deeply ingrained in the personality. I would think they might be even harder to change or influence than drug use!

Raven: Well, first, thank you for the compliment. As for your second comment, I thought that way myself at first, when I saw which psychological factors had been found to be related to drug use. But then Lacey showed me some descriptions of "cooperative learning groups." They've been used in regular classrooms, desegregated classrooms, and classrooms with handicapped or "mainstreamed" children, and have shown effects on some of the very same variables that have been found related to drug use in adolescents. So, it seemed like it might be worth trying with adolescents and preadolescents to see if it did have some effects on drug use and drug attitudes. So we worked up a program and got some teachers to try it in two schools, as Lacey said.

Compass: What do these cooperative groups do? How do they differ from regular classrooms?

Raven: Well, we use the same curriculum as the regular classroom, but we do it differently. Instead of kids working by themselves and maybe competing with others for grades, praise, etc., we try to set it up so that they benefit from each other's learning. We use a method called "Jigsaw," developed by Elliott Aronson. It's called Jigsaw because a unit of curriculum is divided into pieces, which are fitted together by the kids in a group. Say the class is covering a unit on the Civil War. You divide the class into six-person groups, and you divide the Civil War readings into six sections. Each member of each group is responsible for learning one of the sections and then teaching that section to all the other members of that group. Before teaching the section to the other group members, the kids from each of

the groups who have the same sections to teach get together and help each other learn that material and decide on the best ways to teach it to the other group members. Each member of the group becomes responsible for the learning of all of the other group members. The group gets graded as a whole on that unit of the curriculum, so no one individual can benefit unless all the group members learn the material well. Aside from learning the curriculum very well—an approach shown to have positive effects on academic achievement—kids in these classes learn to pay attention to the needs of other kids, to adjust their teaching so that each of the others masters the material. They learn to be concerned about other people, and they learn that they can really make a significant contribution to the welfare of everyone in the group. This helps them to feel better about themselves. If you watch a class that's going well, you can see this happening!

Sizer: Well, we can't use grades as a measure, since the kids are graded by their own teachers. Do you use standardized readiness or achievement tests?

Raven: Sure. Every class level has a broad achievement test at the beginning and end of each school year.

Sizer: Where do they keep those records?

Strait: Oh, on that fancy computer! Do you know that while they're trying to cut back teacher training they're planning to buy an even more expensive one—after only 3 years.

Sizer: What other student records do they keep on it?

Strait: Everything. They keep track of absences, tardiness, disciplinary actions—grades, too.

Sizer: Great. That will cut down on data collection costs if we decide to use those variables. Paper-and-pencil tests are my stock in trade, but behavioral measures are usually the best, provided they are directly related to the objectives. The standardized tests should be good measures of academic change. Absences have been shown to be associated with substance abuse and, in fact, a host of delinquent behaviors. Disciplinary actions speak for themselves.

Strait: We've got to be concerned with cost, because the board won't give us any extra money for the evaluation. That's one of the reasons we called NPERN.

Compass: We'll keep that paramount when we develop the draft evaluation plan. Meanwhile, I'd really like to see one of these classes operating. Are any of them in session now?

Sizer: I'd like to see one, too.

Raven: Yes, there are several, and you'd be most welcome to come and visit.

Strait: Before we set up any specific visits, I'd like your comments on whether or not the program can be evaluated.

Sizer: It seems to me, from what I've heard so far, that a feasible evaluation design could be developed. You seem to have a fairly clear idea of the major variables you are trying to influence, both directly and indirectly, and at least a rudimentary theoretical model that lays out some of the mechanisms of influence. The process in the classrooms sounds fairly well specified and observable, and reliable measures of some of the psychological factors already exist. I'd like to see some of the classrooms in action before making a final decision, but as of right now, I'd say that a decent evaluation can probably be developed. What do you think, Allen?

Compass: I feel certain that a good evaluation plan can be developed, and I'm ready to start on it right now. But first, Lacey commented that there have been other studies that show positive effects on academic achievement. You might be able to persuade the board to use those findings as justification for continuing the program next year, to give us a chance to evaluate it. If you can do that, we can start some of the preliminaries now.

Strait: Well, let me make a suggestion, then. You and Pam can set up some visits to classrooms in the next few days. Then think about it for a while. Read some of this material we've put together on the Cooperate and Progress Project (hands Sizer several documents), discuss it with each other, talk to Pam if you need any more information about the project as it has operated in the past year, or as we're thinking about it for the next year, and then give me a call and let me know if you're interested in working on it with us. Meanwhile, I'll go back to the board.

Sizer: Right.

Compass: Sounds good.

Second Discussion. A dull, cold day in early winter. The same four people are sitting around an oblong table in a meeting room in the evaluators' offices at the university. Tables around the sides of the room are piled neatly with stacks of computer printouts.

Strait: As you know, the board approved our continuation based on our summary of the literature on academic achievement, but we still have to show that it works here. So let's look at the evaluation draft. You folks really did a nice—and, what's even better—a quick job developing the draft of the evaluation plan. I want to say I'm really glad you decided to take this on. I think we're going to work well together. Pamela and I do have some questions about a number of points in your plan, so maybe we can just go through them. My first question is this: why on earth do you have those observers in there? It's going to disrupt the classrooms!

Raven: Lacey, please! Calm down.

Sizer: Well, there are a lot of things we need to look at. We need to look at the psychological change and academic achievement that are considered the most direct outcomes of the program, and we need to look at the more directly prevention-related variables. And, finally, we have to see what's actually going on in the classroom.

Strait: Yes, but the number of hours of observation you're calling for is going to wreck the program. The teachers won't stand for it.

Raven: I'm afraid I agree with that. Remember, we have two major interests. We want to get some ideas about the psychological processes being affected, but we also have to satisfy the board, just to keep the project going.

Sizer: But don't you want to know, in some really well-documented sense, whether it's having the effects you think it has? And if it is effective, aren't you interested in having it adopted by other school districts?

Raven: Well, of course, but—

Sizer: Well, the best way, and certainly the most responsible way to get the project first known, and later adopted or adapted by other districts, is to have its effects clearly and rigorously documented.

Compass: I hate to say this, but I don't think the history of educational fads bears out what you say, except for the "responsible" part. I mean a lot of things have been taken on without any real evidence at all.

Sizer: Well, of course, but surely we don't want this thing to become a fad. If it is shown to be effective, and the reasons for its effects seem to be fairly well understood, then it should be adopted. Short of those conditions, it shouldn't be adopted, at least very widely, no matter how attractive and intriguing it may sound.

Compass: Them's tough words, pardner.

Raven: Actually, I think I agree with you. We don't want this to become a fad—in one year and out the next. We want it to be "solid," and if it takes tight research to make it solid, so be it. But the amount of observation still bothers me.

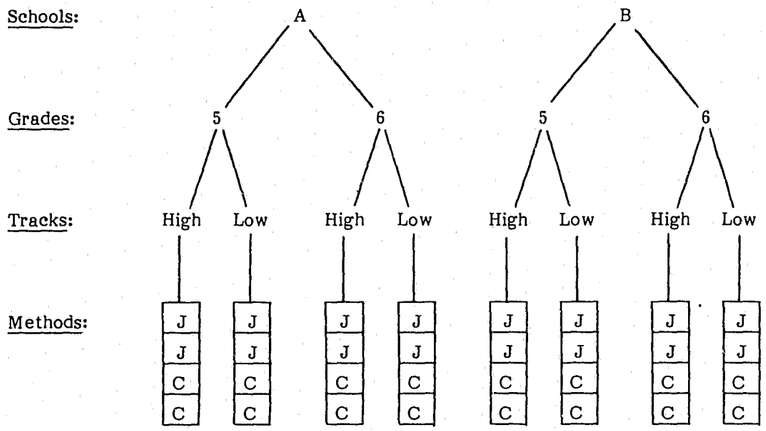
Strait: Well, that leads me to another question. First, I'm a little confused as to how we're going to pick the kids to get the Jigsaw program. Right now, it's in place in a little over a dozen sixth grades in two schools. Some of the teachers involved, however, are actually fifth grade teachers. How are we to actually select Jigsaw and non-Jigsaw students?

Sizer: In the ideal case, we would randomly assign students to the two different teaching methods, Jigsaw and--we'll call it--Control, in each of the two grades. But, we're dealing with intact classroom units, so we have to assign entire classrooms to Jigsaw or to Control. This also has implications in the analysis stage, but we'll get to that later. You also have two tracts—high and low—in each——

Compass: Conrad, I think it might help to draw the design out so they can follow it a bit easier.

Raven: I hope it will help!

Sizer: OK, what we have is called a 2 x 2 x 2 x 2 factorial design, that is with 2 schools, 2 grades (with 8 classes in each), 2 tracks and 2 teaching methods. Like this (going to the blackboard):



The row of boxes represent the actual classrooms that will either have Jigsaw (J) or won't have Jigsaw (C for Control). There are 32 classes in all, 16 in each school. Let's just look at one school first. School A has eight fifth grade classes and eight sixth grade classes. Four of each are in the high track and four of each are in the low track. Since we have four Jigsaw teachers in each grade in the school, we can randomly select two of the four classes at each track-grade combination to receive Jigsaw. The same logic holds true for School B, so that we have a nicely balanced design, controlling for school, grade, and track.

Raven: Is this that tight research I just mentioned? Do we have to control for everything at once?

Sizer: Well, it's as tight as we can make it given the overall situation. There are only two classes of each method at each track-grade-school combination, and 32 classes overall, but the program has to survive before you can get more elegant.

Compass: In response to your second question, to rule out other explanations for any differences we find between the classes, we have to measure and test the effects of other possibilities. For instance, there could be differences between the two schools, if the atmosphere and/or environment differ. With the proposed design, we can also see if Jigsaw seems to work with one grade or track better than the other. If we just lumped all the classes together and selected half to get Jigsaw, we might just get certain effects canceling each other out in the data, showing no overall effect, and have no way of breaking the results down.

Strait: Well, I have a clearer picture of the random assignment, but I still have a question: Isn't the purpose of random assignment to make the experimental and control groups equivalent? Your experimental design now calls for two testing periods during the year. A set of pretests in the fall and a set of posttests in the spring. But if the random assignment has made the two groups equivalent at the beginning of the experimental year, why can't the testing simply be limited to the spring (or post) testing?

Sizer: In the first place, just doing the random assignments isn't enough. You have to have the data to determine whether the random assignment has actually resulted in equivalent groups at the start of the project. And the characteristics which you most want to be equivalent at the start are the ones you're trying to change--the very things you measure as the major outcome variables. While that's sufficient

reason, the fact that we're talking about random assignment of classrooms, rather than individuals, makes assessing the variables at the start of the study even more important. We're talking about a relatively small number of classrooms, 16 experimental and 16 control. Finally, even though our major interest is in individual-level variables, most of the analyses will involve group-level aggregations. It's essential to determine whether the school and class assignment to teaching methods has resulted in equivalence at the individual level.

Strait: That's very convincing, although I didn't understand all of it.

Raven: I have a different kind of a question. I guess I have a kind of proprietary interest in this program. Oh, other people have worked on it with me, but, well, it really was my idea. Anyway, we have carefully constructed it out of several elements which fit together just so. And I get the feeling that this evaluation is looking at one piece, then another piece—sort of pulling the wings out and looking at them one at a time. I just wonder how it's going to arrive at a picture of the whole thing—it's an organic system, not just the sum of its parts. I think you're planning to look at the parts—the individual pieces, but not at the whole.

Compass: That's a very good question, and it's one I sympathize with. I do think we have a way of getting at the program as a whole, but it may be a bit underplayed in the proposal draft—you may not have noticed. Actually, we have two ways of handling it, one more quantitative and one more qualitative. First the quantitative way—although we plan to measure a number of variables individually and one at a time, our analyses won't be limited to looking at them individually, at least not all of the analyses. Many analyses will use multivariate statistical procedures to identify patterns of variables. That is, we'll try to recreate statistically the complexity of the program and the program's effects. But, to really appreciate the complexity, we recommend classroom observation. A number of sensitive qualitative observers will go into the classrooms and observe the general aspects of their social structure, atmosphere, interaction patterns, etc. Initially, these analyses will be done independently of the more statistical analyses. Later, the two sets of analyses will be looked at together. We expect that the qualitative analyses will help add flesh to the quantitative ones, help us in their interpretations, and help identify new variables and new ways to investigate the quantitative data. The quantitative data will similarly provide an empirical anchor for the qualitative speculations. Each, we hope, will strengthen the other.

Raven: OK, here's another question: Your proposal stresses assessing the adequacy of program implementation, process measurement, and the like. I think I understand the purpose in general. I mean, it's nice to know that people are running programs right and all that. But, in the first place, I think I have a pretty good idea about the program already, and in the second place, there's not much you can do about it anyway if it's not being run correctly, is there?

Sizer: There are both program reasons and evaluation reasons for doing a careful assessment of program implementation. You can do something about it if it's not being done right. That is, you can if you know about what's happening. You may have a good idea about how well the program is being done, but you need much more specific information to relate that sense of knowing the classroom to the quantitative data.

Raven: I guess I had a misconception. I thought that once you set up a formal evaluation of a program, you'd be stuck with what you get, and that you couldn't use the research results to alter the operation of the program in the middle of the evaluation.

Sizer: What you had was only a partial misconception.

Raven: I don't know what you mean, but it makes me feel better.

Sizer: Well, if your observation of the program leads you to believe that the initial program plan may be incorrect and that two or three of the program elements, even though they are being implemented well, should be altered or dropped, you should try to restrain yourself. Making such changes might be good for the program (although it would be difficult to document that it was, short of doing a second evaluation), but it would be disastrous for the evaluation. To evaluate a program, the program has to first be definable. The process observations help in the definition. But if the program changes into something different halfway through, the evaluation cannot clearly generate information about the program after the change, as distinguished from the program before the change. Thus, it would be better to note your ideas for changes in the program as they occur in the course of doing the evaluation, and then to test them later, in an evaluation of a revised program (which would also be informed by the results of the evaluation of the initial program).

But, if the process observations show that the program is not being implemented as originally planned, it is perfectly permissible to bring this to the attention of the program implementers and to try to get it changed so that it becomes adequately implemented. If the program is not adequately implemented, it is not the program which is being evaluated, but a distortion of the program.

Raven: But does that really work? Is it really possible to train implementers so well that they all produce similar, and equally adequate, versions of the program? After all, people vary, their skills vary, and their temperaments vary. It seems almost impossible. And if it is impossible, what does that do to your neat little evaluation designs?

Sizer: Well, you're right, that can be a very serious problem. There are some ways of handling it. But before I go into them, tell me, how much variation do you think there is in the way the teachers implement the program now?

Raven: A great deal! All of the teachers are volunteers, of course, but even so, there are great differences. A few of them seem to understand the program completely, are very interested in it, and do it very well. Some others work really hard but, don't quite seem to get the idea. And others really show a pretty low level of involvement.

Sizer: Have you worked much with the teachers who are less good at implementing the program?

Raven: Oh yes, at least we've tried. We do most of our work with the people who want to do it and are willing to work at it, but have difficulties with it. With those who are really not interested, there's not much we can do. I guess what has kept us going is that the program looks so nice with those teachers who do it well.

Sizer: It's a crucial problem, and one of the major uses of the process data, as I just suggested, is to get useful evidence quickly about where and to what extent individual teachers may be going wrong. Of course, an intensive initial training involving class tryouts and frequent feedback is also essential. It's also important for teachers to have a say in the definition of the program—that is, in helping decide the best specific ways to implement the program in the classroom. Do you involve teachers in the planning at all?

Raven: Well, we've had a few teacher representatives work with us. The actual participants get a lot of training, but aren't much involved in planning. I can see that it might be a good idea, though.

Sizer: I consider it essential, for two major reasons--in the first place, it will greatly improve the program. Teachers know the classroom and how to make things work in it better than anyone else. You'll find that they have a lot of useful ideas about the best ways to make the program work. Secondly, teachers who have a real say in defining the program, and it's important that it be a real say and not a token, will become committed to it, involved in it, and will do everything they can to make it work. Teachers who feel that something is essentially being imposed on them—even if they have "volunteered"—are much more likely to be indifferent and even resistant to the program goals.

Raven: That makes sense. But let's get back to the uses of the process data. What kinds of data are you talking about?

Sizer: Several kinds. But before we discuss them, it's important to emphasize that all of the data will be kept confidential at the individual classroom level. And the teachers must be made aware of that. We have to make it clear that the program is on trial, not the teachers. Now, all of the process data stems from observation of one kind or another. The first is done by the trainers. By the end of the training period, there should be a pretty clear idea of how the program should look when ideally implemented. But the trainers' work won't end there. When the teachers go into the classrooms with the program, the trainers must make frequent visits to observe the classrooms as the teachers attempt to implement the program, and then to give fairly immediate feedback. This will be fairly informal observation, although we'll develop a simple observation and feedback form to aid in this process and to make it somewhat comparable from teacher to teacher. But the evaluation staff will also do some more formal and more structured observations, using intensively trained observers. The observers will visit each of the classes on several occasions during the course of the year, and will look for a number of specific indicators of frequency and adequacy of implementation of the cooperative instructional program. The descriptive results produced by these observations will be shown to the trainers, who will be able to use them in their feedback sessions with the teachers.

CONTINUED

1 OF 2

Raven: All that is impressive, and I can see that you've really thought about it, but after all the training, all the visits, and all the feedback, there are still going to be differences in the way different teachers do the program.

Sizer: I'm sure there will, too, but I hope that after all the training, visits, and so on, at least the differences will be in a fairly narrow range. Don't forget, the process data will have an important research function, as well as the program quality control function. In the first place, it will allow us to document fairly rigorously exactly what the program was, as delivered. But aside from that, if differences do occur, we will be able to see what effects these natural variations in program implementation have on the measured program outcomes. They might give some initial evidence about whether some elements (particular teacher skills, for example) of the program are more important than others in producing those effects. We could follow this evidence up with more controlled studies later.

Strait: I'd like to hear more about how you're going to deal with teachers who think you're really evaluating them, or that someone else might use the data for that purpose. I mean, you're going to analyze data, write reports, publish results (if I know you guys). How will teachers know that their data won't be identifiable? Can you make it anonymous when you collect it?

Sizer: No, we can't make it anonymous. For one thing, we want the trainers to have access to it, as I said, to help them improve the program delivery where it is needed. Besides that, for purposes of the data analysis, we'll need to be able to identify all the different kinds of data that come from the same class. But the identity of the teachers won't be given away by any of the reports. Results will be reported on a statistical basis, in terms of relationships between variables, not in terms of the performance of individual teachers. Still, at some level, the teachers will have to trust us. We'll have their data; we'll tell them that it won't be used for evaluation and it won't be revealed. I hope that we'll be able to establish good enough relationships with them so that they will believe us and feel secure and safe with us.

Raven: I think most of the teachers will accept that, if you establish a really good rapport. But I'd like to get back to the process data and analyses for a minute. You've mentioned two uses for it. Are there any others? What benefits will there be to the program?

Sizer: Well, I think both of those uses will benefit the program. Producing data for analyses which will be used to determine the effectiveness of the program is a benefit. Not only will it give an overall evaluation of the effects of the program (in combination with the other data we'll get), but it will give some objective evidence about program components which might need changing or possibly eliminating. The data will be useful, in other words, for making revisions and improvements in the program.

Strait: Do you really think your analyses will be done quickly enough so that we'll be able to use the results to justify continuing the program? I've had experience with program evaluations before, and getting the results out of those folks takes forever!

Sizer: Well, it's hard to make guarantees that involve things you don't have complete control over (like computer crashes), but we'll certainly try. We usually try to phase our work, so that we get some overall results quite quickly (starting with the purely descriptive data), and move later to the more fine-grained and detailed analyses.

Strait: When I see it, I'll believe it. Speaking of the data, though, there are some questions about that that we ought to talk about. You're going to be producing an awful lot of information about this program. How is that information going to be used?

Sizer: Wisely.

Strait: My, but you're reassuring.

Sizer: A long technical report will be submitted to you for your use. We'll prepare a report for the Board of Education. I recommend that in addition to a written report, you prepare a brief presentation to be made at a regular meeting. If you think it's necessary, I'm sure one of us could go with you for technical support. Short summaries of the project and the findings will be sent to the participating schools and possibly other audiences. If the results come out as hypothesized, there will be a great deal of interest in the project and a good chance of its being considered by other school districts. That should be encouraged, if some way can be established to make certain that the program doesn't become distorted in translation.

Strait: Now wait a minute. You're talking as if you're going to be in complete control of what's said and to whom. Don't forget this is our program. You're just being called in to do the evaluation. So I think we should have final say in all matters concerning interpretation and dissemination.

Sizer: I can't agree with that. My assumption has been that we would determine the content of all reports that describe the evaluation, and that you would determine the content of reports that present or describe the program. Reports that do both we could work on collaboratively (and, of course, any evaluation report will need to have at least a brief description of the program). Or, we could get your approval for the portion of an evaluation which describes the program operation and goals. But the description of evaluation procedures, outcomes, and implications is our responsibility and must be under our control.

Raven: That sounds reasonable to me. Besides, I don't think it looks good when a program appears to be evaluating itself. Results are more convincing and credible (especially positive results) when the evaluation is clearly seen to have been done and reported by some independent group.

Strait: You've got a point there. Not a good one, but a point. If those are the only conditions under which you'll take on this job, I guess we'll have to go along with it; but frankly, it makes me a little nervous.

Sizer and Compass: Why?

Strait: If the results are clear, straightforward, and positive, there's no problem. It's when the results are negative, or a little muddy, or "open to interpretation——"

Sizer: Surely you wouldn't want us to minimize or distort negative findings?

Strait: Oh, heavens no. But there are different ways of looking at things. You don't know the ins and outs, the political machinations, the specific catchwords that are bound to set off one or another community group. At least, I would want to have the chance (and maybe this should be formalized) to review any reports you prepare and to make suggestions about wordings, emphases, and the like.

Sizer: I can agree to that, and even welcome it (since you do have such extensive knowledge of your school district and your community), as long as it is understood that any comments or suggestions are advisory and not mandatory. We would certainly consider any of your suggestions very carefully and seriously, and would probably accept most of them, but I don't want to be bound to that beforehand. There's an additional mechanism we could use. If you had any disagreement, you could include your own statement as an addendum to any of the reports.

Strait: That doesn't completely satisfy me, but I can accept it. What about dissemination?

Sizer: What about it?

Strait: Well, since you're going to be preparing all these fancy reports, I think we should take on the job of deciding who they go to, and sending them.

Sizer: We should do that together. I think we should decide, fairly early on, exactly how many reports we want to have, directed to which audiences, and prepared at which times. We should do this long before there are any results. Then, when the reports are ready, we should send them to the audiences decided on earlier, no matter how the results look.

Strait: That's all right in principle, but remember, except for the board, there will be much less interest in the findings if they're negative. Some of those audiences, especially the general ones, just won't care about it if it doesn't tell them something clear and definite.

Sizer: That's probably true, but I still think we should send them all to the audiences which we originally select. Those who don't want to read the reports won't, and no harm will be done.

Raven: I think this discussion is a little silly, since we have to present it to the board anyway, and I know the results are going to be great! Didn't I already tell you that? They're great already (I think). Now, if you don't mind, I'd like to turn to one or two other matters. (General assent.) As you know, in the program as we've been running it so far, we haven't been doing too much formal assessment. But, since our long-range intent is to influence drug use and attitudes about drug use, we can expect to be asking some of the kids in the program some questions about such things. Now, from time to time in the past, we've gotten some pressure—I won't say from whom—to make sensitive information available to certain

people. We haven't done it, but since you'll be collecting more thorough and more systematic data, you can expect to get such pressures even more strongly than we have. How would you handle that?

Sizer: Before we collect any data on any topic from anyone in this project—teachers, students, anyone—we will make it very clear that this information is confidential and will be seen only by project staff and no one else under any circumstances. This includes parents who ask to see data about their children, and teachers who ask to see data about their students, as well as anyone else. We can take on this project only if this is understood from the start.

Raven: That's good, we agree on that. Except the part about teachers seeing their students' data. If its nonincriminating material, like self-esteem scores, feelings of personal efficacy, and the like, mightn't it help teachers to plan the best academic program for their students if they know about some of these characteristics? What would be the harm?

Sizer: All of this is personal information. It may not be incriminating, but it is private. We feel it's essential to assure confidentiality, both because it increases the possibility of truthful responses (since the children can assume that no one who knows them will see them), and because it's a way of showing respect for the integrity of the individual.

Strait: My, my!

Raven: I think I'm going to like working with you . . . except . . .

Sizer: Except what?

Raven: Well, we've had a pretty informal and free-flowing program up to now. We've had some general guidelines, but people have done pretty much what they wanted, when they wanted. Now you're going to come in, make us define the program very specifically, determine what skills are needed, what all the elements are, train a whole bunch of people——

Strait: I've been trying to get you to do that for quite a while, if you remember, Pam.

Raven: Yes, well it just seems the whole character of the thing is going to change. We'll have to be rigid and precise, we'll have to decide on a set of procedures, and then not change for a whole year. I'm afraid all the fun is going to go out of it.

Sizer: Just think of it as reaching a new phase in the life of the program. You have completed the experimental phase, developed some procedures, tried and discarded some, looked at some intriguing hypotheses. Now you've reached a point where these procedures and hypotheses can be put to the test. To do that properly, you have to keep careful control over the definition of the elements of the program, over the ways in which they are operationalized, and over the specifics of their implementation. It may not be the same kind of fun you had when you were first developing the ideas and procedures, but ideas and hypotheses are worthless if they're never put to the test.

Raven: I understand that, in a way, but I can't help wondering whether by standardizing and routinizing the procedures and overwhelming everybody with data collection, we might be stamping out the very elements that may have been most important in making the project successful (and, as I told you, I know it was!) when it was small and experimental—the enthusiasm, the excitement, the uncertainty about where it was leading.

Sizer: Well, in a sense, those are components of the program, along with specific program activities. Any program will be more successful with enthusiasm and commitment than without them. But I hope it will be possible to do the program rigorously and completely without eliminating these "emotional" qualities. Remember, most of the teachers have been doing this for a year. If the thing is handled properly, there is no reason why they shouldn't be as enthusiastic and excited as last year. I think what you have expressed is an important concern that we should all be aware of, and try to take steps to counteract.

Strait: Well, I don't seem to have any more questions just now. Do you, Pam?

Raven: No.

Sizer: Well, we'll refine the evaluation plan to incorporate some of the things we discussed today, and then why don't we get together again in two weeks?

Strait: Suits me.

Raven: Fine.

Compass: Here we go again.

Third Discussion. More than a year later, late spring. Conrad and Allen have just entered Lacey's office, carrying several copies of the evaluation reports. Our four characters sit around the conference table, ready to work.

Raven: From our talk on the phone, I know you have some good data for us; but frankly, I didn't fully understand what you were talking about. I got lost when you mentioned statistical interactions.

Sizer: OK, let's tackle that one by talking about the board's primary concern first — achievement results. To clarify it, I'll make notes on the blackboard as we go along. As you remember, we have a design including (writing on the board):

Factors	Levels
Schools (S)	2
Grades (G)	2
Tracks (T)	2
Methods (M)	2

So, within each of two schools, we have two grades (fifth and sixth); within each grade we have two tracks (high and low); within each track we have two methods (Jigsaw and Control). Well, our question is—does Jigsaw improve academic achievement? Now, that might be the case in only one school, one grade, or one track, or in any of the combinations of these factors. Our goal is to find out statistically if any of the variation in scores can be attributed to any of these. Let me lay out the possible combinations on the board. My laziness compels me to use the abbreviations S, G, T, and M for the factors:

Effects	Significance
Pre-test	n.s.
S	n.s.
G	n.s.
T	n.s.
M	$p<.05$
SxG	n.s.
SxT	n.s.
SxM	n.s.
GxT	n.s.
GxM	n.s.
TxM	$p<.05$
SxGxT	n.s.
SxGxM	n.s.
SxTxM	n.s.
GxTxM	n.s.
SxGxTxM	n.s.

You'd think it would have been easy to simply compare all Jigsaw classes with all Control classes, but our findings show how important it is to look at all the other factors. There is a statistically significant difference in the scores of the two methods (after controlling for the pretest using ANCOVA), but, if you'll look at the TxM interaction, you'll see that this is also significant. This says, in simple terms, that the effect of the method differs between the tracks, or, in the jargon, there is an interaction effect. Even though the Jigsaw classes as a whole differed significantly from the Controls, most of the difference is due to the improvement of the low-track Jigsaw classes.

Now you can see the necessity for testing all combinations of factors. So you can go to the board and say, "Regardless of school or of grade level, Jigsaw classes in the low track scored higher than their Control classes. High-track Jigsaw classes had scores which didn't differ significantly from their Controls."

Strait: So Jigsaw improved academic achievement for the low-track classes and didn't affect it for the high track.

Compass: Exactly. The results for self-esteem are more straightforward. The only significant effect was for method. That is, the Jigsaw classes had, overall, higher scores than the Controls.

Strait: Regardless of the other things—er, factors?

Sizer: Yes, both the other factors and all the interactions were not significant. So on this, you can simply tell the board, "Jigsaw improved self-esteem."

Raven: Then we should be able to satisfy the board. Even though not all students improved academically as a result of Jigsaw, I'm sure they'll see the importance of the change in the low-track classes. That's really exciting! But I don't think that the board will be impressed with improved self-esteem, even though we see it as being associated with behavioral change.

Compass: Well, we did find one difference in actual behaviors. When we went back to the school records and checked attendance, tardiness, and disciplinary actions for the last 4 years, we found that the Jigsaw classes had significantly fewer disciplinary actions this year than the Controls.

Strait: But what about attendance and tardiness?

Sizer: There were no differences in either direction for either of those variables. So on this—

Raven: We can tell the board, "Jigsaw reduced disciplinary actions."

Sizer: Wrong! I said that they had fewer than the Controls—I didn't say that they decreased from previous years. In fact, they increased! But they didn't increase as much as the control classes.

Strait: That's understandable. As students get older, they tend to have more disciplinary actions. What you're saying is that Jigsaw reduced this expected increase.

Sizer: Right! And that's what you can say to the board.

Compass: There's another important element to this. Remember that we have to consider as many plausible alternative hypotheses as we can. Let's suppose that Jigsaw teachers didn't make referrals for the same disciplinary problems, but instead handled them in the classroom. To consider this possibility, we also analyzed only nonclassroom related disciplinary actions. We got the same results. And our observations support this.

Raven: Tell us more about the observations. We certainly got a lot of help from the immediate feedback the observers provided on the implementation of Jigsaw. Some of the teachers improved tremendously in their ability to use it.

Compass: As my partner said, the observational data supported the significant quantitative findings. But more than that, they've provided us with a wealth of information in three general areas, as they relate to the Jigsaw process. They are training, teacher, and student characteristics. The details are covered in our report to you, but I should comment on the highlights. The training would probably be enhanced by increasing role-playing and focusing on teacher versus student control in the classroom. This issue seemed to underlie some of the implementation problems. In fact, several of the teachers said exactly that to the observers.

Sizer: That ties in with teacher characteristics. It might be that better training and teacher selection could be achieved by taking something like authoritarianism into account. But that's a hypothesis for future testing.

Compass: And another one that really interests me is similar to the question of tracks. We know that low-track classes improved with Jigsaw compared to high-track classes. But other student characteristics may cause effects. What about girls compared to boys? Or, what about differences in motivation?

Sizer: It interests me too, but they would need a sizable grant to get to that level of detail, and right now they just want to survive. But it is important to note that the observers saw significant differences between Jigsaw groups even within the same classroom, and that one of the major comments was that groups that had more girls seemed to function better.

Strait: OK, we have the reports and your clarifications. They should guide us in developing the verbal presentations to the board. We're ready for them.

Fourth Discussion. Two weeks later, Sizer and Compass are alone in Sizer's office discussing Jigsaw and wondering what happened at the Board of Education meeting. The phone rings.

Sizer: Hello?

Raven: Conrad?

Sizer: Yes?—Oh, hi Pam, we've been hoping you'd call. In fact, Allen happens to be here right now—let him get on the other line.

Raven: OK, Lacey's on an extension here.

Sizer: Great--so how did it go?

Raven: Terrible, we didn't even get a chance to present it to the board.

Sizer: What! What happened?

Strait: Well, basically the board said they'd simply run out of money and couldn't fund more training regardless of how good the program was.

Compass: Oh,——!

Sizer: I don't believe it. But, when did you hear? Why didn't you go to the board meeting?

Strait: The board president phoned day before yesterday, saying their budget committee had just reviewed the latest fiscal year figures, and there was no way they could continue outside teacher training, for Jigsaw or anything else . . . .

Raven: I'm so depressed. I spent the whole day yesterday letting the Jigsaw teachers know about the board's decision.

Sizer: Hang on a minute, Pam. I want to hear about that also, but I'd like to know the whole story on the board first.

Raven: Right. I'm just still angry . . . .

Strait: So the president said she was sorry but didn't think there was any point in making a board presentation if the decision was already made and took us off the agenda.

Compass: And that was it?

Raven: Well, maybe one or two glimmers in the gloom.

Sizer: Like what?

Strait: The president said both she and another board member—what's his name, Pam?

Raven: Lengenfeld.

Strait: Right. I can never remember him for some reason. Anyway, she and Lengenfeld had both read the full evaluation report we gave them to review before the meeting and were quite interested in the results and might try to help us find some outside support, foundation or whatever.

Raven: But how real can that be?

Strait: Well, I'm not sure. It may just be a bone to soften the blow, but I had a feeling there may be some real interest there--at least she talked as if she had actually read the thing and it seemed to have gotten her more interested. She was asking all kinds of questions . . . .

Compass: Hmmm, that's something to consider. I'm still reacting myself . . . . When I think of the hours we put into it, to say nothing of your time, and the teachers—it's just disappeared down a tube . . . .

Sizer: How did the teachers react, Pam?

Raven: Actually, two ways, when I think about it—maybe that's the other glimmer. Everybody was disappointed, of course, but the thing I found interesting is that two of them—you remember Nancy and Doug from the B school sixth grade?

Sizer: Right—the two who were always asking righteous questions about our evaluation design.

Raven: Those two—anyway, they came to me at the end of the day and said they had been talking about it and maybe there was a way the current Jigsaw teacher group could get together and do some in-house training next year.

Compass: That is interesting.

Strait: What's so disappointing to me is somehow just as the evaluation seemed to be actually helping increase interest, the rug gets pulled out.

Sizer: I know. I was thinking the same thing, but maybe it's not a complete loss. The two of you should be thinking about how to build on what the president and teachers said.

Raven: Believe me, I am. I'm getting all the Jigsaw teachers together next week to talk about it after I've had—and they've had—a little more time to think about it.

Sizer: Yes, I want to think about it, too. Look, I'd like to talk some more with you in a day or so, but Allen and I have a meeting this afternoon we have to prepare for. Could we get back to you?

Strait: Sure. Ah, there was one other thing. When the president called, she mentioned that she didn't quite understand one of the analyses in the report. At the time, I was too hot to even focus much on what she was saying, but suggested she could give one of you a call about it.

Compass: Oh, maybe she was really interested—maybe we could interest her in our coming in for another evaluation.

Sizer: Allen! One step at a time. We and they've both got a lot to consider. If she calls, she calls, but let's sort of let it sit for a few days.

Strait: Right at the moment, if you mention the word "evaluation" to me, I'm likely to see red.

Raven: Evaluation?! Nevermore—

Strait: Pam, please—I thought we agreed you'd stay off that pun—

Raven: Oh, sorry. It seemed just right. Anyway, we'll talk to you again in a few days.

Sizer: Right, say Thursday.

Strait: OK, we'll call in the morning.

Compass: See you then—bye.

Critical Moments in a Media Campaign Evaluation

This vignette illustrates a number of problems that program decisionmakers and evaluators encounter in the usual process of program development and evaluation. Every problem that arises in the unfolding of this drama is shared, although both primary actors see each problem as their own. Even more, many of the problems are seen by each as being caused by the other.

As the drama starts, the immediate problem is a time constraint, caused by a change in the theme of the prevention media campaign. But time is the fundamental resource, and its limits increase the awareness of conflicting and unclear goals.

Characters:

- o Beverly LeBeau. The young founder of LeBeau Associates, a media production company specializing in public service mass media campaigns, and project director on the State-funded media campaign for Project Straightalk, a new, three-year alcohol abuse prevention demonstration project.
- o Walter Stauback. A program evaluation specialist and project director on the separate State contract to conduct a "third-party" outcome and impact evaluation of Project Straightalk.
- o Alice Stauback. Walter's wife.

Beverly LeBeau walked into the staff lounge and flopped into the armchair, saying to two of her key people, "He didn't look too happy, but I'm going to meet with him again tomorrow."

Beverly is director of Project Straightalk, the new, highly publicized mass media campaign to prevent alcohol abuse by teenagers. Beverly has already produced four public service media campaigns, two of them on drug abuse prevention. She knows how to deal with the many people who can help or hurt a project like Straightalk. She knows how to manage tight production schedules and budgets. And she designs effective media products—creative, hard-hitting spots that grab the audience and deftly deliver the message. Beverly strives to meet the commercial advertisers on their own ground, with high-quality production values and messages that speak to people.

Beverly also prides herself on being a realist. She is resigned to the fact that public service money comes with many strings and that a big part of her job is keeping her projects from becoming entangled. Straightalk is State-funded through a contract between the State Alcohol and Drug Abuse Agency and her "media production shop." The contract requires that she deal every day with bureaucrats, advisory groups, evaluators, and other pains-in-the-neck. But knowing there are no "free lunches" in the public sector, Beverly is usually able to stay philosophical. Sometimes on a particularly frustrating day, she fantasizes about Michael Anthony appearing at her door with a seven-figure check and saying, "Beverly, just go do it the way you know it should be done." However, Beverly knows that the work and the shackles are inseparable.

Today promised to be one of those bad days. Beverly was not looking forward to her first major meeting with the "outside" evaluator, Walter Stauback, since she and her staff had decided to change the campaign theme. Like Beverly, Stauback had written a proposal in response to a State request for proposal and had won the evaluation contract. That contract was huge, almost half the size of the 3-year, $950,000, media contract. Because of its size, Beverly knew that the State was serious about the evaluation.

Five months have passed since both contracts were awarded, and for different reasons both Lebeau and Stauback have been under stress during those months. Beverly has felt the pressure to firm up the campaign theme so that scriptwriting and production can proceed on schedule. This means constant coordination of the creative process with market research and the project's advisory board. The original theme, the one that had been presented in the proposal and had won Beverly the contract, bombed in the early research. Small groups of carefully selected teenage volunteers had been brought together to discuss the theme "Alcohol is a drug!" and to see rough storyboards of television spots based on the theme. Beverly had developed the theme after reading surveys showing that many young people regard alcohol, and beer in particular, as a natural, innocuous, and harmless way to get high. "What's the problem? It's only a beer," was the attitude suggested by the survey data. In the proposal, Beverly had written, "Beer is regarded as the psychoactive equivalent of a soft drink by a sizable proportion of American youth."

The young volunteers in the discussion groups, called focus groups, yawned at both the theme and the storyboards. Instead of responding, "I didn't realize that!", the teenagers reacted with "Of course," or "So

what?" The beer drinkers in the groups, even those least experienced, just didn't believe that they were taking any serious risks. Their own experience had convinced them that they could drink without encountering trouble. And the nondrinkers, what few there were, already regarded alcohol as a drug; "I don't need a crutch to have a good time" was their most typical response. None of the kids seemed to think that Beverly's theme would change anything or anyone.

The State's reviewers, and later, the project's advisory panel has endorsed the campaign idea. But the kids had not, and it was the kids who counted. Beverly had not been too upset because the theme had subsequently proved barren for developing a good campaign. So Beverly and her staff had closeted themselves for 2 days and emerged with a new idea. There was no time or money to test the new theme as the old one had been tested, but Beverly had learned a lot about kids from the earlier focus groups and she was absolutely convinced that the new theme would work. Besides, she and the staff had hit upon a tremendously exciting format for the TV spots, one that would deliver the message with great visual power.

Beverly's project officer at the State Alcohol and Drug Abuse Agency, Molly Sorensen, hadn't been enthusiastic about the revisions; she wasn't sure that all of the project's goals would be directly addressed by the new theme. Beverly persuaded her to approve the changes by pointing out that the project's timetable would have to be revised if further delays were encountered. Since the beginning, Molly had emphasized that the project must produce all the deliverables on schedule. Beverly was even able to persuade Molly that another meeting of the advisory panel would be an unnecessary delay; the advisor's reactions to the revisions could be more quickly and efficiently gathered via the mail.

Only when she had gotten Molly's approval in writing did Beverly call Walter Stauback to tell him that the campaign's theme had been revised. Walter reacted with understandable anger—he had spent many hours with Beverly clarifying project objectives, monitoring the development of scripts, and discussing the evaluation plans to make sure they would be responsive. Walter was also under the gun. He wanted the pretest questionnaire to focus upon the campaign strategy, so a good deal of his work thus far might need redoing. But questionnaires had to be delivered to the survey firm within 10 days. The pretest survey was scheduled to begin in 6 weeks in both the nearby experimental city and the highly similar comparison city on the other side of the State.

As a gesture of good will, Beverly had offered to drive the 20 miles to Walter's office to explain the changes and to help determine their implications for the evaluation.

A few minutes into their meeting, Beverly realized that Stauback was threatening her stereotype of evaluators. Even under the strained circumstances, Stauback laughed occasionally. He spoke English and not just "Research." He was trying his best to understand Beverly's new ideas about the campaign. To her surprise, Beverly found herself enjoying the conversation.

Stauback: Let me see if I've got this straight. You're saying that now you want to put across the message that "Alcohol is for losers. The only way to be a winner is by working for it."

LeBeau: That's the basic idea. It's time to stop dancing around the critical point. In the long run, the only way to really feel good about yourself and to succeed is to work hard at the things that are important to you. Maybe some people will say it's puritanical, but it's true. One of the hidden dangers of regular drinking is that it causes kids to waste time they could be spending stretching themselves in some way. It also undermines their ability to push themselves. And too many kids rationalize that beer is OK, thinking it only has a "little" alcohol.

Stauback: So you're primarily looking to change kids' attitudes toward beer, especially their perceptions of the costs of using it—costs to their character and competence, not physiological or legal costs. At the same time, you want them to get the idea that personal success and satisfaction come only through hard work.

LeBeau: That's right. The message has two components. If possible, we won't just be telling them, we'll also be showing them. There's not much variation in the way kids get loaded, but hard work comes in many forms. Athletics, arts, scholarships, business—there are plenty of paths for kids to take. Showing how a kid can work hard in one of these directions will be the positive side of each script. Contrasting hard working kids with kids drinking beer—cutting back and forth between the two—pits the positive against the negative. In each spot, hard-working kids grow, sweat, hit and miss, progress, achieve something and feel good, while the drinkers continue to cruise, listening to music or playing the arcades, complacent, stagnant, falling behind.

Stauback: You can show that in a 30-second spot?

LeBeau: I think so. It'll be tricky and tight, but I think so. We can do it with the TV spots, not the radio spots. Radio requires a different approach—same message, but we will have to tell it rather than show it.

Stauback: What about the other objectives? What about knowledge gains? And which behavior changes are you looking for now—reductions in first-time use, in experimental use, or in regular use?

LeBeau: I guess we'll have the biggest effect on abstainers or kids who have just started drinking. We've read the research articles you gave us showing that kids who already drink regularly aren't influenced by mass media.

Actually, Walter's questions about other project objectives had surprised Beverly a bit, so she was pleased that she remembered the research studies and appeared to take the question in stride. The truth was that for several days Beverly had not been thinking at all about "changing behavior," or increasing "knowledge," or about anything except the new campaign idea and how to effectively translate it into TV spots. Walter's questions about objectives had reminded Beverly of the terms of her contract with the State, which specified that the media campaign was to "increase specific knowledge of the pernicious effects of alcohol use, promote greater understanding of the risks and thereby reduce the abuse of alcohol by young people ages 12 to 18."

Beverly wondered for a moment whether she could be criticized for ignoring the contractual objectives. Legally she was covered—she had Molly Sorensen's formal signoff and had effectively neutralized the advisory panel—yet she still felt a twinge of anxiety that perhaps she had neglected or overlooked something truly important. But there simply wasn't time for indecisiveness or backtracking, and the new spots were going to be the best she had ever done.

A half-hour later, Walter Stauback decided to cut the meeting short and schedule another one with LeBeau for the next day. Walter was upset and he needed time to think. With great enthusiasm, LeBeau had described in detail the scripts for four different spots. LeBeau was a gifted storyteller and Walter had appreciated the visual and dramatic impact of each script. However, LeBeau's impressive presentation did not alleviate Walter's increasing concern; rather, it added to his worries. Walter could see that Beverly had invested much time and energy in the scripts and was firmly committed to the new concept. He could understand how the new theme might be a major improvement on the old, but from his own perspective the new theme did nothing to solve the complex, intertwined problems that plagued not only the evaluation but the entire project.

That night Walter asked his wife's advice, as he usually did when he was considering major decisions. Alice was a wonderful listener. Often she simply asked a question here or there and let Walter find his own solution.

Walter: The biggest parts of the problem are the unrealistic expectations and the lack of time. First, the State's goals for the campaign are pie-in-the-sky. Mass media campaigns do not produce major attitudinal changes, let alone behavioral changes. The State people think that changing kids' decisions about alcohol use is like changing decisions about which soap or toothpaste to buy. The media people do, too. Show the kids the spots a few times and they'll straighten right up. But decisions about whether to use alcohol are a lot more complex and hard to influence than choosing a brand of tissue. These are not superficial choices like Kleenex or Scotties; these are behaviors that depend on dozens of considerations. In the last few weeks I've reviewed nearly a dozen evaluations of public service mass media campaigns and not one found a major shift in behavior.

Alice: Have you explained this to them?

Walter: Not really. I didn't realize it until I'd read the evaluations, and I'm positive they don't want to hear the bad news. And who am I to tell them about the media or alcohol use? The media people half believe in the theory that information changes attitudes and attitudes change behavior. They also believe "Link it to sex or success and it will sell." I'm not sure what results they expect, but they certainly aren't worried whether the campaign will be successful.

As for the State people, they want to show the legislature and everyone else that they're doing their job, which means changing behavior, I guess. They seem most concerned that all the "deliverables"—the products—get produced and get produced on time.

Alice: You don't think they have <u>any</u> chance of succeeding?

Walter: It all depends on how you define success. The media people can get their message across. They can get kids to remember and understand their campaign idea if they do a good job. Maybe—maybe—they can get some attitude shifts, especially if they can keep the message in front of the kids for a long time and if they can focus it on a specific attitude. And they may be doing that with their new idea. But I wouldn't bet on any behavioral change, even if they have a huge budget for buying air time, which they don't have a hundredth of. They're spending most of their money on TV, producing 30-second spots and buying air time, but TV is a very inefficient and expensive way to reach teenagers. Teenagers watch less TV than anybody else. I think they would get a lot more for their money if they concentrated on radio, billboards, and buscards. Even school newspapers.

Alice: TV's a lot more exciting to them, I'll bet. There's one thing I don't understand. You think they are making big mistakes, but really, none of this makes your job harder, does it?

Walter: It makes my job easier. If my primary responsibility is measuring changes in general behavior and attitudes regarding alcohol, I can just go ahead and finish the pretest questionnaire and run the pretest survey without worrying too much about what their theme is or what the particular spots will be like. Measuring the general or ultimate effects, if there are going to be any, is easy. I've already gotten most of the general questionnaire items I need by pulling them from previous surveys and evaluation studies. Measuring the specific or immediate effects of the particular theme requires that I know exactly what they're going to be saying or doing, so I can include questionnaire items that show changes from pretest to posttest in kids' recall or recognition of the theme, in specific kinds of knowledge or concerns, and so on. Those items I have to write myself and try out to make sure they work.

Alice: How can you do that? You're out of time. I thought you had already finished the questionnaire.

Walter: I thought it was finished—until they changed the theme. Time is the real killer here. The media people are being forced to rush into production before they should, and I'm being forced to run the pretest prematurely. The State thinks it's protecting its investment by holding us to the timelines, but it's ensuring that the money will be squandered.

Alice: Didn't you know that the time frame would be tight before you bid on the project?

Walter: I knew it and I didn't know it. When you're writing a proposal, you tend to go along with what's demanded and to adopt the requester's perspective. You're hungry and you want to please. It's different afterwards when you have to live with the day-to-day pressure. In actuality, it's never as simple or smooth as you hope it will be beforehand.

Alice: So what are your options?

Walter: Obviously the smart choice is to stay on the sidelines and do the general outcome evaluation. That's certainly the easy thing to do. The alternative is to make trouble for everyone including myself, to tell the media people where I think we're all making mistakes and see how they respond. If they react positively, I'll do my best to focus the evaluation on their final product. But they can't afford to listen to me—and I can't afford to do anything either—unless the State backs off on the time schedule.

Alice: I have a hunch you've made up your mind already.

Staub[illegible]k: Yes. Maybe I'll open all of this up with LeBeau tomorrow.

The next afternoon, Beverly had two reactions to Walter's concerns. One was irritation. She just didn't have time to deal with this, even with the part that made some sense. But she was also surprised and impressed that Walter cared enough about the project to have wrestled with these issues so seriously.

LeBeau: Look. I'll be straight. I think you've got some good points, but that you're way offbase on some others. But really that's irrelevant. We just don't have time to redesign anything. And you don't, either.

Staubach: You're right, unless we can renegotiate the schedule and the deliverables. We can go to Molly together if we want to. What have we got to lose?

LeBeau: A lot. For one thing, the time you and I take discussing all of this, and for another, the time we spend talking to Molly. Not to mention the dues we'll pay one way or the other over the next two and a half years for scaring her and helping her to see that things are more screwed up than she realized.

Staubach: Maybe so. Maybe so. At least tell me your reactions to what I've said.

LeBeau: OK. I'll make it short and sweet, and we can go from there.

One. You're obviously right about the time crunch. I need the extra time just as badly as you do. You're absolutely right.

Two. I guess I don't really believe that this project will produce behavioral change by itself, but I do think it will change attitudes and awareness. And that's a significant result in my book. Even if the campaign affects only one or two of dozens of factors, maybe that's worthwhile. If kids clearly or more deeply understand the risks of drinking—that's important. It may not pay off behaviorally in the short run, but maybe in the long run it will. Kids don't really understand the type of risk we're focusing on.

Three. I don't think we are relying too heavily on TV, although I'll admit that TV's where the professional payoff is greatest for us media types. Remember, we can aim the spots at who we want by putting them into the right shows and time slots. We're buying air time, not asking the stations to give us public service time. You know, the 6:00 a.m. and 2:00 a.m. time slots. We'll buy time on the programs that give us greatest "reach and frequency," which means the greatest number of exposures to the spots by the greatest number of kids for each dollar we spend.

There's another point that you've got to understand about TV. We want people other than the kids themselves to see the spots. We want parents, older brothers and sisters, teachers, you name it, to see the spots. We want the message talked about in the home, in school, wherever, and we want it understood by everyone—so that it will be supported from all sides. TV is the way to get people talking about something like this, because it is the mass medium. If we're lucky, and if we handle this right, the TV exposure will stimulate some newspaper and magazine coverage, maybe even some TV news coverage—publicity that will be priceless for spreading and supporting the message. So don't sell us short. TV is the way to make a lot of things happen.

Four. I do want you to do the specific evaluation. We need that level of precision to know what really happens. It makes me a little nervous, but I'm deeply curious to know how much we really get across to kids. I sure don't want to put all our eggs in the behavioral-change basket.

Staubach: I need more time if I'm going to do a specific evaluation. I'll need to know precisely what you're going to be doing all along. You'll have me looking over your shoulder for 2 more years.

LeBeau: I understand. That's OK with me. And I know that I'll have to delay the start of the campaign so that you can finish the pretesting first.

Staubach: Let's go see Molly.

LeBeau: Let's go see Molly.

Working as a team, Beverly and Walter were able to renegotiate the time frame for the project. Their success came not so much from the astuteness of their reasoning as from convincing the State staff that a specific evaluation would be in their interest as well. After all, their agency's reputation would not be enhanced by a general evaluation that showed no effects. With measures of specific outcomes added, the evaluation was much more likely to supply some sort of evidence that could be used to justify the State's investment. Of course, Beverly and Walter's cause was also aided by the fact that they were not asking for more money or a reduced workload, just for a revised schedule.

Beverly and Walter came to understand and trust each other more as they continued to work closely and talk honestly about their ideas and concerns. Problems arose often, but most could be handled to their satisfaction. And their growing respect for each other helped them accept the occasional sacrifices each had to make for the other.

CHAPTER 7: POLITICS AND SCIENCE IN PREVENTION PROGRAMING

(What Really Goes On . . . .Outside)

Evaluation of social programing, like the programing itself, does not exist in a political vacuum. To the other elements defining the context of social programing—the source of funds, the organizational foundations of the program, the constituency created by the program, and its social setting—evaluations introduce their own political necessities.

Evaluation has always been part of the learning process by which social organizations profit from lessons of the past and evolve into stronger, more effective institutions. Anthropologically, the strongest motivation of all social organizations has been self-preservation, and those that have survived over long periods of time have learned their lessons well.

Today, it is difficult to think of evaluation simply as a natural learning process. Beginning with Suchman's classic text (1962) and building on a historical foundation of educational evaluation, evaluation research as we know it today has emerged as a new discipline, blending knowledge of economics, operations research, and almost every aspect of the social and psychological sciences. Concomitant with this evolution have been the wide-ranging social programs launched by the Great Society legislation of the middle 1960's, which called for evaluation at every level of planning and programing. This recent history has cast evaluation into a special light, sensitizing evaluators and program personnel to the political implications of evaluation. It has become such a specialized dimension of social programing that one can lose sight of its role as the basic learning which accompanies all healthy programing, whether special evaluation research studies are funded or not (see Bittner 1972, for further discussion of this point).

This volume, as well as this chapter, focuses on the interpretation of evaluation as a formal study, rather than as a naturally occurring tool for learning. Of course, the formal evaluation study should also help those managing a prevention program to learn and to make that program more effective.

The strongest political aspect of an evaluation study is its potential threat to the survival of the evaluated program. In times when funds for even basic social services--education, health care, and public safety—are in short supply, the threat to funds for recently conceived social services such as drug abuse prevention is even greater. In a political climate when every competing program is being carefully scrutinized, negative findings in an evaluation report can endanger a program's very survival.

But even though programs and the funding agencies must continue to rely on evaluations to learn how well the prevention programs are performing, neither the evaluators nor the programs themselves need be helpless victims of circumstances. The central question addressed by this chapter, therefore, is how, in an increasingly changing political and economic context, one can have sound evaluation that supports the growth of alcohol and drug abuse prevention programs and that helps them survive and improve rather than provide ammunition for their opponents.

One approach to this issue, in harmony with the messages of preceding chapters is presented below.

- First, it is important to understand in advance the political problems associated with the evaluation of alcohol and drug abuse prevention programs.

- Second, it is important for the program manager and the evaluator to arrive at an open, shared understanding of their personal and professional goals for the evaluation so that it can be accomplished in an atmosphere of mutual trust.

- Third, it is important to develop a comprehensive plan before the evaluation starts. A critical element of that plan concerns how the political implications of the evaluation research are to be addressed—spelling out the complementary roles, in this regard, of the evaluator and the program manager.

- Fourth, throughout the evaluation the evaluator and the manager maintain a close working relationship, so that they can solve, to their mutual satisfaction, the political issues which are likely to arise at each stage of the evaluation.

- Finally, to the extent possible, all other significant decisionmakers outside the program should also be included in this process. Advanced planning is essential, but it can only go so far in anticipating the manner in which these political forces actually develop around an evaluation. Real effectiveness in dealing with these issues must arise from continual interaction with external powers, which initial understanding and planning can do much to assure.

Another purpose of this chapter is to show how to present evaluation data, results of which are almost always ambiguous. That is, data seldom point to a prevention program either as a resounding success or as an abject failure. Usually, they point up strengths and weaknesses in a complex fabric of findings and interpretations. The limitations, seen in proper light, provide opportunities for improvement; and the strengths highlight the achievements that the program has already accomplished.

The manager and staff of a program can be expected to examine findings which point in a variety of directions and discover the lessons that can be learned. But persons outside of a program are less likely to ponder a complex pattern. The news media especially like to have their stories etched in black and white. Therefore, this chapter suggests ways in which managers and evaluators can present complex, ambiguous evaluation results simply, in a manner that benefits the program and satisfies the need of more remote audiences.

It is assumed that the evaluator has undertaken to assess program effectiveness within a framework that the program itself defines—that is, in terms of the program's goals. Ideally, the evaluator is detached, and willing to give the program a fair test of its effectiveness. But the tacit (sometimes explicit) understanding is that the evaluation will accept the goals as the program defines them and, in terms of the underlying theory of alcohol and drug abuse prevention, will relate those goals to the problems of the participants. As Carol Weiss has stated in generic terms (1975, p. 19):

> First, evaluation research asks the question: How effective is the program in meeting its goals? Thus, it accepts the desirability of achieving those goals. By testing the effectiveness of the program against the goal criteria, it not only accepts the rightness of the goals, it also tends to accept the premises underlying the program. There is an implicit assumption that this type of program strategy is a reasonable way to deal with the problem, that there is justification for the social diagnosis and prescription that the program represents. Further, it assumes that the program has a realistic chance of reaching the goals, or else the study would be a frittering away of time, energy, and talent. These are political statements with a status quo cast.

This initial willingness to see the world as the program sees it, at least provisionally, is a major political stance that most evaluators take when they do an evaluation. This stance must go even a step further; namely, evaluators should be committed to seeing the results of their work used to strengthen the program whenever possible. This commitment is the foundation of the mutual trust and understanding that are essential if evaluator and manager are to work together with external forces to deal with the many issues surrounding an evaluation.

The remainder of this chapter is organized into five sections:

- Four Case Studies
- Issues Relating to Values
- Issues Relating to Evaluation Design
- Issues Relating to the Presentation of Findings
- Concluding Guidelines.

For several reasons, the chapter focuses on outcome evaluation, with only occasional references to process and impact evaluation. Most external political issues arise from outcome evaluation, primarily because it is the type with which non-evaluators are most familiar and for which they have the clearest

expectations. Process evaluation results are typically used within the program context, and impact evaluation results have the same external political ramifications as outcome results.

FOUR CASE STUDIES

The issues raised later are illustrated here with examples drawn from the evaluations of four prevention programs conducted by the author or his associates. Obviously, these case studies do not reflect the full scope of prevention programing. All involved programs were designed to prevent drug abuse in youth, adolescents, and young adults. A great deal of contemporary drug and alcohol abuse prevention programing focuses on other special populations.

Because of the sensitive nature of the issues being discussed, the four case studies are anonymous. All identifiers have been changed, and some fictional illustrations have been added.

Project Commune

Project Commune was an early intervention project, providing individual counseling, a limited amount of group counseling, and referrals to other programs for specialized help. It served high school students and young adults who were experimenting with drugs and were self-motivated or were encouraged by their families, teachers, or friends to seek help before more serious drug use caused real harm. The setting was a suburban university town, Los Verdes, Arizona, providing the program with a white, middle-class clientele. The most interesting feature of the program was that it was based on Maoist philosophy and was run by a collective of seven female managers, the "Committee", no one of whom was officially more in charge than any of the others. The principal evaluator was a male, and both outcome and process were evaluated.

The Chinese Youth Club (CYC)

The Chinese Youth Club was a storefront program located in the Chinatown area of Big City, California. It served a population of secondary school students who had recently immigrated to Big City from Hong Kong, Southeast Asia, and mainland China. The program used the facilities of neighboring schools and provided tutoring, Chinese arts, sports programs, and individual and group counseling to the students and their families. The students lived in an inner-city community characterized by a considerable amount of drug use, drug dealing, and gang membership on the part of Chinese youth and others. The program's clientele did not have a history of any drug use on entering the program. The program was evaluated from both process and outcome perspectives. The program manager was Sue and the evaluator was Elliot.

The Mexican-American Youth Alliance (MAYA)

MAYA was a prevention outgrowth of a community-based heroin treatment program. After a number of years of providing effective treatment of addicts in this Mexican-American community, the members of the community sought to prevent the development of heroin addiction by working with secondary school youth. They provided a Chicano prevention counselor in the three junior high schools and the one senior high school that served this inner-city Chicano community in Central City, Texas. The prevention workers conducted values clarification sessions in social studies classes, provided individual counseling during the day, and conducted a cultural club for Chicano youth after school which included sports, arts and crafts, outings, and group counseling. Maria was the program manager and Thomas was the evaluator. Process and outcome evaluations were undertaken.

The New Life School

The New Life School served Saddle Creek, New Jersey, a large bedroom community of a major eastern city. Like Project Commune, it was an early intervention program, helping secondary school youth who had begun to experiment with drug use. It provided an alternative school setting, which was strictly enforced as drug free, and in which students could reestablish their commitment to doing well in school. It also provided counseling groups for parents. The clientele were black and white middle-class students. They spent a year away from home in this specialized school to prevent limited experimentation with drugs and alcohol from

blossoming into a full-blown drug-oriented lifestyle. The school was evaluated with both process and outcome evaluations. The program manager was Sharon, and the evaluator was Michael.

ISSUES RELATING TO VALUES

The Evaluator Has Values

Although most evaluators strive to be objective, they inevitably bring their own values into the evaluation. Beware of evaluators who deny this, for they are unlikely to know their own values and, therefore, cannot take them into account in efforts to be objective.

Managers must know the evaluators' values and be able to discuss them openly and frankly. Often evaluators feel some cultural distance between themselves, the program, and its setting, even if they are from the same culture. For example, The New Life School serves a middle class suburban community on the east coast, and Michael—the evaluator—grew up in a suburban middle class community in the midwest. Not only are the two communities geographically different, but also youth culture has undergone a dramatic transformation in 20 years. In addition, because the program manager and staff averaged about 10 years younger than Michael, he felt out-of-tune to some degree with the staff, and even more so with the students.

The cultural distance becomes much greater when the manager, the staff, and the clients come from a cultural background distinctly different from that of the evaluation team. Consider the Chinese Youth Club in which all staff and clients were recent immigrants to the United States—all within the past 12 years, many having been in the United States less than a year. The evaluator, Elliott, on the other hand, grew up in a small, rural university town in Northern California. His family background was white and middle class, as was most of his hometown.

Most of the CYC staff and about one-third of the students came from Hong Kong. Until the normalization of relations with China and the lifting of immigration restrictions, the majority of the Chinese immigrants to Big City came from Hong Kong. But since the political shift, nearly three-quarters of the immigration to Big City is from the mainland. The Hong Kong Chinese speak English well and are comfortable dealing with occidentals. In contrast, mainland immigrants usually have no knowledge of English and are more timid with occidentals, at least until they become familiar with the language and the culture.

Through his upbringing and his own tastes, Elliot had developed an affinity for Chinese culture and, therefore, felt comfortable working with Sue and her staff. He probably would not have felt as comfortable had the program been staffed by Chinese from the mainland. As a result, he was inclined to be favorable towards the program, a bias that was nonthreatening to Sue and the CYC.

On the other hand, Elliot's research assistant—Robert—was an immigrant Chinese working on his doctorate at Big City University. He was inclined to be critical of the way the CYC operated, and would have liked a more professional staff, with advanced degrees in counseling or education. Although Elliot recognized these feelings in Robert, he did not feel that he knew him well enough to discuss them. Sue and the CYC staff seemed confident that the tone of the final report would be in Elliot's hands, and that he would filter out excessive negativism on Robert's part.

A program with a strong political orientation cannot ordinarily find an evaluator with a shared outlook; it can, therefore, expect to feel some discomfort with almost any evaluator.

Mutual openness is important with respect to this first issue. In instances where the manager selects or participates in the selection of the evaluator, the manager should request that the evaluator identify those values relevant to the evaluation, especially any that relate to the program's goals, methods, and cultural background. If a candidate evaluator seems unwilling to be frank, seems uncommunicative, or expresses values that make the manager uncomfortable, rejecting the candidate might be wise.

Time and resources probably do not permit an exploration of the values of all members of an evaluation team. Normally, however, because the principal evaluator will have the greatest impact on the evaluation and on the manner in which results are presented, understanding that person's values is normally sufficient.

One actual instance illustrates how disastrous the consequences can be of failing to recognize a bias. Two principal investigators were awarded a grant to evaluate a national, multi-site program for juvenile

delinquents. These investigators held strong personal theories of delinquency and privately expressed their hope that these programs would turn out to be failures. They therefore undertook this evaluation to "prove" the programs ineffective. The results confirmed their expectations; the published outcome was exactly what they had wanted.

The phenomenon of researchers' finding what they are looking for is not always so blatant. Even when evaluators have only a latent belief about how things should turn out, the results will quite likely support the validity of that belief. Citing excellent psychological research demonstrating the frequency of this phenomenon, Martin Orne has labeled it the "demand variable" (Orne 1962; Orne and Evans 1965). To the extent that managers can control the situation, they must ensure that no "demand variable" exists to cloud the evaluation results.

And the Program Has Values Too

Of course, an effective collaborative relationship requires openness on the part of the manager as well as the evaluator, although the two parties need not share the same or similar values. What is necessary is that they understand each other's values and that the values of neither party work against a reasonable evaluation. Often the evaluator and the manager have strikingly different values, but both parties have agreed to respect their differences as best they can.

Project Commune provides a striking illustration. In this rare instance a drug abuse prevention program founded on a Maoist feminist philosophy was funded by a State criminal justice planning agency. The grant required that the program secure an objective outside evaluation. The seven managers approached a friend at a local university, who helped them find an evaluator, George, who then hired a small staff and designed a process and outcome evaluation study for Project Commune.

It is inherently problematic to deal with more than one manager. In this case there were seven, all nominally equal to each other—a structure which George had to respect. However, the situation was made somewhat easier because the managers' deeply held extreme political views were remarkably similar, obviating much of the internal value conflicts which might ordinarily have been expected.

George was at the time a rather liberal Democrat, but from the perspective of a Maoist, his position was not much different from an extreme right-wing Republican. So from the start, all accepted the gulf separating their outlooks and values. To work together, they negotiated a compromise around the distinction between process and outcome evaluation. The process evaluator would, of necessity, have to get close to the program, whereas the persons collecting the outcome data needed to maintain their objectivity and did not need to "infiltrate" the program. George, in conjunction with the Committee, selected a woman graduate student in sociology at the local university to work half-time as the process evaluator since only another woman could probably have secured the trust of the Committee and the staff. Although not a radical, the woman had strong liberal views, and was regarded by the Committee as co-optable. In fact, to some extent, she was co-opted as the study progressed, casting some doubt on her objectivity. However, given the political nature of this program, the selection of a woman may have been a necessary condition for process evaluation data to have been collected at all.

This illustration provides a clear example of how an evaluator and a group of managers solved a difficult situation of dissimilar value orientations and were able to carry out an effective evaluation. Mutual respect for each other's values, formed during an initial collaboration, made it possible for the two parties to work together throughout the evaluation. In general, the degree to which the evaluator and manager can understand and respect each other's values, the more likely they are to sustain mutual trust throughout the evaluation. Mutual trust is essential for working through thorny political problems that typically beset the presentation of evaluation findings for a program in the public eye. Thus, establishing reciprocal understanding and trust is a critical first step in dealing with the politics of evaluation.

The Community and the Political Leadership May be Watching

Prevention programs operate in a context of community values, of significant bureaucrats, and of political leaders. This larger, external context is usually foremost in people's minds when they think about the politics of evaluation. The values internal to the program and to the evaluation interact with these external values in the resolution of the evaluation's political issues.

The evaluation of the MAYA prevention program illustrates issues associated with a concerned community. In this instance, the Chicano community, with serious heroin addiction problems, had been neglected by city agencies. A politically aware and creative group of young men and women conceived the idea of getting a grant to set up a heroin treatment program. They were successful, and the MAYA program came into being. The founders, however, were not good administrators, and the requirements of the State funding agency forced them to hire a professional administrator, Maria, who came from a Chicano drug abuse treatment program in Big City, California. Soon after her arrival at MAYA, Maria applied for a prevention grant.

The community was uneasy. It did not want to relinquish control of program administration to a professional and an outsider. The second grant, the prevention grant, also affected the operation of the agency, including the requirement to let a substantial contract for an evaluation. In time, community members on the board of directors were replaced by members from some of the agencies that MAYA dealt with, including a deputy superintendent of schools, a probation officer, and a member of the sheriff's department, all of them Anglos. Gradually, Maria felt constrained to act as a bridge between two cultures with little mutual understanding—the local Chicano community and the Anglo, middle-class bureaucracy that provided the funding. In many instances, it seemed as though actions that pleased one constituency only upset and confused the other.

Thomas, the evaluator, felt at once beset by this strain and mistrust when he arrived to evaluate the MAYA prevention project. To make matters worse, because of its distrust of Maria's commitment to evaluation, the State funding agency had specifically selected Thomas as an evaluator. But Thomas and his staff were Anglos, only one of whom had experience dealing with Chicanos and could speak a little of the local Spanish dialect.

On the positive side, Thomas and Maria soon realized that his presence and Anglo background could help give the prevention component of MAYA credibility with the Anglo funding source. The community, however, was anxious that the Anglo influence and the professional character of Maria, her staff, and half of the board of directors not undermine MAYA's focus on Chicano concerns, values, and culture. These were the shared concerns of Maria and Thomas as they mapped out the evaluation.

Whereas the MAYA program needed to work within the concerns of the local community, the New Life School focused on the politics of the school system and the board of education. The New Life School had been founded--over the superintendent's objection that the school system was doing all that was required--because of the personal commitment of two board members. Once established it also had strong support from the Assistant Superintendent for Alternative Schools, under whose authority the program fell.

The evaluation was planned and undertaken by the Division for Program Assessment, who hired Michael, an outsider evaluator, to evaluate the prevention school. Michael and his staff were hired by a competitive procurement conducted by the division. The New Life School had been underway for a year when the superintendent's office decided to have it evaluated, with the expectation that the findings would be available to the board of education in time to consider the school's refunding.

Michael first encountered Sharon, the manager and the principal of the New Life School at a meeting in the office of the Assistant Superintendent for Alternative Schools (Sharon's boss and mentor). The meeting also included the director of the Division for Program Assessment, thus creating the potential for conflict between program administration and evaluation. At the time of this first meeting, Michael was fairly new to the scene and only slightly aware of the political history of the school. He did feel that the meeting was strained, but could not immediately understand the source of the conflict.

After a little investigation, and development of a closer collaboration with Sharon, Michael began to sort out the nature of the political pressures. It seemed clear that the "pro-school party" consisting of several board members and the assistant superintendent, were looking for a favorable evaluation. The staff of the Division of Program Assessment were neutral, and wanted only to see the evaluation carried out professionally. Although the superintendent and a few asssociates were probably slightly hostile to the program because of the manner in which the board had pushed it on them, their negative feelings did not seem very strong, and they were willing to support the program if the board continued to want it.

The case of the New Life School is typical of many instances in which a prevention program has drawn considerable attention to itself at the time of its founding, resulting in some polarization of key political forces. At the same time, most political situations are complex. It is often most clear who the committed supporters are. Other key actors, often neither for nor against the program, may be somewhat threatening because they cannot be relied on to support the program if findings are not favorable. Usually there is also a third camp, which continues to bear a grudge against the program. These individuals do not necessarily lean on the evaluation for negative conclusions, but they would probably be pleased at such an outcome.

Such forces need to be understood and sorted out before an evaluation can be undertaken since they will come into play when a report is released.

The World of Macro-Politics

Macro-politics may affect any social field, but at times changes at this level are especially radical. The budget cuts for social programing now in effect could alter the very structure of prevention programing. Major support responsibility has now devolved upon the States, a few of which are enjoying exceptional wealth because of fuel severance taxes while most are facing serious fiscal problems. The resulting picture, especially in the poorer States, is one in which drug abuse prevention programs must compete for limited Federal, State, and local tax dollars with a wide range of health programs, most of which have strong medical and consumer constituencies. In such a climate, prevention programs need extraordinary support to maintain and expand their funding base. History has shown over the past two decades that favorable evaluation results are seldom, if ever, a deciding factor in such debates. But favorable evaluation results can be added to other kinds of supporting information to build a more compelling case for the continued support of prevention programing. In this context, sensitivity to the larger political picture takes on an unusual degree of importance for evaluations.

ISSUES RELATING TO EVALUATION DESIGN

Specific versus Generic Prevention

Anyone in the prevention field comes to realize that the categorial boundaries by which Government agencies address the world of education, health, and human services often make it difficult to encompass real world problems. Prevention of alcohol and drug abuse provides an especially poignant example of how the "official" versions of the world differ dramatically from the experience of programs dealing with prevention "on the street."

Preventing behaviors destructive to the individual's health and well-being, and potentially destructive to others, of which drug abuse prevention is just one aspect, is by its nature a unified generic problem. Evidence from a number of research studies suggests that among adolescents, alcohol and other drug abuse are associated with each other and with delinquency, teenage pregnancy, problems of family life, and poor school performance (Jessor 1979). Problems demanding prevention initiatives are found among young adults, the middle-aged, and senior citizens, each with their own peculiar generic mix. A look at the Federal bureaucracy reveals that intrinsically related prevention activities have been funded by the Alcohol, Drug Abuse, and Mental Health Administration (ADAMHA); by other agencies of the Department of Health and Human Services (DHHS) concerned with aging; by the Department of Justice; and by the Department of Education. Several other Federal and State offices, agencies, and institutions have funded research and demonstration projects relating to one aspect of prevention or another.

In this context, local programs have at times shifted their emphasis from one dimension of prevention to another, shifting, for example, from drug abuse to delinquency prevention and doing a credible job of both. Some progress has been made linking prevention efforts involving drug and alcohol abuse at the Federal, State and local levels.

Program managers generally recognize that their prevention efforts, in most instances, have generic impacts broader than alcohol and drug abuse prevention alone. Program effects across the range of destructive behavior depend on the nature of the prevention modality and the risks associated with a particular population being served. In addition to drug abuse in our four case studies, the risks of destructive behavior include alcohol abuse, delinquency, and failure in school.

The model of drug abuse onset and other destructive behaviors proposed by the Jessors (see, for example, Jessor 1975; Jessor and Jessor 1975a; Jessor and Jessor 1975b) suggests that changes in destructive behavior form a predictable pattern. Thus, a genuine change in an adolescent's lifestyle away from drug abuse would probably be accompanied by changes in other aspects of life such as school attendance, academic performance, and the tendency to commit delinquent acts and status offenses and other disruptive behavior. This model, therefore, justifies a program's efforts to correct behavior more generically, rather than to focus simply on drug and alcohol abuse. For certain preventive strategies, therefore, it may be important to collect clusters of appropriate prevention outcome data to understand the degree that prevention efforts result in broadly based life changes.

In three of our four cases, additional data were collected on delinquent and acting out behavior (CYC, MAYA, and the New Life School). In two of these cases (CYC and New Life), information was collected on aspirations toward the future (another dimension of the Jessor model); and for the New Life School, detailed information was also collected on school attendance and academic performance. In all instances the kinds of clustering of outcomes that one would expect from the Jessor's model were noted.

For a program with high public visibility, the collection of a wide range of outcomes may be advisable. The ability to demonstrate outcomes in a number of areas of public concern may be helpful in developing a broad-based constituency and in selling the program for future funding. The selection of outcome measures may have significant political overtones and should be a collaborative effort of the evaluator, the program manager, and other decisionmakers.

Control Over the Evaluation Report

Evaluators, in general, are rewarded, in part, by having their work read, used, and appreciated. A spectre that hangs over the evaluation field is that the commissioning agency might suppress the report and prevent the evaluator from making the findings public. Such suppression may be reinforced by highly restrictive language in the evaluation study contract which gives the contracting agency complete control over the findings and any reports produced. However, once word gets out that an agency has exercised such authoritarian control over a report, it may be difficult for them to contract with reputable evaluators in the future.

Understandably, of course, managers are concerned that an evaluation report will contain material that in their view is totally misleading or erroneous, and that they will not have an opportunity to detect such problems before the final version of the report is published. Or, even if managers do see a draft, they worry that evaluators will cling stubbornly to erroneous views, and that needlessly damaging or misleading reports will see the light of day, without any opportunity for the manager to express a dissenting opinion.

This problem can be avoided if, at the design stage, the evaluator and manager work out a mutually acceptable set of guidelines to govern the preparation and issuing of publications. Following is an example of the way such guidelines might be drawn up.

- The evaluator agrees to show the manager a final draft of any reports or articles which are to be published concerning the study to allow the manager to review and comment.
- The program manager agrees to review and comment on any draft materials in a timely manner and to comment frankly on the draft.
- The evaluator agrees to consider carefully the manager's comments and criticisms, to make appropriate changes in the text of the draft, and to show these changes to the manager.
- If the manager continues to have serious reservations about the contents of the draft, even after all the changes which the evaluator is willing to make have been made, these dissenting opinions may appear as an addendum to the report. If the material is to be published in a journal or book form, where there is a serious concern that misrepresentations may damage the program, the manager should have the right to insist on anonymity of the program.

Guidelines like these assure the evaluator of a right to present findings in all appropriate channels and assure the manager of means to protect the program's interests. Even when the program and the evaluator are on harmonious terms, as was the case with the CYC evaluation, such guidelines are best expressed formally.

The Selection of Goals to be Measured

Another major concern is whether the stated goals of the program are the goals actually pursued. The author once participated in an evaluation of a drug abuse treatment program in which the published goal was to help adolescents and young adults stop using drugs. Soon after beginning the evaluation, he was amazed to find himself sitting in on an employment interview in which a candidate for a staff position was being rejected in part because she did not take enough drugs. The actual goal of this program turned out to be to legitimize what the program regarded as appropriate drug use behavior in that community. Any evaluation which had judged the program in terms of its stated treatment goal would have been completely out of tune

with reality. The program would have appeared a failure to external powers, and the manager and staff would have found the evaluation totally irrelevant.

This issue also arose with respect to both the outcome and the process evaluations in the case of the New Life School. In the outcome evaluation, the program's stated goal was to help secondary school students stop using the drugs with which they were experimenting. In her review of the draft evaluation report, Sharon, the manager, stressed that the program goal was to ensure that students spend the school day in a drugfree environment, rather than to try to stop their drug use in nonschool hours. This change in the program goal had apparently occurred sometime between the proposal to the school board and inception of the evaluation study. The outcome evaluation had measured a goal that no longer applied to the program. Much effort could have been saved had the evaluator and the manager fully discussed the program's goals and objectives during the design of the evaluation.

Michael, the evaluator, partly at the request of the Director of Program Assessment, had focused a major share of the process evaluation data collection on assessment of the counseling component at New Life School. He later discovered that Sharon and her staff were not professional counselors and did not regard counseling as a primary component of the program. They were teachers and had concentrated on those elements they could best deal with, such as discipline, attendance, and academic performance.

Obviously, Michael could have been more efficient had he carefully reviewed his plans with the funding agency and Sharon before going ahead with the evaluation. Instead, his priorities were set by the funding agency representative, who wanted the New Life School evaluated in terms of its published objectives. The situation would also have been helped had Sharon reissued the statement of objectives, so that the school administrators responsible for the evaluation could understand the intent of the program.

Are the Tools of the Evaluation Appropriate?

Another technical concern with important political implications is the relationship between the evaluation methodology and the objectives of the program. In the evaluation field, certain focal areas have received the most attention in terms of measurement, instrumentation, and analysis. Three factors combine to create a dilemma in the measurement of program goals and, therefore, in the ability of the program to be evaluated:

- o Some existing instrumentation does not cover all variables of interest.
- o Some existing instrumentation may have debatable validity or reliability.
- o Rarely are evaluation resources sufficient to develop and refine new instruments based on unique program goals.

The evaluator may have to select an instrument that does not correspond exactly to program goals. This problem arose in every one of the four case studies examined in this chapter, and in two instances it had serious political ramifications.

In the CYC, a focal objective of the program was to work with the immigrant parents to help them understand their neighborhood street conditions. The Chinese parents lived in an insular world; they knew almost no English, could communicate only with other Chinese adults, and spent most of their waking hours working in factories and restaurants.

The evaluator could not locate an instrument that would assess changes the program tried to produce in parent knowledge, attitudes, and behaviors regarding child-rearing practices. The manager pressed this point because it was such a central goal of the CYC program. The failure of the evaluation study to document achievements with the parents undermined the credibility of the program with the head of the State funding agency.

In the case of New Life School, the main goal was to maintain a drug-free environment during school hours. Unfortunately, the evaluator was unaware of any instrument which measured the prevalence of drug use during a specified portion of the day, so that no attempt was made to evaluate this particular objective. Overall prevalence of drug use was assessed using a standard instrument. But the inability of the evaluation study to focus specifically on the central goal of the New Life School had a consequence—the manager felt acute political repercussions when the evaluation could not "prove" attainment of a major objective.

The manager must understand that only rarely will an outcome evaluation provide existing instrumentation tailored to the program. Therefore, managers and evaluators must assess in advance which goals and objectives the available instruments will measure accurately and which they will measure poorly or not at

all. Anticipating this imbalance, they should design the overall evaluation to minimize negative political implications by communicating evaluation constraints to external decisionmakers and negotiating mutually acceptable evaluation objectives.

Similar problems arise with respect to process and impact evaluations. A problem for process evaluations is that adequate methods are seldom available for recording the substance of the prevention modality as it is actually implemented. The political implications of instrumentation problems are usually not as far-reaching for process and impact evaluations because public administrators and the community have much less experience with these.

ISSUES RELATING TO THE PRESENTATION OF FINDINGS

Throughout the preceding discussion on politics and evaluation, reception of the final report has received emphasis, even though the issues concerned mostly predesign and design phases of the evaluation study. Usually, the politically sensitive issues of prevention programing do not come into play until the evaluation findings are reported outside program confines. In smaller studies, such as our four cases, this usually occurs after the study is completed and the final report is prepared. Larger, longer term studies may report findings from time to time throughout the course of the evaluation.

If the recommended planning occurs, and if the evaluator and manager have developed a collaborative relationship, then a strong foundation is laid for dealing with any political issues that arise when findings are presented to the community and to concerned public administrators.

The Need for a Positive Approach

Evaluation results are almost always ambiguous. (See Weiss 1975 for a fuller discussion of this point from the perspective of the evaluation of social programs in general.) In fact, evaluation results were somewhat ambiguous for our four case studies, as evidenced by one aspect from each:

- o Project Commune revealed a sharp decrease in drug use among participants who stuck with the program; however, many of those who entered the program left long before they had completed it. Those who left early showed no change in drug use.

- o CYC gave a similar picture. Recently arrived immigrant youth, especially boys, tended to begin experimenting with drugs and other forms of acting-out behavior. If they were regular CYC attendees, this experimentation was short-lived, and they continued to be essentially drug free. If, however, they left the program at or before this point, they sometimes adopted a destructive lifestyle, based on association with Chinese street gangs who both used and sold drugs—a pattern common for both boys and girls.

- o The MAYA program definitely helped boys reduce acting-out behavior. However, Chicano teenaged girls in Central City were "over controlled." The impact of such experiences as values clarification was to encourage the girls to act out more, including more experimentation with drugs—although their overall level of experimenting and of acting out was less than that of the boys, both before and after the program. Comparison group girls acted out less and took fewer drugs than did program girls; whereas comparison group boys acted out considerably more and were considerably more likely to use drugs than were program boys.

- o The New Life School finding was that program youth—based on a number of sources of evidence but not strictly on outcome data—did experience a drugfree school day. The attendance record and the quality of the school work for the program students was considerably better than those for the comparison group students. But the quantity of overall experimentation with drugs was unchanged throughout the program year for both program and comparison group students.

In all four instances, the program could be judged to make an important contribution to drug abuse prevention. However, these findings could also be presented to emphasize the aspect and to make each of the programs appear a failure. Note that in each case we are considering only one central ambiguity; other findings showed similar patterns, making a more complex tapestry than we can deal with here.

In each study, the evaluator was committed to a positive approach, trying to help the program build on its accomplishments and improve its programing. In two of the four cases, CYC and New Life School, the

program was able and willing to take advantage of the negative findings and make important course corrections in program strategy. However, Project Commune and MAYA became entangled in problems with their communities sufficiently serious to produce the demise of both programs. They never had the opportunity to try to correct deficiencies in their program strategies.

In both instances, the process evaluation tried to place the problems with the community in perspective to help the program understand and deal with them. Project Commune's managers did not take the written observations of the evaluation seriously, perhaps because of the lack of trust between the evaluator and the seven radical managers, growing out of their ideological gulf. MAYA's community problems were so far advanced by the time the evaluation was underway that a solution to the problem was probably no longer possible.

If possible, managers should select evaluators with commitment to constructive use of evaluation findings. Evaluators who approach their work primarily as "judges" and who classify programs into only two categories—successes and failures—are out of tune with the ambiguous character of most evaluation results. When such evaluators bring with them a generally negative outlook, they can be quite destructive and should be avoided.

The Presentation of Findings

Even if the evaluator and the manager are prepared to deal with ambiguous findings internally and to make them a point of departure for constructive change, presentation of ambiguous results to the funding source, to concerned public administrators, and to the community is still difficult. In all four cases, some community groups were interested in the findings, and in two of these the interest even attracted media attention. In three of the four cases a State-level funding agency was interested in the effectiveness of the program. In the fourth case, New Life School, there was an interested local funding source. In all four cases the evaluation results could affect the current funding agency's decision to continue program support. Finally, with respect to all four cases, other important public administrators were potentially interested in obtaining the evaluation findings.

One approach was tried in each case study to help clarify evaluation findings and enhance their potential for use by external forces. Summaries and presentations were prepared that minimized the complexity of the findings and presented them constructively. The case summaries were proactive, while the two kinds of presentations—to funding agencies and to public bodies—were reactive. It is always desirable for the manager and evaluator to chart a more proactive campaign to disseminate findings.

Responding To Audiences Creatively

The evaluator and the manager must be sensitive to the breadth and character of the issues of concern to a potential audience and to stress these issues in their presentations, even if those issues were less critical when the evaluation was originally designed. For example, a prevention evaluation started several years ago and only now about to present findings may not have paid much attention to cost-benefit issues. But recent dramatic reductions in Federal support to health and human services have made cost-benefit arguments crucial. Changing circumstances may require organizing even data collected for other purposes to make as compelling a case as possible. Managers and evaluators need to have considerable flexibility.

Some other ways to present evaluation findings in their broader context are to:

- o discuss the community's prevention service needs and the program's overall contributions to meeting them
- o present the findings to illustrate the human pathos of the program context
- o capture the enthusiasm that participants, their families, and interested community members may spontaneously express toward the program.

Written reports, even concise general summaries, may not be an effective way to communicate program accomplishments to members of the general community while creative use of other media can help reach a broad audience.

CYC provides an illustration of the innovative use of media for reaching the community. The agency rented the elementary school auditorium across the street for a Sunday afternoon meeting. The choice of time was critical, because a large percentage of adult men in the community worked in restaurants evenings, and many women worked in garment factories on Saturdays. Sundays were the only days during which both men and women were available for such a meeting.

The immigrant Chinese adults were too tired from working 60, 70, and more hours a week, to want to attend a meeting about CYC; but it was important, given the politics of Chinatown in Big City, to obtain the interest and support of the community. The manager hit on the idea of showing a popular Chinese movie free to the persons who attended the Sunday afternoon program. The resulting meeting was a total success. About 300 adults from the community attended. They saw the first half of the movie. Then during a break the manager and her staff presented some of the evaluation highlights in a manner interesting to the community. The evaluator was introduced to the audience, although he did not make a presentation because he did not speak Cantonese. After the half-hour of CYC presentations, the remainder of the film was shown. Afterward, refreshments were served in the school cafeteria. During the refreshment period the manager and staff mingled with the audience and discussed the program with them. As a final attraction, participants' paintings, calligraphy, and other arts and crafts were exhibited in the foyer.

Subsequent feedback indicated that the afternoon affair had made a strong positive impression. The resulting support filtered through the active Chinatown grapevine and was helpful in suppressing opposition from competing programs that regarded CYC as a threat to their sources of funding. CYC illustrates how the presentation of evaluation findings can involve creative, sensitive approaches.

Dealing with the News Media

In some instances, the program is the focus of media attention whether it wants it or not: New Life School, MAYA, and Project Commune were all sought out by the newspapers and radio and television news reporters. The CYC program, however, wished to obtain favorable coverage for itself, and sought out news coverage in the local Chinatown newspaper and the Chinese radio station in Big City.

Whether contacts with news media are reactive or proactive, keep in mind the following two considerations and deal with the media appropriately.

First of all, remember that the news media sieze upon drug abuse data. Newspaper editors like to build their headlines around such material. Almost invariably some information regarding the prevalence and incidence of drug use (and possibly of delinquency or other kinds of destructive behaviors) will appear in the report of an outcome evaluation. The media tends to blow this information out of proportion, distorting the real meaning of the findings.

To counter this tendency, the evaluator must develop approaches that play down such statistics or their uniqueness. He might mention, for example, that such levels of drug use are typical for adolescents in the area. The important thing is to anticipate a focus on drug use data, and to prepare responses designed to refocus attention by helping news people place the matter in perspective.

A second concern when dealing with the press, radio, and TV is the media's tendency to prefer simple, either-or findings. They often base a story on answers to a few questions asked in the course of a five-to-ten minute telephone conversation. This almost always results in serious oversimplification of the findings, often to the detriment of the program.

The manager and the evaluator should not allow themselves to be trapped in this no-win situation. If reporters seek information about the evaluation and/or about the program, they should insist on a face-to-face meeting in which the reporters are willing to commit at least 30 minutes of their time to talking about the program. If they have serious professional intentions, the reporters will probably agree. If not, it is safe to assume that the potential story would not have been very helpful in presenting the program to the public.

Assume that the media will be interested. Even if such interest seems unlikely at the time the evaluation is being developed, unforeseen circumstances can arise that draw the attention of the media, and put the manager and the evaluator on the spot. For example, MAYA did not expect media coverage. Central City had no Chicano-oriented news media, and Chicano programs seldom attracted the attention of the Anglo-dominated news media. Near the end of the evaluation, however, a murder occurred in the Chicano community—an organized crime assassination—and the manager was inadvertently connected with the event. Suddenly MAYA was briefly in the news. The manager and evaluator were both sensitive to the program problems that such coverage entailed. Although they had not planned how they would deal with news reporters, they held a meeting and mapped out a strategy. Their coordinated approach was effective, and they received in-depth favorable coverage from Central City's two newspapers, from a major television station, and from an important radio station.

CONCLUDING GUIDELINES

Four conclusions summarize the major points in this chapter and organize them into broad guidelines to help the evaluator and manager deal with evaluation politics:

- Political issues can subject the evaluation team and the program to considerable pressure, especially when the evaluation findings become public. To counter these pressures, the evaluator and the manager must develop a strong collaborative relationship based on trust, respect, and understanding. Such a relationship arises from an open sharing of relevant values and a joint exploration of the larger context of values in which the evaluation program is embedded.

- Evaluations tend to focus on the stated objectives of a program, using tools which are available to the evaluator. An effective evaluation, which will both strengthen the program and sustain it through political storms, is based on a sound design developed collaboratively by the evaluator and the manager; both parties must also understand the implications of the methods selected and their relationship to the program's goals and objectives.

- Effective evaluation requires appropriate communication of findings to all interested parties, including the program, the funding source, concerned public administrators, and the community. The evaluator and the manager must put their joint effort into developing and carrying out creative and appropriate means to communicate the findings. Evaluations presented in a positive light can do much to help a program gain support and evolve into a more effective resource for the prevention of drug abuse.

- Although the politics which surround evaluations can be a set of thorny problems, they can also be a source of opportunities. If the manager and evaluator work together to face these issues with appropriate planning and full awareness of the political context, the program, if actually effective, should be able to maximize public and funding support.

The author wishes to share his appreciation to his colleague, Robert Emrich, of the General Electric Company, for his wise observations on the topics discussed in this chapter.

REFERENCES

Aiken, L. ed., Prevention Evaluation Research Monograph, Outcome Volume. Rockville, MD: National Institute on Drug Abuse, 1981, draft manuscript.

Alkin, M. Evaluation: who needs it? who cares? Studies in Educational Evaluation, 1975, **1**, 3 (Winter), pp. 201-212.

Bird, T.; Beville, S.L.; Carlson, O.; and Johnson, G. A Design for Youth Policy Development. DHEW Publication (OHDS) 78-26042, 1978.

Bittner, E. Summary address. In: California Council on Criminal Justice. Account of the Proceedings of the Criminal Justice Research Conference (1972). Sacramento, CA.

Bukoski, W. J. Drug abuse prevention evaluation: A meta evaluation process. Paper presented at the 1979 Annual Conference of the American Public Health Association. November 4-6, New York City, New York.

Campbell, D.T., Degrees of Freedom and the Case Study, Comparative Political Studies, 1975, **2**, pp. 178-193.

Cantor, J.; Kaufman, N.; and Klitzner, M. Four Steps to Better Objectives. Madison, Wisconsin: Wisconsin Department of Health and Social Services, 1981.

Cline, H. and Sinnott, L. What can we learn about organizations? In: S. Ball, ed., Program Evaluation. San Francisco: Jossey-Bass, 1980.

Cronbach, L.J.; Ambron, S.R.; Dornbusch, S.M.; Hess, R.D.; Hornik, R.C.; Philips, D.C.; Walker, D.F.; and Weiner, S.S. Toward Reform of Program Evaluation. San Francisco: Jossey-Bass, 1980.

Davis, H.R., and Salasin, S.E. The utilization of evaluation. In: E.L. Struening and M. Guttentag, eds., Handbook of evaluation research, Vol. I. Beverly Hills: Sage, 1975.

Delbecq, A. Contextual variables affecting decision making in program planning. Decision Sciences, 1974, 5, pp. 726-742.

Delbecq, A.; Van de Ven, M.; and Gustafson, D. Group Techniques for Program Planning — A Guide to Nominal Groups and Delphi Processes. Chicago: Scott, Foresman, Inc., 1975.

Fitz-Gibbon, C., and Morris, L. How to Calculate Statistics. Beverly Hills: Sage, 1978.

French, J. F. and Kaufman, N.J., eds., Handbook for Prevention Evaluation. Rockville, MD: National Institute on Drug Abuse, 1981.

Glaser, E.M. and Taylor, S.H. Factors influencing the success of applied research: A study of ten NIMH funded projects. Los Angeles: Human Interaction Research Institute, 1969.

Hage, J., and Aiken, M. Social Change in Complex Organizations. New York: Random House, 1970.

Hays, W., and Winkler, R. Statistics: Probability, Inference, and Decision. New York: Holt, Rinehart, and Winston, 1971.

Jessor, R. Predicting time of onset of marijuana use: a development study of high school youth. In: Lettieri, D.J., ed., Predicting Adolescent Drug Use: A Review of Issues, Methods and Correlates. National Institute on Drug Abuse. Washington, D.C.: U.S. Government Printing Office, 1975, pp. 283-298.

Jessor, R. Marijuana: a review of recent psychosocial research. In: Dupont, R.L.; Goldstein, A.; and O'Donnell, J., eds., Handbook on Drug Abuse. National Institute on Drug Abuse and the Office of Drug Abuse Policy, Executive Office of the President. Washington, D.C.: U.S. Government Printing Office, 1979, pp. 337-355.

Jessor, R. and Jessor, S.L. Adolescent development and the onset of drinking: a longitudinal study. Journal of Studies on Alcohol, 1975a, 36, 27-51.

Jessor, R. and Jessor, S.L. The transition from virginity to non virginity among youth: a social-psychological study over time. Developmental Psychology, 1975b, 11, 473-484.

Kiresuk, T.; Larsen, N.; and Lund, S. Management and evaluation in a knowledge transfer context. In: R. Levine; M. Solomon; G. Hellstern; and H. Wollmann, eds., Evaluation Research and Practice: Comparative and International Approaches. Beverly Hills: Sage, 1981.

Lindblom, C.E. The policy-making process. Englewood Cliffs, N.J.: Prentice Hall, 1968.

Lippitt, R.; Watson, J.; and Westley, B. The Dynamics of Planned Change. New York: Harcourt, Brace, and World, 1958.

National Institute on Drug Abuse. Management Effectiveness Measures for NIDA Drug Abuse Treatment Programs, by Rufener, B.L.; Rachal, J.V.; and Cruze, A.M. Washington, D.C.: U.S. Government Printing Office, 1975.

National Institute on Drug Abuse. The Drug Abuse Instrument Handbook, by Nehemkis, A.; Macari, M.A.; and Lettieri, D.J. Washington, D.C.: U.S. Government Printing Office, 1977.

Orne, M.T. On the psychology of the psychological experiment. American Psychologist, 1962, 17, pp. 776-783.

Orne, M.T. and Evans, F.J. Social control in the psychological experiment: antisocial behavior and hypnosis. Journal of Personality and Social Psychology, 1965, 1, pp. 189-200.

Patton, M. Q. Utilization-Focused Evaluation. Beverly Hills: Sage Publications, 1978.

Reichardt, C.S. The analysis of covariance (ANCOVA) and the assessment of treatment effects. In: L. Aiken, ed., Prevention Evaluation Research Monograph, Outcome Volume. Rockville, MD: National Institute on Drug Abuse, 1981, draft manuscript.

Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 1974, 66, 5, pp. 688-701.

Suchman, E. A. Evaluation Research. New York: Russell Sage Foundation, 1962.

Suchman, E. Evaluative Research: Principles and Practice in Public Service Programs. New York: Russell Sage Foundation, 1967.

Tharp, R., and Gallimore, R. The ecology of program research and evaluation: a model of evaluation succession. In: L. Sechrest, ed., Evaluation Studies Review Annual, Volume 4. Beverly Hills: Sage, 1979.

Waller, J.D. and Scanlon, J.W. The Urban Institute Plan for the Design of an Evaluation. Working paper 3-003-1, Washington, D.C.: The Urban Institute, March, 1973.

Weiss, Carol H. Evaluation Research: Methods of Assessing Program Effectiveness. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Weiss, C.H. Evaluation research in the political context. In: Struening, E.L. and Guttentag, M., eds. Handbook of Evaluation Research, Vol. I. Beverly Hills: Sage Publications, 1975, pp. 13-26.

Weiss, C.H. Between the cup and the lip. In: W.A. Hargraves; C.C. Attkisson; and J.E. Sorensen, eds., Resource Materials for Community Mental Health Program Evaluation (2nd ed.). National Institute on Mental Health. Washington, D.C.: U.S. Government Printing Office, 1977.

Weiss, J.A., and Weiss, C.H. Social scientists and decision makers look at the usefulness of mental health research. American Psychologist, 1981, 36, pp. 837-847.

Yates, B.T. Improving Effectiveness and Reducing Costs in Mental Health. Springfield, IL: C. C. Thomas, 1980.

END