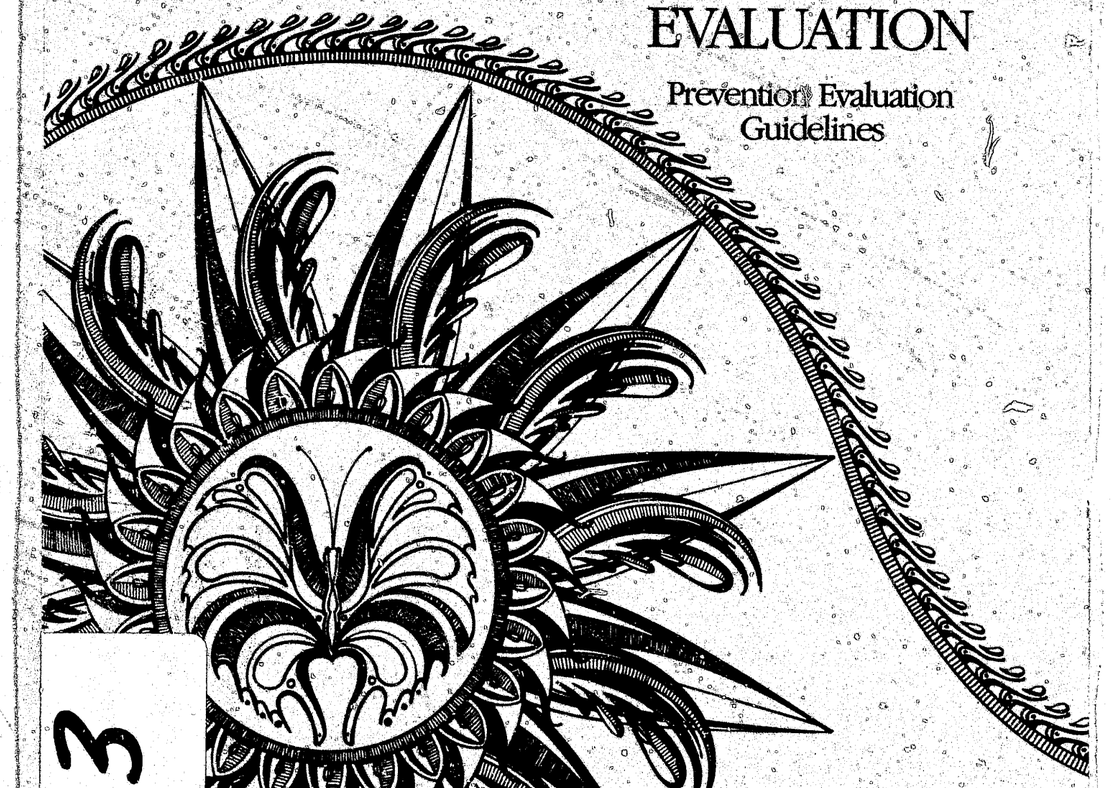
National Institute on Drug Abuse

HANDBOOK FOR PREVENTION EVALUATION

Prevention Evaluation Guidelines



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

HANDBOOK FOR PREVENTION EVALUATION

Prevention Evaluation Guidelines

Edited by
John F. French
and
Nancy J. Kaufman

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

National Institute on Drug Abuse
Division of Prevention and Treatment Development
Prevention Branch
5600 Fishers Lane
Rockville, Maryland 20857

FOREWORD

This volume represents a significant first step in the development of scientifically tested prevention programs that meet the needs of States and local communities. The Handbook for Prevention Evaluation has been written to assist evaluators working cooperatively with program managers to apply their skills in the assessment and improvement of school, community, and family based prevention programs. Much has been written in the scientific literature concerning the evaluation of action programs, such as prevention; however, this text serves as a first in that it directly applies the state-of-the-art in evaluation and program design to drug and alcohol abuse prevention programs.

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 this effort, a system for evaluation has 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. Effective prevention will be the goal of NPERN.

William J. Bukoski, Ph.D.
Deputy Chief
Prevention Branch
Division of Prevention and
Treatment Development
National Institute on Drug Abuse

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 Agencies of Wisconsin, New Jersey, and Pennsylvania, under Contract Number 271-78-4607. Acknowledgment is made to Lawrence S. Burns for his professional advice and consultation. William J. Bukoski, Ph.D., served as the NIDA project officer.

John F. French is with the New Jersey Division of Narcotic and Drug Abuse Control and Nancy J. Kaufman is with the Wisconsin Bureau of Alcohol and Other Drug Abuse.

DHHS Publication No. (ADM)83-1145
Printed 1981 Reprinted 1983

Cover Illustration • Noël van der Veen

FOREWORD

This volume represents a significant first step in the development of scientifically tested prevention programs that meet the needs of States and local communities. The Handbook for Prevention Evaluation has been written to assist evaluators working cooperatively with program managers to apply their skills in the assessment and improvement of school, community, and family based prevention programs. Much has been written in the scientific literature concerning the evaluation of action programs, such as prevention; however, this text serves as a first in that it directly applies the state-of-the-art in evaluation and program design to drug and alcohol abuse prevention programs.

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 this effort, a system for evaluation has 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. Effective prevention will be the goal of NPERN.

William J. Bukoski, Ph.D.
Deputy Chief
Prevention Branch
Division of Prevention and
Treatment Development
National Institute on Drug Abuse

CONTRIBUTING AUTHORS

James R. Beniger
Princeton Unversity

Terry C. Bloom
University of Wisconsin-Madison

Nicholas Braucht
University of Denver

Brenna H. Bry
Rutgers University

William J. Bukoski
National Institute on Drug Abuse

Royer F. Cook
Institute for Social Analysis

Robert L. Emrich
Pacific Institute for Research and Evaluation

John F. French
New Jersey Division of Narcotic and Drug Abuse Control

Teh-wei Hu
The Pennsylvania State University

Nancy J. Kaufman
Wisconsin Bureau of Alcohol and Other Drug Abuse

Lynne D. Kaltreider
The Pennsylvania State University

Michael Klitzner
Wisconsin Bureau of Alcohol and Other Drug Abuse

Erich Labouvie
Rutgers University

Nancy McDonnell
The Pennsylvania State University

Barry Milcarek
New York State Office of Mental Health

Judith Mausner
Medical College of Pennsylvania

Gregory Muhlin
New York State Office of Mental Health

Art Perlman
Wisconsin Bureau of Alcohol and Other Drug Abuse

Arthur Richardson
State University of New York

Judy Schector
Wisconsin Bureau of Alcohol and Other Drug Abuse

John T. Soper
University of Wisconsin-Madison

Elmer L. Struening
New York State Office of Mental Health

John Swisher
The Pennsylvania State University

David Twain
Rutgers University

TABLE OF CONTENTS

Page

FIGURES

TABLES

PART I
INTRODUCTION

The major purpose of prevention evaluation is the formation of a body of knowledge, with empirical and theoretical foundations, that defines what prevention strategies work, with which groups of people, and at what level of effectiveness. The usefulness of this data base is readily apparent. Evaluation research of prevention programs provides information about the strengths and weaknesses of a particular prevention strategy, and yields qualitative and quantitative findings that can be used to improve or enhance moderately effective strategies, redesign or implement more effective strategies, discard strategies that have been shown to be ineffective, and plan prevention programs that address unmet needs of individuals within a target population.

Essential to this evaluation research perspective is the recognition that if the field of prevention is to continue to develop as a formal discipline, new knowledge needs to be applied in the program setting. And utilization of evaluation research findings stands as the ultimate criterion of success or failure of evaluation research studies. Evaluation research that addresses the relevant program issues of the day by testing appropriate evaluative research questions is encouraged; for, these findings will either support or refute the operational assumptions that constitute the existing knowledge base of effective prevention programing.

In 1978, the National Institute on Drug Abuse (NIDA) funded a consortium of states (Wisconsin, New Jersey, and Pennsylvania) to (1) determine the needs in state and local prevention evaluation and (2) design a National Prevention Evaluation Resource Network (NPERN) that will meet these needs. Some of the most pressing needs identified in this study include a larger body of evaluation information, skilled technical assistance, and guidance in conducting drug abuse prevention evaluations. Although NPERN was initially a NIDA initiative, the network has received enthusiastic support from alcohol abuse prevention programs and the National Institute on Alcohol Abuse and Alcoholism (NIAAA). NPERN philosophy, models, products, and services are relevant to the concerns of alcohol abuse prevention programs and responsive to their needs.

The purpose of the Prevention Evaluation Guidelines is to provide a technically sound, logical, and useful frame of reference for the acquisition of new knowledge in prevention. Chapter 1 provides the reader with a general orientation to prevention concepts and program strategies, and chapter 2 delineates the model for prevention evaluation. The remaining chapters of this publication cover technical areas of special interest to the prevention evaluator, for whom this particular monograph has been developed. A second publication is being written for the prevention program manager, and will delineate the important role to be played in the evaluation process by the prevention program specialist.

Both publications clearly illustrate the evaluation philosophy and consultative practices recommended by the National Institute on Drug Abuse (NIDA), the National Prevention Evaluation Resource Network (NPERN), and the National Institute on Alcohol Abuse and Alcoholism (NIAAA). This philosophy is based on the premise that technically sound and programatically useful evaluations are best accomplished when the prevention program specialist works in a professional partnership with a well trained prevention evaluator. Working side-by-side, the prevention program specialist and the prevention evaluator bring their unique skills to the task of designing an evaluation that is tailored to the needs of the program.

This team approach to prevention evaluation will result in an evaluation design that is sensitive to the specific outcomes intended by the prevention program and consistent with the cultural context in which the program operates. It is NPERN's philosophy that attention must be given to culturally relevant issues in designing meaningful and constructive evaluations. The result of this cooperative relationship is evaluation that yields technically sound data useful for program improvement and enhancement.

This evaluation perspective is pivotal for the implementation of a National Prevention Evaluation Resource Network (NPERN), in which State and local prevention program staff participate fully with experienced prevention evaluators to design evaluation plans that meet the information requirements of individual programs.

ORGANIZATION OF THE GUIDELINES

The Guidelines has a four part organization which in effect parallels the sequence of steps an evaluator undergoes in conducting an evaluation. Part I consists of this introductory chapter, followed by a model for evaluation--really the core of the Guidelines. Part II (chapters 3-5) discusses the indicators and measures used in the three levels of evaluation, that is, process, outcome, and impact. Part III concerns methodology. It begins with a presentation of general methodological issues and proceeds with a consideration of specific methods appropriate to the evaluation of process (chapter 6), outcome (chapter 7), and impact (chapter 8). Chapter 9 discusses many of the major statistical issues facing the evaluator. It then describes the most important data analysis strategies available, and discusses their range of applicability. Finally, Part IV (chapter 10) addresses the critical topic of utilizing evaluations: how to increase the likelihood that your evaluations will be used.

NPERN GUIDELINES

CHAPTER 1: INTRODUCTION

The prevention evaluation guidelines which follow were developed as part of the NPERN contract to address the needs of skilled technical assistance and guidance in the development and implementation of prevention program evaluations. The Guidelines provides a consistent frame of reference for people conducting evaluations of drug abuse prevention programs. It specifies in detail the principles, procedures, and techniques recommended for evaluations. The Guidelines is primarily aimed at the experienced evaluator, in particular the NPERN evaluation consultant to state and local prevention programs. However, it is anticipated that the Guidelines will also be useful to individuals without an extensive knowledge of evaluation procedures and methodologies; such people may include prevention program staff and administrators. In fact, a major objective of the Guidelines is to facilitate and improve communication between prevention and evaluation personnel. Close collaboration between these groups at all points during the evaluation is viewed as a prerequisite to obtaining useful program feedback. The Guidelines is also intended to foster a level of consistency among evaluations that may aid in advancing our state of knowledge. However, the great diversity in strategies and programs for preventing drug abuse precludes the establishment of rigid evaluation standards. Therefore, the Guidelines should be thought of as a set of recommendations.

THE CONCEPT OF PREVENTION

Much debate has taken place regarding an appropriate definition of prevention. The sample of definitions that follows perhaps conveys a sense of the various philosophies that underlie primary prevention programs:

> Prevention is an active process of creating conditions that promote the well-being of people.
>
> --Associates for Youth Development, Inc.

> Prevention includes purposeful activities designed to promote personal (emotional, intellectual, physical, spiritual, and social) growth of individuals and strengthen the aspects of the community environment which are supportive to them in order to preclude, forestall, or impede the development of alcohol and other drug abuse problems.
>
> --Wisconsin State Drug Abuse Plan

> Primary prevention is directed at reducing the occurrence or incidence of alcohol, drug abuse, and mental health disorders. This goal is achieved through the promotion of physical, mental, and social growth toward full human potential. Prevention activities are directed towards specifically identified high-risk groups within the community who can be helped to avoid the onset of mental and emotional dysfunctioning and to inhibit the use of alcohol and drugs.
>
> --Alcohol, Drug Abuse and Mental Health Administration

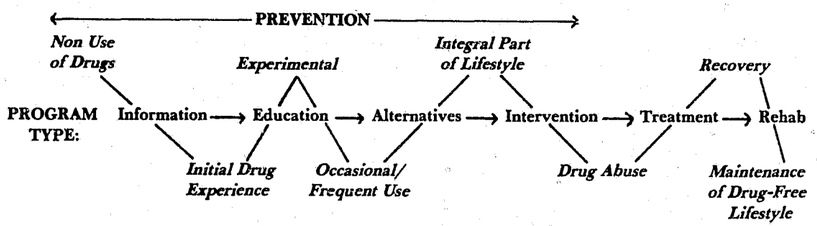
Another way to break down the concept of health promotion is to consider the community as well as the individual. We are accustomed to think of an individual's health, both in terms of treatment and building resistance, but we can extend this to the community. Often people succumb to ill health in part as a result of forces in the social context. Such could include unemployment, insensitive institutions, including schools, or prevalent attitudes which reinforce unhealthy behaviors. If this is the case, then it makes sense to design programs which deal with these factors.

--Vermont Alcohol and Drug Abuse Division

When drug and alcohol abuse were first adressed as major social problems, attention was directed toward individual abusers. However, in attacking only the most visible manifestations of the problem--mostly through treatment and rehabilitation programs--only a short-term, partial solution was provided. It failed to confront wider ranging societal issues. Consequently, increasing attention is being focused on prevention programs, i.e., those that attempt to "preclude, forestall, or impede" the development of the problem.

NIDA places prevention, treatment, and rehabilitation activities along a Drug Abuse Program Continuum, as follows:

Figure 1. NIDA Drug Abuse Program Continuum



Within this framework, specific programs or program types targeted at particular audiences can be identified. Typical of prevention programs are drug education in schools, recreational alternatives, peer counseling, employe wellness programs, community action projects, and information and referral centers. Treatment and rehabilitation programs include therapeutic communities, outpatient counseling, and methadone maintenance. It is important to stress that this continuum is not intended to restrict prevention activities to any one point. Instead, it indicates for each point the modality which is expected to have the major effect. This does not negate the potential influence of other modalities at a given point, since any modality can be expected to have at least some influence.

To help define alcohol problems and appropriate intervention points, NIAAA has adopted a public health model to provide a conceptual framework for prevention planning. The model suggests that a disorder is caused by a combination of factors--host, agent, and environment. It is thought that intervention at any of these points can prevent a disorder from occurring. Loosely adapted to alcohol, prevention strategies can address the individual drinker (host) and his or her attitudes, beliefs, knowledge, or behavior; some characteristic of alcohol (agent) itself that would include availability, taxation, or advertising; and the environment, both social and physical, in which people drink or must function after drinking. Most work in the past has focused primarily on the host or individual. Prevention specialists are now looking at the availability of alcohol and the environment in which drinking is done as appropriate targets of prevention programing.

MODALITIES OF PRIMARY PREVENTION

The Drug Abuse Program Continuum indicates the four modalities of primary drug abuse prevention identified by NIDA. These are: information, education, alternatives, and intervention. Note that a given prevention program can, and often does, incorporate more than one modality.

Information programs are rooted in a belief in the human's potential for self-improvement through responsible decision making. The basic notion is that, given adequate information, people are likely to make rational choices and adopt behavior deemed socially desirable. Thus, information programs are designed to provide accurate, honest, and timely information about all types of drugs and their effects on the human system. Information is basic to good prevention programing, but may be ineffective or counterproductive if not founded upon clearly articulated communication objectives that use market research data to tailor specific messages to target audiences. Information programing includes such activities as appropriately targeted media campaigns, flyers, posters, brochures, or drug information seminars for youth, parents, and other target groups.

Education programs are based on the theory that people are powerfully driven to satisfy certain basic needs, such as love, security, and self-identity. Ordinarily these needs are met socially and in ways that are self-enhancing. However, people who are deficient in such critical skills as decision making, problem solving, and communicating have more difficulty meeting their basic needs. Unless these deficiencies are corrected, people may opt to satisfy their needs through undesirable behavior, such as drug and alcohol abuse. Education programs assist these individuals to develop or improve their critical life skills. The usual setting for education programs is in the schools, although they may take place in the general community or workplace. Prevention curricula often include learning activities that promote the development of skills in decision making, coping with stress, values awareness, problem solving, interpersonal communication, and intrinsic motivation. Career awareness and planning are important adjuncts to curricula, as are parent education and involvement.

Schools initiating prevention programs give consideration to humane settings, strong counseling components, and the provision of curricula relevant to the needs of young people in today's world. A popular innovation is the integration of drug and alcohol prevention units within health education or related subject areas such as social studies and science, where the discussion of drug use and human behavior or drug effects on human physiology is appropriate. The selection of specific prevention curricula and other education activities is dependent on the perceived needs of parents, children, and educators within the target community.

The theoretical basis for alternatives programs is that participation in activities which foster a positive awareness of self and others and which offer exposure to a wide range of enjoyable and rewarding nondrug activities will deter drug and alcohol abuse. Alternatives programs provide challenging positive growth experiences in which people can develop the self-discipline, confidence, personal awareness, self-reliance, and independence they need to become socially mature individuals. Alternative-prevention programs are designed to offer positive alternatives to drug taking behavior through a potpourri of community activities. Alternative programs provide opportunities where young people, especially, can increase their range of experiences, learn craft skills, be assured about their own value, and savor the lasting satisfaction that comes from being an involved, responsible, and trusted member of the community. A critical ingredient in a youth alternatives program is that young people have a very real form of ownership and self investment in the program. These programs foster constructive peer pressure, created by working together on meaningful tasks that meet a community need and that are recognized by community leaders as a valued contribution to the improvement of the quality of life within that city, town, or county.

The objective of intervention programs is to help individuals assess their problems and seek solutions to them. Intervention involves giving assistance and support to people during critical periods in their lives, when person-to-person communication, sharing of experiences, and empathetic listening contribute to a successful adjustment to personal (including drug and alcohol related) or family problems. Intervention techniques include personal and family counseling, hotline assistance, and "rap" sessions. A recent and promising approach trains young people to work in prevention through such strategies as peer counseling, peer tutoring, cross-age tutoring (older child to younger child), and the creation of new peer groups.

To these four modalities, NIAAA has added two additional categories, environmental change and social policy change.

Environmental change programs seek to identify and change social and physical environmental factors that influence drinking behaviors and patterns. Change methods may include modifying drinking settings or settings which are unsafe for intoxicated individuals, insulating others from undesirable drinking behaviors, or changing reactions to drinking behavior. Environmental change strategies may be developed in conjunction with educational approaches. Examples include providing transportation for drunken individuals or making food and nonalcoholic beverages available at parties or in college pubs.

Social policy changes may influence drinking behaviors in the general population. Strategies for modifying social policies include: changing laws, regulations, and enforcement procedures governing alcohol availability and distribution; changing taxation policies; modifying alcohol advertising practices; or reducing possible negative consequences of drinking through, for example, consumer product safety or automobile safety regulations. These approaches are most often carried out by a governmental unit or institution, and should be monitored for impact on the occurance of alcohol problems.

Though each prevention modality can, in and of itself, be an effective tool in the prevention of drug and alcohol abuse, reviews of prevention evaluation research studies suggest that a comprehensive prevention approach which includes more than one modality is more effective than a single mode program in changing factors that may lead to drug and alcohol abuse, and in moderating the use of alcohol and a variety of drugs. In effect, a prevention program needs to include a number of prevention modalities.

EVALUATION OF PREVENTION

Evaluation generally refers to activities undertaken to measure a program's success in achieving specified objectives. This definition is useful because it highlights some of the formidable obstacles facing prevention evaluation. First, as noted earlier, the strategies and policies underlying various prevention programs are so diverse that any specifying of objectives is necessarily difficult. Second, the communication that is needed between preventors and evaluators to designate objectives and develop a workable evaluation plan is all too frequently absent. Third, measures and methodologies for data collection and analysis may not be readily available. Despite these and other obstacles, the benefits of evaluation are becoming increasingly apparent. Evaluation can suggest areas for program improvement and provide a rational basis for allocating limited resources. It can help develop a sound body of prevention knowledge concerning theory, modality, technique, and ways of pinpointing target populations. Hence, the increased federal interest in evaluation. The justification for funding prevention programs is that they will improve the quality of health in the community at reasonable cost. NIDA's and NIAAA's prevention grants programs are devoted to validating drug and alcohol abuse prevention approaches through evaluation research. For example, NIDA's evaluations of drug abuse media campaigns are providing feedback on public information strategies. NIAAA uses evaluation findings to select model programs for replications. The funding of the NPERN system and the technically innovative Prevention Evaluation Research Monograph (PERM) series signal an era of increased interest and commitment to prevention evaluation.

The potential uses and users of prevention program evaluation are as diverse as the strategies. Uses which may be made of quality evaluations include:

Program Feedback--providing ongoing information to guide a project's operation. The feedback may help to:

- Increase client satisfaction
- Reexamine theoretical assumptions
- Redirect program goals and objectives
- Review program operations
- Improve results
- Improve responsiveness to the community.

Accountability--demonstrating to others that a program is worthwhile. Establishing accountability may:

- Aid program monitoring
- Justify future funding
- Enhance public relations
- Aid in allocating agency funds
- Meet funding or legal requirements
- Increase cost effectiveness and benefit.

Program Development--testing prevention theories, modalities, and techniques as used with target populations in order to expand the body of prevention knowledge. It may:

- Provide state of art data
- Assist in comparing prevention techniques
- Help programs choose among different techniques
- Help planners design programs
- Provide a mechanism for replicating new techniques and programs.

Potential users of prevention evaluation include (1) prevention programs and (2) communities, legislatures, and society at large. Some specific uses are indicated below:

Programs

- Fulfill funding or legal requirements
- Determine if the program is doing what was intended
- See if the program is worthwhile
- Compare the effectiveness of prevention techniques and question underlying assumptions
- Undertake rationally planned growth or changes in program direction.

Communities, Legislatures, and Society

- Find out whether tax dollars are being wisely spent
- See if social problems are being adequately addressed
- Determine resources (time, money, staff, facilities) needed to run an effective program.

The extent to which the above "markets" will be served depends to a large extent on the quality and utility of the evaluations--and on the individuals who plan, conduct, report on, and determine the utilization of results. It is hoped that the Guidelines will contribute to the realization of this potential.

CHAPTER 2: MODEL FOR EVALUATION

INTRODUCTION

NIDA's Prevention Branch has developed an evaluation research model that is applicable to any of the four drug abuse prevention modalities (information, education, alternatives, and early intervention) and any of the five primary targets (individuals, peers, families, the school, and other significant social institutions) (Bukoski 1979). The model, illustrated in Figure 2, features three levels of evaluation: process, outcome, and impact. Each level has its own set of indicators of effectiveness and its own appropriate evaluation methodologies.

Process evaluation refers to an assessment of a prevention program that includes identification of the target population, a description of the services delivered, the utilization of resources for the programs, and the qualifications and experiences of the personnel participating in them. Process evaluation attempts to capture in "still frame" the dynamics and characteristics of an operational, ongoing prevention program.

Outcome evaluation is concerned with measuring the effect of a project on the people participating in it. This includes youths, parents and families, youth workers, teachers, and so on. Outcome evaluation attempts to answer the question: "What has this program produced relevant to changes in the lifestyles, attitudes, and behaviors of those individuals it is attempting to reach?" In essence, outcome evaluation tries to determine if a prevention project has met its own objectives.

Impact evaluation explores the aggregate effect of prevention programs on the community as a whole. The community may be defined as a school, neighborhood, county, city, state, region of the country, or the nation. The purpose of impact evaluation is to gauge the effects of numerous drug abuse prevention programs operating within a geographic boundary, or of an individual drug abuse prevention program operating over an extended period of time. Impact evaluation assesses a variety of macro-indicators relating to drug and alcohol abuse at the community level. In contrast to the other two levels of the model which are directed at assessing a specific program, impact evaluation measures the generalized effects of drug and alcohol abuse prevention programs operating within the totality of the community. Measuring changes in the quality of health within a community is the focus of impact evaluation. Potential indicators of impact relevant to drug and alcohol abuse prevention include changes at the community level in the prevalence and incidence of drug and alcohol use, related mortality and morbidity, rates of juvenile delinquency, drug and alcohol related accident rates, or changes in institutional policies and programs.

Figure 2 summarizes the essential features of NIDA's prevention evaluative research model and also shows examples of evaluative research methodologies or techniques that are appropriate to each level of evaluation. For example, the Cooper evaluation model is appropriate for process evaluation; experimental or quasi-experimental approaches are appropriate for outcome evaluation; and epidemiologic research is appropriate for impact evaluation. The key element of this evaluative research model which has applicability to all types of prevention programs is the emphasis given to measurable indicators of program success. At each evaluative level (process, outcome and impact), the suggested indicators of effect are empirically based (for example, drug and alcohol use levels, drug and alcohol related accident rates) and demonstrate the observable changes in individuals, institutions, or communities, brought about by the prevention program.

Figure 2. 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

The Guidelines operationalizes NIDA's prevention evaluative research model. It offers a conceptual framework for presenting evaluation issues, strategies, and methodolgies. It also illustrates a recommended procedure by which prevention program evaluation may be conducted.

The Guidelines focuses on two kinds of issues pertinent to evaluation, those which (1) are necessary for effective evaluation of drug and alcohol abuse prevention and other human service programs, and (2) reflect more broadly on the current thought regarding procedures and strategies that will enhance the quality of evaluation research.

NEED FOR A MODEL

Numerous surveys of evaluations in human service areas, including drug and alcohol abuse prevention, consistently find that:

- Few evaluations are performed in response to previously stated decisionmaking requirements.
- Many evaluations suffer from serious methodological deficiencies.
- Most evaluations focus on outcome, with little or no information on program process or on impact within the community.

There are three major concepts critical to the effectiveness of drug and alcohol abuse prevention evaluations. First, the field of evaluation research has developed a wide range of methods and strategies, building on the many scientific disciplines that have contributed to the evaluation of human services, namely, psychology, sociology, anthropology, political science, statistics, operations research, economics and computer science. Evaluators working in the field of drug and alcohol abuse prevention need to be aware of this body of knowledge and its appropriate application.

Second, evaluators need to know the strengths and weaknesses of various methodologies (designs, measures, data analysis techniques). This is essential to selecting appropriate methods and in utilizing findings.

Third, techniques exist which can enhance the likelihood that evaluation findings will be utilized. Evaluators need to be aware of these techniques and assume the responsibility for applying them.

In addition to these needs, there has been and remains pressure from many sources (taxpayers, Federal and State agencies, legislators) for more effective evaluation in all the human services. Drug and alcohol abuse prevention, because of its recent emergence as a human service field, is especially in need of effective evaluation in order to demonstrate the importance of adequately funding programs and projects. In part because of this pressure, people in the field are especially receptive to efforts to improve the quality of evaluations.

The Guidelines adresses the above concepts, so important to the effectiveness of drug and alcohol abuse prevention evaluations. One objective of the Guidelines is to provide a broad survey of evaluation technology so as to acquaint evaluators and their customers with the range of options available, and thus aid them in securing the required information efficiently and effectively. Another objective is to increase the ability of evaluators to recognize both the usefulness and limitations of their findings. The results of even the most sophisticated research are likely to have some limitations, which the user must be aware of in order to make reliable use of the information. Toward this end, the Guidelines will review the principal sources of bias in evaluation research, and indicate the best remedies and approaches for dealing with them. Where a critical bias has not been controlled for, or significant error weakens a result, the Guidelines will point out what limited use can be made of the flawed results.

With respect to utilization, the Guidelines takes the position that it is the evaluator's responsibility to increase the likelihood that new knowledge will be applied by decision makers. Evaluators must do what they can, within reason, to assist programs to effectively

use evaluation results for program improvement. Elsewhere in the Guidelines are detailed procedures for implementing evaluation findings.

The Guidelines does not present a logical theory of evaluation; it attempts rather to organize information concerning evaluation into a particular framework. This framework is designed to be sufficiently specific to guide evaluators in the conduct of effective and useful evaluations yet flexible enough to encourage incorporation of new developments in prevention programing and evaluation technology.

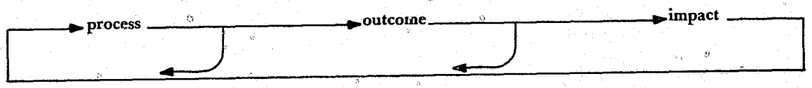
EVALUATION PARAMETERS

The Guidelines proposes three major parameters of evaluation. This organization is appropriate for evaluation regardless of the point at which formal evaluation activities are begun. The three parameters are: levels of program evaluation, type of evaluation information, and target area.

LEVELS OF PREVENTION EVALUATION

The levels of evaluation refer to the successive phases in the development of information in an ideal evaluation effort. This can be represented in the following systems diagram:

Figure 3. Evaluation Levels



The relationship between the three levels of evaluation as depicted by figure 3 is incremental in that a sound process evaluation cogently describes the program whose level of effectiveness is measured during an outcome evaluation. The resulting knowledge concerning the effectiveness of a specific prevention program (or programs) operating within a community then forms the basis for assessing the generalized changes (impact) in drug use or drug and alcohol related indicators at the community level that have occurred over time. Also considered during the impact evaluation would be factors, events, or conditions that have a causal or associative relationship to the level of drug or alcohol use and related behaviors, which are indicative of the quality of health of a community.

Each level of evaluation produces valuable data needed for prevention program improvement. Given the specific information needs of a particular program, the emphasis given each evaluative level may vary. For example, a prevention program at an early stage of its development may be more concerned with the thorough documentation of program events resulting from a sound process evaluation, rather than a comprehensive analysis of program effects produced by an outcome evaluation. Likewise, not all prevention programs would require, given their stage of development, a measure of their contribution to the improvement of the quality of health within a community as produced by an impact evaluation. However, the importance of each level of evaluation to the one following it is clear. That is, sound process evaluations form the basis for meaningful outcome evaluations, which are essential for relevant impact evaluations conducted at the community level.

Process Level

Process information reflects the inputs that go into a program, the patterns in which these inputs interact, and the transactions that take place within the program. Information such as participant and staff characteristics, physical plant characteristics, and financial resources, as well as the theory on which the program operates, needs assessment, policy development, and program design activities are all examples of program inputs. Information derived from the socio-political environment is also considered to be important evaluative information because of its potential contribution to subsequent evaluation and its use as a basis for record keeping systems. Other assessments on the process level may include a description of services rendered, the decision-making structure, patterns of interaction among participants and staff, and so on.

Outcome Level

Information gathered during this phase of program evaluation typically is addressed to specific program objectives concerned with change in participant behavior, attitudes, values, or knowledge. The major objectives in all prevention program modalities concern the reduction of inappropriate drug and alcohol use. At the same time, different prevention programs have unique objectives relating to the particular theories underlying them. For example, some programs attempt to deal with a variety of risk states associated with drug and alcohol abuse, such as low self-esteem, irresponsibility, alienation, and poor school performance. And this list is far from exhaustive.

Impact Level

Information gathered in this phase relates to longer term, generalized results of program operations. The manner in which impact data are relayed is a function of the community needs and problems which gave rise to the prevention program in the first place. That is why such broad issues as changes in incidence and prevalence in drug and alcohol abuse and in community competence to deal with these problems are frequently addressed in impact evaluation. Such changes impinge directly on inputs to the program.

INFORMATION TYPE

The type of information obtainable in an evaluation is necessarily constrained by the availability of data; but it is also a function of the evaluation design and the choice of analytic technique. The Guidelines identifies three types of information: descriptive, comparative, and explanatory.

Descriptive information is the easiest to obtain and frequently can be drawn from program records. However, program records are often inadequate. Thus, development of a management information system comprised of descriptive information categories is a perfectly legitimate first component of an evaluation.

Comparative information relates variables thought to significantly effect program functioning, but without assigning causality. Obtaining this type of information usually requires more elaborate design, more time, more cost, and more justification to management than obtaining descriptive information.

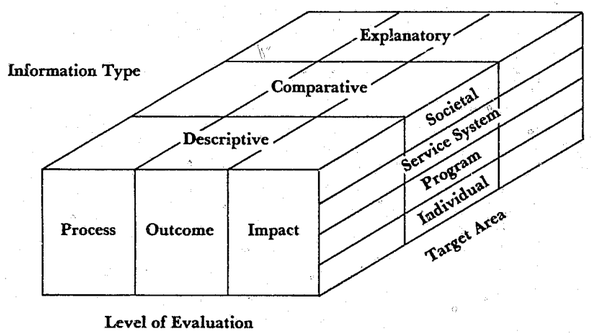
Explanatory information attempts to answer the question "why does the program work?" The development of this type of information requires still more sophisticated design, theory testing and basic knowledge building than for comparative information.

TARGET AREA

Maintaining a systems oriented focus, it is important to realize that evaluation can be directed at different targets or subsystems of the overall program. The level of focus can influence significantly the type of question asked. The most common targets of analysis are the individual, face to face group, program, service system area, and, finally, components of the general society. The well-publicized success or, more realistically, the failure of one individual in one program can have significant repercussions throughout the system and may influence policy at the societal level. Conversely, a decision at a high level can dramatically influence the behavior of individuals in local programs.

Figure 4, is a matrix of the parameters discussed on the preceding pages. It attempts to depict the possible interactions between, and combinations of, level, type of information, and target area of evaluation. The matrix is presented to illustrate that there is a potential for meaningful analysis within each cell. However, some cells are infrequently, if ever, found in evaluations. The choice of cells in any particular evaluation depends upon the needs of decision makers and the availability of resources.

Figure 4. Matrix of Evaluation Parameters



SYSTEMS CHANGE USING EVALUATION: PROGRAM DEVELOPMENT

Properly employed, evaluation ensures that program development will be a rational process, anchored in theory and based on the constant supply and assessment of feedback to programs. It follows that the maximum potential effectiveness of evaluation will be realized if evaluation has a role from the first stages of program development. But in reality, actual introduction of an evaluation into a program can occur anytime in a program's life, and the point at which it is introduced has implications for the type of feedback. For example, planning or initiating evaluation in the earliest phases of program development makes possible the collection of data which would otherwise be unavailable later.

Given the link between program development and evaluation, it is useful to examine five major phases of program development and the evaluation issues associated with each. The phases are listed below:

- Needs assessment
- Policy development
- Program design
- Program initiation
- Program operation.

The first three phases may be considered planning operations, whereas the last two are implementation activities. A similar classification will be used in the discussion of the evaluation plan. Each phase has associated with it a major issue for program evaluation that may not be explored or even understood if the evaluation is not introduced until sustained program operation is achieved. A brief discussion of these phases and their associated evaluation issues follows.

The needs assessment phase of program development is a planning activity which attempts to establish whether and to what extent certain previously defined problems and needs exist in a community and which subgroups are affected. The major issue for program evaluation at this point is one of external validity. That is, program ineffectiveness can result from incorrect assessment of the problem. Specifically, the evaluator must realize that

no matter what program is eventually put into operation, it should have a valid needs assessment as its foundation.

The policy development phase establishes the goals and specific objectives for the local intervention or program area. The issue for evaluation here is one of construct validity. In this instance, either the causal theory may be inappropriate or it can be improperly translated into policy (that is, improper translation into independent or dependent variables). There may not have been, for instance, appropriate understanding and consideration given to certain community values and other critical factors in the socio-political environment.

The program design phase involves transforming policy into significant characteristics of the program (for example, the target population, personnel qualifications, intervention methods, and other program aspects). Again, evaluators must be aware of a construct validity issue. Program policy may be appropriate, but the program itself may fail because of an improper translation of policy.

The program initiation phase calls for the translation of theory into action. It is then that the program is implemented. Many evaluation practitioners believe that it is in this phase that program evaluation data collection first takes place. In other words, there is a difference between the evaluation that takes place during needs assessment, that which takes place during policy development or analysis, and that which begins with the implementation of program activity. The focus of evaluation in the program initiation phase is on the identification of participants, resources, and constraints. The major issue for evaluation at this time is one of external validity. Program design may be appropriate, but the program may still fail due to improper implementation of the design.

The program operations phase involves those critically important internal transactions which are a major focus of management information systems. The predominant issue for evaluation activity during this "process" phase is one of internal validity. Program implementation (initiation) may be appropriate, but the program may fail anyway because of faulty management (for example, high staff turnover and insufficient supervision).

The major issue for evaluation in relation to program results, both outcome and impact, is one of conclusion validity. Program operations may be appropriate, but failure still may result from the influence of external factors. In addition, throughout all five phases of program development, statistical conclusion validity is an issue--it may lead to unclear or misinterpreted outcome or impact data.

EVALUATION PLAN

These Guidelines are based on the proposition that any assessment of program value must be made in the context of community need and alternative strategies for meeting those needs. The ideal evaluation activity is as responsive as possible to the socio-political environment surrounding the program activity, as well as to the needs of program decision makers.

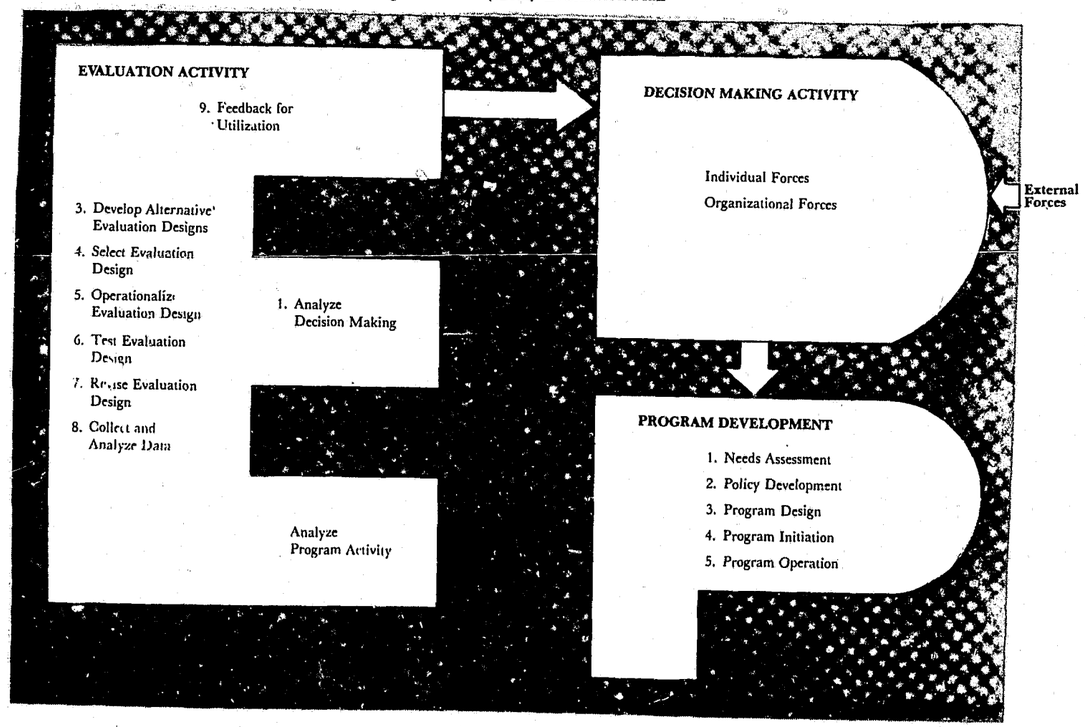
The ideal evaluation plan[1] is seen as consisting of nine sets of activities, each of which builds on preceding activities. Feedback to program decision makers and evaluators, and consequently utilization of results, can occur from any activity and thereby provide for continuous modification of program and evaluation activities. (See Figure 5.)

The basic premise in the implementation of the evaluation plan is that chances for an effective and useful evaluation to occur are maximized when a skilled evaluator works in a cooperative fashion with an equally skilled program staff member. A collaboration of this sort will produce an evaluation plan that is most responsive to the information needs of the program.

The evaluation activities are listed below in the order in which they normally occur:

1. Analysis of decision making activities

Figure 5. The (Ideal) Evaluation Plan



2. Analysis of actual or intended program activities
3. Development of alternative evaluation designs
4. Initial selection of a design
5. Operationalization of the design
6. Field test of the evaluation plan or revisions of the plan
7. Revisions resulting from the field test
8. Collection and analysis of data
9. Utilization of information resulting from interpretation of collected and analyzed data.

ANALYSIS OF DECISION-MAKING ACTIVITY

Ideally, the objectives or purposes of an evaluation will determine the type and amount of information to be collected and analyzed, as well as the appropriate uses that can be made of evaluation results. The NPERN model stresses that these objectives or purposes should be related to the needs of the users. Thus, the first step in the evaluation process is to identify the primary users and assess their needs, for example, their requirement for information relating to specific decision making activities.

Next, the evaluator and the decision maker should specify the kinds of information or indicators that are relevant for the decision-making activity and the amount and detail of information that is necessary. It can be assumed that there will be a tendency to over identify information "needs." Thus, the next component in this task is to differentiate information that is desirable from that which is essential. One way to do this is to assess the expected impact of the information, or its absence, on decision making and program activites.

A final step involves determining the quality of data that will be acceptable to and used by the decision maker. Quality of data is controlled by the evaluation design, measurement procedures, and analytical procedures. The question is whether or not the decision maker requires information best provided by quasi-experimental design, qualitative assessment techniques, or a true experimental design.

ANALYSIS OF PROGRAM ACTIVITY

An effective evaluation requires a program that has: (1) testable program assumptions, (2) clearly specified and measurable objectives that address specific risk states to drug and alcohol use, and (3) documented program strategies. Collaboration of program personnel and evaluators in the analysis of program activity substantially increases the possibility that the program will meet these requirements and that there will be a commitment to use the results.

The analysis of program activity, coupled with a study of decision making, provides the information needed by an evaluator to develop alternative designs. The analysis seeks to identify basic characteristics of the processes of the program, and its operating relationship to the ideals of planners, legislators, and others. Opinions and values may be challenged and revisions may be required.

The conceptual basis of the program should be clearly understood. This includes the assumptions or hypotheses on which the program is based and the rationale for the modalities in effect. The evaluator should know what the assumed dependent and independent variables are, and how the various program strategies are intended to effect the changes identified in the objectives.

Program objectives should be stated in quantifiable terms that describe (1) the changes being sought, (2) the degree, extent, or pattern of change, (3) how the changes will be measured, and (4) the time frame within which the change is expected to occur.

The documentation of program processes or activities is important to the evaluator because of the implications they have for certain dimensions of evaluation. Program recruitment, referral, or intake procedures all shape the design to be used in a program evaluation. The manner in which services are delivered, let alone the objectives and the content of the service, can affect the type and timing of measurement and the unit to be measured, as well as the costs and quality of data. The development and the maintenance of a good record system is one way that a program can ready itself to contribute to effective evaluation. Design and establishment of a data base that provides an accurate picture of a program's inputs and processes should be one of the first steps taken in an evaluation effort. Such data are most useful in planning the evaluation.

DEVELOPMENT OF ALTERNATIVE EVALUATION DESIGNS

The preceding activities provide the information needed to design a feasible evaluation plan. Many texts on evaluation research stress the need for evaluation research to model itself along the lines of classical experimental designs. While such designs have an important role in outcome and impact evaluation, they are of limited use in process evaluations. Furthermore, there are alternative approaches to evaluation that may make important contributions to decision-making and may be more appropriate than the classical approach, given time and resource constraints or the dynamics of the program being considered.

Designing an evaluation requires that choices be made carefully among information options, which are themselves subject to time and resource constraints. Ideally, the evaluator should prepare several workable evaluation plans that will meet the identified needs of the decision maker. The plans will likely vary as to the following: type of information (explanatory, descriptive, associative); timing of measurements (including both frequencies and intervals); measurement techniques (interview/questionnaire, observation, archival); qualitative versus quantitative assessments; single versus multiple measures; and--obviously --who and what is measured. At issue is the quantity and quality of information to be produced and the costs associated with each. Many drug abuse prevention projects are funded for less than $50,000 per year, and this may cover the cost of an evaluation as well as the expense of operations. A project of this scale usually can afford to spend at most a few thousand dollars on an evaluation. In addition, such projects also may be able to contribute staff hours and time of the administrator. Despite financial limitations, evaluators should be able to assist such a project, perhaps by obtaining the bulk of the desired information from already existing records.

INITIAL SELECTION OF AN EVALUATION DESIGN

To enable the decision maker to make an informed choice among alternative plans, the evaluator should rank the plans according to criteria relating to the decision maker's needs identified in step one (for example, the level of confidence associated with each design, resources required, and other advantages and limitations). This process may result in changes in previously identified needs and considerations so that additional design development may be necessary. In effect, the development-selection processes may require several iterations until an initial, feasible evaluation plan is selected.

PUTTING THE EVALUATION DESIGN INTO AN OPERATING CONTEXT

Having selected an evaluation design, the evaluator and program personnel will "operationalize" the plan. Instruments need to be selected or developed, and design elements of sampling, data collection, data analysis, and utilization procedures specified and incorporated into a time frame. Appropriate roles for evaluators and program personnel are also spelled out.

One strategy for ensuring that an evaluation is intimately tied to project development and that the results are understood and utilized by decision makers working with the project, is to build an active role for project personnel in all phases of the evaluation. The role of project staff can vary greatly. They may conduct the actual evaluation, or they may work closely with an outside evaluator.

The role of the evaluation consultant too may vary. In some cases it will correspond to that of the independent evaluator. Where the project staff assume a primary role in the evaluation process, the evaluator may function as a guide or resource person--s/he may introduce appropriate technical options and help with the design of the evaluation and the selection among alternatives. S/he may also provide training and technical assistance to enhance or complement the skills of the project evaluation staff.

FIELD TEST OF THE EVALUATION PLAN

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 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 decision maker.

REVISE EVALUATION DESIGN

Following the field test, evaluators and program personnel should review the plan and its initial operation to determine what, if any, revisions should be made and what procedures should be followed to implement the full scale evaluation.

ROUTINE DATA COLLECTION AND ANALYSIS

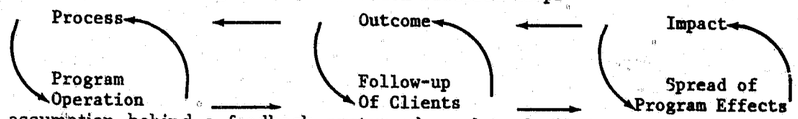
Implementation of the evaluation process on a full scale requires routinized data collection and analysis. As ideally envisioned in this model, data will be produced and interpreted in a scheduled series of oral and written reports, along with special reporting as required. It should be noted that during this stage of the evaluation process the preceding evaluation activities may be continued or repeated. This is a major feature of the incremental evaluation plan--learning is open ended and no phase is ever completed.

UTILIZATION

The evaluation cycle is completed with the feedback and utilization of results from routinized data collection and analysis. Utilization is the final test of any evaluation model. In most of the social services, the history of program evaluations is characterized by scant use in decision making at any level. The field of drug abuse prevention is no exception. Although no systematic analysis has been performed to explore the reasons for this failure, it is commonly assumed that a major cause relates to unmet expectations of the decision makers for whom the studies were intended. The evaluation plan recommended in these Guidelines has a feedback system at its heart. It stresses that program decision makers should be involved throughout the planning and execution of the evaluation, thereby ensuring that their information needs will be addressed. Evaluations that provide periodic feedback in the form of reports that include quantitative data are especially well-suited for this purpose. In addition, the decision maker who helps design the data presentation will be more likely to accept data-based implications.

Feedback loops are one of the distinguishing features of evaluation. Therefore, the pattern and function of feedback loops should be designed or negotiated in advance. Figure 6 portrays a typical feedback loop system for different levels of evaluation.

Figure 6. Evaluation Feedback Loops



The assumption behind a feedback system planned to facilitate program improvement is that the elements and timing of the critical points in the loop should be predetermined to the extent possible. Furthermore, the potential implications of possible negative findings from alternative courses of action should be outlined.

If evaluation results are to be utilized, the organization for which the results are intended must have an internal climate that is receptive to new information. The evaluator therefore has the responsibility to try to develop a climate of receptivity within the receiving organization. This does not mean that the evaluator must be a full-fledged organizational developer capable of transforming an organization that is dead set against incorporating evaluative findings. It does mean that on a limited scale the evaluator is expected to work with program staff to improve the climate for the utilization of the evaluation findings.

A multiplicity of organizational levels usually surround a particular prevention program, and should be taken into consideration when planning an evaluation. The model is intended to help administrators and evaluators appreciate the variety of uses that can be made of the evaluation results if it is planned and designed properly. Modifications in design and activities may occur in subsequent cycles of the process, thus encouraging the use of the findings by decision makers at alternative levels.

In addition, the evaluator should bear in mind the model presented in this chapter includes the influence of external forces upon the decision making activity. Rarely, if ever, are policy decisions made solely on the basis of empirical evidence. Given the current state of evaluation research, a legitimate basis can always be found for questioning policy implications derived from the findings of evaluations. The more attention paid to enhancing utilization, the greater the likelihood for evaluation efforts to have impact on the field of drug and alcohol abuse prevention.

DECISION MAKING ACTIVITY

The evaluator must keep in mind that in addition to the evaluation results themselves, there are a host of other forces which affect decision making. Often these other factors are more influential than evaluation findings. They are the byproducts of the socioeconomic milieu in which decisions are made.

These forces can be categorized in terms of how they relate to the individual decision maker, and to influences inside and outside the organization. (Overlaps obviously exist between these categories.)

Individual forces. The personality and leadership style of decision makers have major impact on the way evaluation findings are accepted. His/her perception of how the organization will accept particular findings, his/her commitment to change both in a general sense and as regards the particular problems addressed by the evaluation, and the persuasiveness which the decision maker brings to the organization all affect an evaluation's potential to bring about change.

Organizational forces. Not only are the individual characteristics of the decision maker important but just as salient are the ways in which s/he is viewed by others within the organization. The perceived power and credibility of the decision maker, stemming from professional authority and personal prestige, will influence the extent to which evaluation findings will be accepted and implemented.

External forces. Extra-organizational forces--essentially those of the community and of funding sources--as well as the general belief system of the prevention field have powerful influences on the degree to which evaluation findings are accepted. Community action can support or hinder program change; inaction on the part of the community reflects a lack of interest, a desire to maintain the status quo, or simply poor community organization directed at the problem area. Obviously, the relative importance of drug abuse prevention as a community issue will have a strong effect on how evaluation findings will be received and acted upon within both the community and the organization.

A full discussion of leadership styles and their effect on organizations may be found in Cartwright and Zander (1968), while a comprehensive discussion of the utilization of knowledge, including annotated bibliographies, may be found in Putting Knowledge to Use (Glaser and Davis 1976). The topic of utilizing evaluation findings and the role of the evaluator in this process is more fully discussed in chapter 10.

PROCEDURAL GUIDELINES FOR THE EVALUATOR

The Guidelines has discussed an ideal evaluation in broad terms. But more specific detail is needed to guide the evaluator through the conduct of an evaluation study. A comprehensive set of such details has been developed by Stufflebeam (1978).* He lists 68 guidelines for conducting evaluation studies, divided into seven classes--conceptual, sociopolitical, contractual/legal, technical, utility, administrative, and moral/ethical. Rather than paraphrase his work, the following is taken directly from the original article:

CONCEPTUAL GUIDELINES

1. Definition -- Achieve mutual understanding among the participants in the evaluation study of the definition of evaluation that will guide the study.

2. Audiences -- Specify the full range of audiences that are to be served by the study.

3. Object -- Clarify what object is to be evaluated (e.g., a teacher, a textbook, a certain federal project, a given technique, a specified drug, or an institution).

4. Purposes -- Clarify the purposes of the study (e.g., selection of persons or groups to participate in a program, allocation of funds, modification of a program, interpretation of program outcomes, or public relations).

5. Questions -- Clarify the audiences' questions (e.g., which practitioners or institutions are in most need of assistance? What areas of an institution or program are most deficient? Should a particular innovative practice be adopted?).

6. Information -- Determine what information (e.g., employer judgments, patient records, students' test scores, and correspondence files) will best answer the audience's questions.

7. Agent -- Identify all those persons whose cooperation will be needed in order to conduct the study.

8. Strategy -- Characterize the evaluation approach that will be used (e.g., a case study, an experiment, a survey, a judicial hearing, or an expert panel).

9. Flexibility -- Insure that the evaluation strategy will allow for the discovery and investigation of new questions as the evaluation develops.

10. Standards -- Define the standards that will be used to judge the study.

SOCIOPOLITICAL GUIDELINES

11. Involvement -- Involve the audiences in planning and implementing the evaluation.

12. Internal Credibility -- Convince those whose work will be evaluated that the evaluation will be conducted and reported in an objective, fair and workmanlike manner.

13. External Credibility -- Convice key external audiences that the evaluation will be conducted impartially and reported completely and honestly.

14. Subject Cooperation -- Secure cooperation from all persons whose participation will be required to successfully complete the evaluation.

15. Communication -- Communicate throughout the evaluation with the staff and all the audiences.

16. Public Relations -- At appropriate times, inform the press and public about the intent, methods, and results of the evaluation.

CONTRACTUAL/LEGAL GUIDELINES

17. Commitment -- Insure that the evaluation findings will be used honorably before agreeing to do the study.

18. Products -- Specify the products and services that the evaluators are to produce.

19. Schedule -- Reach agreement on a realistic schedule for the evaluation activities and reports.

20. Finances -- Reach agreement about a realistic budget and financial constraints.

21. Facilities -- Reach agreement on the office space and equipment that are needed to conduct the study.

22. Personnel -- Reach agreement about who will perform what evaluation functions.

23. Protocol -- Agree on what communication channels will be used and what policies and rules are to be followed in conducting the evaluation.

24. Security -- Agree on procedures for keeping the evaluation's basic information secure from unauthorized use.

25. Informed Consent -- Secure permission from those who are expected to supply personal data to the evaluation.

26. Arrangements -- Reach agreement on any special conditions of data collection that will be necessary in order to meet the evaluation's sampling and treatment assumptions.

27. Editing -- Reach agreement on who will have final editorial authority.

28. Release of Reports -- Reach agreement on who may release the evaluation reports and who will receive them.

29. Value Conflicts -- Reach a clear understanding about how conflicts over the criteria to be used in developing conclusions will be dealt with.

30. Renegotiation -- Define procedures for reviewing and renegotiating the formal agreement (e.g., if there are costs overruns in certain budget categories or if unforeseen factors make it desirable to modify the evaluation design).

31. Spin-off -- Reach agreements about how the evaluation can be used to train evaluators, do research on evaluation, and aid the audience to develop their own evaluation capabilities.

32. Termination -- Agree on conditions for terminating the contract.

TECHNICAL GUIDELINES

33. Investigatory Framework -- Adopt a methodological strategy that is suited to the study purposes.

34. Independence -- Establish procedures for maintaining an independent perspective.

35. Sampling -- Define the population of interest and employ an appropriate sampling plan.

36. Attrition -- Provide for replacing persons who drop out of the study sample in a systematic and representative manner, or at least for keeping a record of the dropouts.

37. Description -- Fully describe the object of the evaluation as it exists and/or evolves during the study.

38. Instrument Validation -- Insure that the data-gathering instruments that are used in the study are valid for the purposes of the study.

39. Data Standardization -- Insure that those data that are to be aggregated are gathered under standard conditions.

40. Data Cleaning -- Remove as many scoring and coding errors as possible from the data before analyzing them.

41. Data Management -- Code, store, and retrieve obtained data according to a systematic plan.

42. Aggregation -- Combine data at those levels that are of interest to the audiences.

43. Computer Program -- Insure that the computer programs used are appropriate and are not in some way flawed in the output they provide.

44. Analysis Plan -- Have a competent statistician verify that the data analysis plan is appropriate and sufficient.

45. Interpretation -- Insure that the norms and standards that are used for interpreting the results of the evaluation are appropriate for the evaluation's purposes.

46. Report Editing -- Edit the evaluation reports before finalizing and publishing them.

47. Complexity -- Do not reduce the problem under study to a false simplicity.

48. Re-analysis -- Make the obtained data available for independent re-analysis.

49. Audit -- Obtain an independent appraisal of the evaluation report.

UTILITY GUIDELINES

50. Audience -- Identify the audiences of the evaluations, stay in touch with them, and periodically update your appraisal of their information needs.

51. Objectives -- Insure that the objectives of the evaluation match the information needs of the audiences.

52. Targeted Reporting -- Address each audience's information needs directly.

53. Precision -- Communicate the findings of the evaluation in clear and precise language.

54. Conclusiveness -- Provide conclusions and recommendations in the evaluation report.

55. Media -- Use that combination of media and methods that will best help each audience understand the evaluation findings.

56. Applications -- Help the audiences to see how they can apply the evaluation findings to their practical situations.

ADMINISTRATIVE GUIDELINES

57. Staff -- Staff the evaluation with qualified personnel.

58. Orientation and Training -- Acquaint the persons who are to participate in the evaluation with their responsibilities, and train them in the procedures required to carry out these responsibilities.

59. Planning -- Plan the evaluation systematically and collaboratively.

60. Scheduling -- Maintain up-to-date projections of what evaluation activities involving what persons will occur at what times.

61. Control -- Monitor and control the evaluation activities so that the evaluation plan gets implemented.

62. Economy -- Monitor and control the use of time and resources so that they are used as wisely and efficiently as possible.

MORAL/ETHICAL GUIDELINES

63. Self-Esteem -- Treat the people whose work is being evaluated with dignity and respect.

64. Comparisons -- Place the evaluation of an object in proper perspective by contrasting its strengths and weaknesses with those of other objects that are in competition with it for funds.

65. Common Good -- Issue reports that reflect the best interests of a free society, but not the self interests of any group.

66. Value Base -- Report judgments that represent broad, balanced, and informed perspectives.

67. Pretext -- Be careful not to allow clients to claim falsely that evaluation results are the basis for prior decisions.

68. Social Value -- Report evaluation results in the context of how well the object of the evaluation meets human needs.

The CONCEPTUAL GUIDELINES are concerned with how evaluators conceive evaluation. These guidelines recognize that evaluations typically are team activities. If the team members are to communicate and collaborate effectively, they must share a common and well-defined conception of evaluation and how they will apply it. In order to develop this common view of evaluation guidelines, the members of an evaluation team should answer ten general questions: What is evaluation? What audience will be served? What object will be evaluated? What purposes will be served? What questions will be addressed? What information will be needed? Who will do the evaluation? What strategy will they follow? How will flexibility be maintained in the evaluation? What standards will be used to judge the evaluation?

The SOCIOPOLITICAL GUIDELINES reflect the fact that evaluations are carried out in social settings and are subject to a variety of political influences. Unless the evaluators deal effectively with the people who will be involved in and affected by the evaluation, these people may cause the evaluation to be subverted or even terminated. The actions that the above guidelines direct evaluators to take in order to avoid and deal with sociopolitical problems are to (1) involve those who will be affected by the evaluation in planning and conducting it, (2) convince those whose work will be evaluated that the evaluation will be done fairly, (3) convince external audiences that the evaluation will be done impartially and reported honestly, (4) secure advance agreements to participate from those who are expected to cooperate in the evaluation, (5) maintain communication throughout the evaluation with the staff and audiences, and (6) conduct a sound program of public relations for the evaluation study.

The CONTRACTUAL/LEGAL GUIDELINES denote that evaluations need to be covered by working agreements among a number of parties, both to insure efficient collaboration and to protect the rights of each party. The underlying assumption is that successful evaluation requires that evaluators, sponsors, and program personnel collaborate. If this collaboration is to be effective, it needs to be guided by working agreements. If these are to hold, they often must be formally constructed, as in a contract or a memorandum of agreement. Such formal agreements should reflect possible disputes that might emerge during the evaluation and should give assurances to each party concerning how these disputes will be handled. The 68 guidelines above direct evaluators to reach agreements with their clients on the evaluation's: (1) purposes, (2) products, (3) schedule of activities and reports, (4) budget, (5) office space and equipment, (6) personnel, (7) protocol, (8) procedures for keeping information secure from unauthorized use, (9) use of human subjects, (10) arrangements for collection data, (11) editorial authority, (12) rules for releasing information, (13)

plan for dealing with value conflicts, (14) rules for modifying the evaluation plan and budget, (15) authority to use the evaluation for research, development, and training purposes, and (16) rules for terminating the evaluation.

The TECHNICAL GUIDELINES reflect the fact that evaluators must solve many technical problems that have to do with collecting and reporting information. These problems pertain to the general investigatory framework, the sources of evaluative information, the instruments and procedures for gathering data, the ways and means of organizing and analyzing the obtained information, and the media and methods to be used in reporting the results. Seventeen specific guidelines have been presented for dealing with these problems.

The UTILITY GUIDELINES are intended to insure that evaluations will be planned and conducted in order to meet the information needs of the evaluation's audiences. These guidelines direct evaluators to ascertain their audiences' information needs; formulate evaluation objectives that respond to these needs; report directly clearly, and conclusively to each audience; use reporting media and methods that best serve each audience; and help each audience to apply the evaluation findings.

The ADMINISTRATIVE GUIDELINES reflect the fact that evaluations are often complex undertakings that require much careful planning and coordination. To help evaluators administer their projects, the guidelines remind evaluators that they should staff their evaluations appropriately, orient and train their staff, plan their activities thoroughly, manage the implementation of the evaluation plan, and be careful to conserve the resources used to conduct the evaluation. Moreover evaluators should implement these administrative guidelines in such a way as to enhance the ability of the client agency to improve its long-range capabilities to manage evaluation studies.

The MORAL/ETHICAL GUIDELINES emphasize that evaluations are not merely technical activities but are also performed to serve some socially valuable purpose. Determining the purposes to be served inevitably raises questions about what values should be reflected in the evaluation. Deciding on value bases also can pose ethical conflicts for the evaluator. The final set of guidelines in the above list direct evaluators to treat all people involved in the evaluation with respect and dignity, present their evaluation of an object in proper perspective, conduct their evaluations so as to advance the cause of a free society, report judgments that reflect the pluralism in society, not allow their evaluations to be used improperly to justify past decisions, and evaluate objects in terms of how well they meet human needs. (The original version of the article from which this section was taken appeared under the title, "Meta Evaluation: An Overview," by Daniel L. Stufflebeam published in EVALUATION & THE HEALTH PROFESSIONS, Vol. 1, No. 1 (Spring 1978) pp. 17-43 and is reprinted herewith by permission of the publisher, Sage Publications, Inc.)

CONCLUSION

The model presents a framework for improving the quality of evaluations. The focus of the model is program evolution--a continual search for improved ways of achieving a specific objective, facilitated by feedback mechanisms. At the heart of the NPERN evaluation process is the reality that the likelihood of producing an effective and useful evaluation is increased when a skilled evaluator works cooperatively with an equally skilled prevention professional. Prevention evaluation should be a multifaceted, incremental, and iterative process.

ENDNOTES

1 This approach borrows heavily from: John D. Waller and John W. Scanlon. "The Urban Institute Plan for the Design of an Evaluation." Working paper 3-003-1. Washington, D.C.: The Urban Institute, March 1973. (Copies may be obtained from either author at the Performance Development Institute, 1800 M. St., N.W., Suite 1025-South, Washinton, D.C. 20036).

REFERENCES

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.

Cartwight, D., & Zander, A. (Eds.) Group-dynamics: Research and theory. New York: Harper & Row, 1968.

Glaser, E.M., & Davis, H.R. Putting knowledge to use: A distillation of the literature regarding knowledge transfer and change. Los Angeles: Human Interaction Research Institute, 1976.

Stufflebeam, Daniel L., Meta evaluation: An overview. Evaluation and the Health Professions, 1978, 1 (1), 17-43.

PART II

INTRODUCTION

Issues pertaining to data collection and measurement are relevant for all levels of evaluation. The selection of indicators, instruments, measures, and data sources in large part depends upon the purpose and modality of the prevention program. One of the most basic considerations in selecting outcome and process indicators is that they reflect the central goals and objectives of the program. For each program objective, at least one outcome and one process indicator should be identified and developed. The care with which this link must be established cannot be overstated. For example, if the program is attempting to change drug-related attitudes, it is important to specify which attitudes are to be changed as well as the activities or services that are intended to bring about that change. Too often, a scale of drug attitudes is selected almost at random without reference to the attitudes the program is attempting to modify.

In the ideal model set forth in chapter 2, goals and objectives that articulate the theoretical basis for specific prevention programs are established in the policy development phase of the program. An alternative, perhaps more typical scenario finds an evaluator in a situation in which the program development process is less formal--with program objectives poorly articulated. In such a case, the objectives must be defined before trying to identify process, outcome, and impact indicators. When faced with having to define program objectives in the course of an evaluation, it is important to review carefully those program components designed to facilitate desired changes. And on the other hand, one's resultant objectives should represent realistic outcomes of the program's activities. There should be clear, logical linkages between activities, objectives, and indicators. When instruments are selected to operationalize indicators, the linkages should extend as far as item content; that is, even items or clusters of items should be linked to activities and objectives.

It is recommended that the evaluator formulate a conceptual framework indicating how program events are hypothesized to effect specific, measurable changes. The framework should be based on the evaluator's own perceptions of the program, the staff's perceptions, and evidence from the prevention literature. The framework should be used by the evaluator and program personnel as a guide to determine which indicators and measures should be included in the evaluation.

In addition to the conceptual framework, the information requirements of the user of the evaluation and the resources available to conduct the evaluation are two other important factors that influence the scope and depth of the evaluation. Since the users' information requirements vary in importance, they should be made explicit, with priorities spelled out.

There are alternative methods that can be used to gather data for specific indicators. These vary considerably in reliability, validity, depth of coverage, and cost. The choice of method should reflect both the priority being given to the indicator and the resources available. For example, if the process by which youths are taken into a community based program needs to be thoroughly analyzed, one may want or need to approach this by means of ethnographic methods which have considerable costs, that is, using highly trained observers. If only a cursory picture of youth intake is required, then a few questions in a process evaluation inventory might suffice. Even within a particular method, great latitude often exists with regard to how extensively and carefully a particular indicator or measure can be studied.

In the following pages is a discussion of indicators, instruments, measurements, and data sources that parallels the trilevel framework of the evaluation model. Process is discussed in chapter 3, outcome in chapter 4, and impact in chapter 5. This format enables the evaluator and the program decision maker to focus on the level most appropriate for their needs.

CHAPTER 3: PROCESS EVALUATION - INDICATORS AND MEASURES

INTRODUCTION

The approach to evaluation in this section is based on the position that in addition to the why's and wherefore's of program success or failure, a thorough analysis and understanding of program operation is crucial to process evaluation.

Process evaluation can be used for a variety of purposes. The continuous feedback it provides is important for internal monitoring, which in turn can be used to guide and direct resource allocation, organizational decisions, and ongoing program development. Externally, process evaluation contributes significantly to accountability and replicability.

In general, process information strengthens an outcome evaluation in at least two ways: first, it can provide clues as to how an outcome evolved and underscore the vital components active in producing the results; and second, along with appropriate control or comparison grouping, it can help to determine whether a particular intervention by itself caused something to happen or if other factors also were involved, and if so, what they were.

To illuminate where and how new factors are introduced into a program, or how prior factors change as a program evolves, process may be studied as an end unto itself. Understanding the influence the program has on participants as both the program and participants change over time may pose problems for the evaluator. Therefore, methods for describing this influence should be carefully reviewed, referenced, and illustrated. Process evaluations are much like case studies, in that a major problem is to decide how to limit the range of inquiry. Boundaries must be placed around what is relevant and useful for the evaluation effort.

The specific needs of the intended user are the most important guides in determining the depth and breadth of a process evaluation. These needs, which should be made explicit in the evaluation plan, will be reflected in its design in terms of the type of information (descriptive, associative, or explanatory) to be collected, and the components of process to be addressed. These components in turn should bear some relationship to the strategy and structure of the program being evaluated.

Unless the explication of program process is recognized as an objective of evaluation per se, the situation that has prevailed in this field in which almost no evaluation adequately portrays what actually takes place in a program may continue. This is particularly lamentable when an outcome evaluation suggests that a program is effective, but there is no clear understanding of what the program is or how it works, so that it can be replicated.

Two issues must be adressed if one is to provide an adequate description of program process. First, one must select indicators, measures, and methods that span the various phenomena that make up a program. The decision as to how to provide appropriate coverage for a specific program must be made in accordance with the special circumstances of that program. For example, a well-balanced picture of a program might have some information in each of the major categories of input and process variables, that is, variables concerning the staff, participants, physical resources, organizational structure, program relationships, and program service delivery.

The second issue to be addressed is that of program characteristics--they must be specified as precisely as possible, and preferably quantitatively. To be effective, a process evaluation design should include methods that afford precision in the specification of process. This is especially true for those indicators and measures deemed the most critical in replicating the program or the most important in secondary uses.

In order to understand what happens in a program, it is necessary to know what goes into one. The framework used to assess inputs should be one that makes it easier to identify variations between or differences among programs and modalities. The clearest way to do this is to build the analysis of inputs from least common denominators. These common denominators, insofar as substance abuse prevention programs are concerned, may be categorized as human resources, physical resources, contextual variables, and program specific variables such as organizational structure, service patterns, and staffing. This is the framework with which we will discuss prevention program inputs and processes.

PROGRAM INPUTS

HUMAN RESOURCES: STAFF AND PARTICIPANTS

An analysis of staff and participant variables is an essential component of process evaluation because of the effect human resources have on program functioning and outcomes. A discussion of relevant variables follows.

The number of staff members and participants is obviously a prima facie indicator of program size. Fulltime workers must be differentiated from parttime and voluntary staff. In each case, all work hours per person should be taken into consideration, as the nature and direction of the program can be influenced strongly by volunteer staff, the ratio of full to parttime staff, and so on. Because staffing patterns will vary throughout the life or cycle of a program, it is important to specify the time or time period for which the staffing pattern applies.

Ascertaining the number of participants may be less straightforward, since program records tend to be less accurate and comprehensive for participants than for staff. Therefore, the evaluator should specify the procedures or assumptions followed in deriving the number of participants. Useful indicators for this purpose include average number of participants per month, numbers of program starters, completers, and dropouts, ratio of program starters to completers and completers to dropouts, and number or percentage of participants completing specific segments of the program. Again, the time related aspects of the enumeration should be specified.

Basic demographic data such as age, gender, marital status, racial/ethnic background, religion, socioeconomic status, residence, family composition, and education have implications for both outcome and process evaluations. Their often-assumed role as intervening variables is important for outcome evaluation, but these aspects of human resources also influence program process. Items for assessing this information can be selected from the demographic data section of the Drug Abuse Instrument Handbook (NIDA 1976). This information is frequently available from standard agency personnel records, but two important cautions should be observed in using it. One is the tendency to make assumptions about the staff and participants based solely on images gathered from the demographic data. The other is an ethical and legal responsibility to respect the privacy of the individual--agency records can be perused only upon consent of the employees or volunteers in question. The so-called Buckley Amendment (P.L. 93-380) has since 1972 protected access to individuals' records in any programs receiving State or Federal funds.

The degree of detail desired for family composition or family background variables may be greater for participants than for staff. The Drug Abuse Instrument Handbook also contains a selection of items and instruments useful in measuring family relationships, as a complement to the "harder" family composition and residence data. The Family Environmental Scale (Moos 1974) is especially well-suited to describe interpersonal relationships among family members.

In assessing the qualifications and special skills of staff members, it is wise to take more than academic background and work experience into consideration. Informal training such as workshops, apprenticeships, hobbies and interests, community involvement, life experiences, independent study, and so on, can reveal abilities which otherwise might remain unnoticed. Of special value to prevention programs is staff experience with or knowledge of drugs, alcohol, and users, along with training in such prevention techniques as values clarification, affective education, behavioral interventions, and healthful alternatives to drugs. Experience in organizing and public relations also can be significant for program operations.

The parallel to staff skills and qualifications for participants would be a set of extra-family behavioral indicators designed to provide information on the experiences participants bring to the program. Of interest in this category are school or community based indicators such as grade point average, class standing, attendance and disciplinary records, group affiliations, employment, leisure activities, and of course, drug use experience.

ATTITUDES, VALUES, BELIEFS, AND KNOWLEDGE

An understanding of, and the distinctions between, attitudes, values, beliefs, and knowledge is crucial to evaluating prevention activities. Beliefs are statements about reality that are accepted as true by individuals based on logic, empirical observation, tradition, faith, or a combination of these. Knowledge can be taken as belief which has been subjected to sufficient empirical testing that the perceived likelihood of the truth of the statement is so great as to preclude its rejection when directly challenged. Programs based on providing information (knowledge) typically measure changes toward beliefs supported by logic or evidence and away from those founded in tradition or faith.

Values are those general elements of belief systems that help to define what is good or right, and which express, through judgment, what is desirable or worthwhile.

Attitudes are learned orientations toward objects or situations which, like value judgments, possess evaluative properties. The difference between values and attitudes is in the greater concreteness and specificity of the latter. Attitudes can be taken as behavioral expressions of values and beliefs. Because of their greater specificity and more direct expression through behavior, attitudes tend to be measured more frequently than are values.

Both operational definitions and the relationships between attitudes, values, beliefs, and knowledge are heavily debated by various schools of thought. One of the major issues involves determination of which of these are dependent and which are independent variables. That is, are attitudes and values shaped by beliefs, knowledge, and behavior, or do attitudes, values, beliefs, and knowedge shape behavior? (See Goodstadt 1975.) The issue is compounded by the fact that all of these concepts must be operationally defined by behavior of one sort or another, whether in response to a questionnaire or through direct observation. Given the fact that there is overlap among them, there is further compounding in terms of the construct validity of any test measuring them. The ongoing nature of the debate notwithstanding, it is recommended that information on drug-related knowledge, attitudes, behaviors, and beliefs be collected for participants and staff as an input in process evaluation.

Participant drug and alcohol attitudes are assumed to be indicative of current or probable substance use, thereby having implications for outcome evaluation. Of interest in process evaluation is the degree to which attitudes toward drug and alcohol usage and users (and drug and alcohol abuse prevention) are shared among participants and staff, and in the larger organization or community in which it is located. Expressed intent to use or not use particular drugs in the future might prove to be a powerful predictor of use, as has been demonstrated with cigarette smoking (Fishbein 1977).

Items which may be of use in measuring attitudes toward drug and alcohol users and usage are the Baer Marijuana Attitude Scale, Suitability for Treatment Scale, Fisher et al. Marijuana Attitude Scale, High School Students' Opinions, Attitudes, Knowledge, and Experience...Concerning Drugs Scale, and the Kandel Study of High School Students--Student Questionnaire, Wave I. All of these are referenced in the Drug Abuse Instrument Handbook. Also useful is "Attitudes Toward Drugs," found in Accountability in Drug Education: A Model for Evaluation (Abrams, Garfield, and Swisher 1973).

Attitudes toward prevention are as important as those toward drug and alcohol use itself. Often these may be inferred from information defining attitudes toward drug and alcohol use. The Addictions Research Foundation in Toronto, Canada, has developed a scale designed to ascertain staff attitudes toward prevention (Goodstadt 1979). Again, it is the extent to which attitudes toward prevention of drug and alcohol abuse are shared by staff and by participants that is most relevant for process evaluations.

It is unlikely that at the outset of a program participants will know much about the particular strategy being employed. However, the staff should be expected to have both knowledge of and attitudes toward the use of specific prevention strategies such as values clarification, decision making, and interpersonal skills development. Positive staff attitudes toward a program's strategy can enhance its effectiveness by ensuring that the program moves along its intended course. Consistency of response and congruence with other personality factors, especially those gleaned from observation and interview, should be considered carefully.

In addition to the above, staff attitudes--toward the community, the participants and their families, other staff members, and administrators--are significant. Studies in the field of psychotherapy have shown that responsibility, perseverance, openness, concern, acceptance, and positive regard are all associated with effective therapy (see Rogers 1961; Rogers and Dymond 1954); these same qualities may enhance the performance of staff members in drug and alcohol abuse prevention programs.

Closely related to attitudes are the values which staff and participants hold. Such global concerns as purpose of life, the meaning of right and wrong, and orientation toward life styles all shade one's perception of others. Views about the work ethic, expression of emotions, use of power, importance of aesthetics, and so forth are also important determinants of program harmony. Additionally, of particular concern to prevention programs are views toward deviance, rebelliousness, self-control, sociability, and security needs. The need for attention to differing attitudes, values, beliefs, and knowledge is brought into sharp relief when considering prevention programs targeted at minority populations. Research suggests that value and belief conflicts between minority programs and "mainstream" funding sources are a major source of stress for programs at all levels (see, for example, Fogel 1975; Kemnitzer 1973; Myers 1979). Developing means of improving compatibility between systems which reflect different world views may be viewed as a major initiative for the field.

Both staff and participants can be expected to bring role expectations to the program that will influence their own performance and that of others. Written program proposals and job descriptions can provide some clarification on intended staff roles. However, these roles may or may not match the role expectations--a mismatch that has serious implications for morale and program functioning. Unrealistic or unrealized performance goals may lead to frustration and self-condemnation, possibly resulting in high rates of staff absenteeism, tardiness, and turnover.

PHYSICAL RESOURCES

Physical resources may often be enumerated in terms such as square feet of space, numbers and types of equipment, supplies, materials, transportation, and so on. They may also be related to program functions. Use of each type of resource can be disaggregated and analyzed, or related either to broad program functions (management/administration and service delivery) or to more specific program activities (tutoring, museum trips, camping) which are performed in order to achieve specific program objectives.

Another aspect of physical inputs important in program evaluation is pattern of use, that is, whether the resource is used once in the course of the program and retained for use in subsequent programs, used once and then discarded, or used frequently throughout a program.

Many of the data required to enumerate and analyze physical inputs may be found in program records. However, additional probing or data collection will probably be required if more detailed matching of resources to specific activities and objectives is attempted. In such cases, program staff or administrators may be able to recall details on resource utilization. Another, more costly alternative, is to design collection activities to measure this aspect of physical resource utilization.

Program records, proposals, and budgets should yield information on program funds, including amounts received, their sources, and intended and actual expenditures. This information is essential for estimating direct program costs. Indirect, in-kind, shared, social, and marginal costs can be estimated from other input information (as discussed in the preceding pages). Cost analysis can be used to integrate analysis of many of the human and nonhuman resource variables. It is a limited integration in that "hard" descriptors (numbers, hours, and so on) of human resources tend to be emphasized. However, if the objectives of the process evaluation include efficiency determinations, then cost analysis is essential. The analysis of inputs in terms of costs also is essential for cost-effectiveness and cost-benefit evaluations.

CONTEXTUAL VARIABLES

The community and institutional environments in which a prevention program operates directly affect its workings and effectiveness. The discussion that follows deals mainly with the community environment; institutional settings are discussed in a later section.

If a formal needs assessment has been conducted during the program development phase and reviewed or updated thereafter, many data on contextual variables should be readily available to the evaluator. The same environmental characteristics that engender a need for prevention services also affect the program's functioning. These relate to the general social, economic, and population structure of a community. Indicators used to describe socioeconomic and demographic structures include the age distribution of the population, its racial/ethnic composition, and the predominance of any subcultures and of power or interest groups. Other indicators are per capita income, unemployment rates by age, sex, or racial/ethnic group, income distribution, major sources of employment in the community, and the percentage of families receiving welfare assistance. Still other indicators of need might be derived from the percentage of single parent families, mobility patterns, education and literacy rates, and general health as indicated by mortality and morbidity rates for key diseases and illnesses. Data for these indicators may often be obtained from special surveys and records maintained by county, city, and State offices that provide human services to the target area. U.S. Census records (disaggregated by neighborhood or enumeration districts) may also prove useful.

The incidence and prevalence of "social problems" as indicated by court or police files on public drunkeness, driving while intoxicated, crimes against property, crimes against persons, and drug arrests all are in their own way helpful environmental parameters that may affect prevention programs and programing.

Arrests for drug use, possession, and sale is obviously a crucial indicator of the extent of the "drug problem" in a region or community. To this can be added medical information (obtained from hospitals and hospital emergency rooms, physicians, and medical examiners--similar to that reported through the DAWN system). The mere presence of crisis or treatment centers provides both an indicator and an awareness of the problem, at least among a portion of the population.

Community attitudes toward youths or groups of people with significant unmet needs, as well as attitudes toward drug and alcohol use, users, and prevention, are directly related to the functioning of a prevention program. Such attitudes can be ascertained by means of community surveys, or inferred from local news articles, records of meetings of community organizations, and so on. The frequency of discussion, the kinds of proposals made, opinions voiced, actions taken--all of these provide information about a community's attitude toward the prevention of drug and alcohol abuse. Even more important from a program's point of view is the extent to which the community and its leaders encourage and support the development and operation of prevention programs.

PROGRAM VARIABLES

If process evaluation is to provide answers to the why's and wherefore's of a program's success or failure, it must have structure. It must be more than sporadic "warm body counts" and reporting of effort. The structure or framework that is recommended in the Guidelines builds on the preceding analysis of inputs. It combines (or borrows) techniques and perspectives from a variety of disciplines. The resulting perspective is a composite. It lies somewhere between that of sociology's treatment of individuals as members of groups and organizations, and management science's emphasis on the product of organizational interaction. The resulting indicators of process are constructed by means of qualitative and quantitative assessments, which in turn are based on measurement techniques that have been developed by the parent disciplines.

When conducting a thorough process evaluation, it is recommended that indicators be constructed using subjective and objective measurements of various types, the goal being to reflect five aspects of a program's functioning. These are:

- Organizational structure and patterns
- Program service delivery
- Program-participant relationships
- Participant-staff relationships
- Staff-Staff and staff-program relationships.

Two additional considerations should be noted. First, each aspect of program functioning can be developed further than is done in the Guidelines. Additional development would illuminate the relationships among the aspects, but it might result in overlap or duplication. In some cases, a component might appropriately be discussed under more than one aspect. Second, "process" denotes something dynamic, something ongoing. Some approaches to evaluating process lend themselves more readily than others to continuing coverage. But even static indicators and techniques can be employed at more than one point in time.

ORGANIZATIONAL STRUCTURE AND PATTERNS

The literature on organization behavior suggests a variety of approaches or models that can be followed to analyze and design an organization. These include variations on the classical pyramid model, the neoclassical organic and behavioral models, and the modern systems based models. The systems based models are most consistent with the approach to evaluation that is recommended in the Guidelines. However, it is apparent that these incorporate many elements that were developed by the classical and neoclassical schools. A prime example is structural analysis.

Interest in organizational structure can be traced to some of the earliest writings on organizational analysis. Structural analysis adresses dimensions such as authority and communication and staff line relationships in terms of their horizontal and vertical characteristics within the organization, and relationships of the organization to superordinate, subordinate, and coordinate organizations. Both formal and informal patterns can be assessed under the guise of structural analysis. However, emphasis has been placed on formal patterns as diagrammed in organization charts. Although the use of organization charts and structural analysis may be considered a static approach to understanding program processes, it provides important benchmark information against which informal and actual program patterns can be measured. These patterns, which are discussed in more detail in subsequent sections, can be analyzed by means of sociometric techniques that utilize observation or interviewing techniques for data collection.

Another aspect of organizations emanating from the classical school of organization analysis pertains to the organization of work in terms of patterns of jobs and degrees of specialization. Again, it should be noted that there may be differences between formal, intended roles and informal, actual roles. The former may be assessed through job descriptions, complemented with interviews with administrators or managers. Actual, or unintended, roles may be analyzed by means of interviews and observation.

Systems based approaches to organization analysis integrate analyses of the formal aspects of organizations with the physical environment in which the organization functions; informal relationships, patterns, and roles; and the personalities and status of the individuals. Linkages among parts of the system are analyzed in terms of communications, balance, and decision making (Scott 1967). Each of these linkages in turn has various dimensions which can be analyzed in detail.

One useful and comprehensive way of describing organizations, using a case study approach, has been presented by Cline and Sinnot (1980). Drawing on previous work by Nielen (1977), they have found it helpful to describe organizations using five interdependent dimensions: task, function, information, fiscal, and personnel.

The task dimension describes the organization as a set of work assignments interconnected by authority and accountability relationships. Major tasks assigned to different units must be identified, along with specific activities undertaken to accomplish each task, as well as the supervisor-supervisee relationships involved in these activities. Data collection is accomplished through interviews, observations of work activities, and existing documents such as job descriptions.

The function dimension describes the organization as multiple operating units interconnected among themselves and the organization's environment by the ways in which they act and react in relation to one another. To examine this dimension, the evaluator determines how the various units within and outside the organization interact by specifying the content of the stimuli and responses that characterize these relationships. Data sources are interviews, observations, and documents such as organizational charts and annual reports. This dimension is distinguished from the task dimension in that the latter focuses on activities within units, while this dimension describes how they interrelate and, in concert, achieve the organization's goals.

The information dimension describes the organization as a structure of decision points connected via data channels. Examination of this dimension is best done through mapping both formal and informal data channels by tapping into a sample of the communication lines between decision points. This is probably the most sensitive and difficult area of this multidimensional approach.

The fiscal dimension depicts the organization as monetary resources connected by budgetary and accounting relationships. Resource allocation within the organization must be determined. In order to obtain details of budget deliberations and decisions, existing documents must be analyzed, and interviews conducted with those who prepare and approve budgets. All financial statements must be carefully reviewed. A major difficulty in describing this dimension is the possibility of "hidden" resource allocations, which might not be easily identified.

The personnel dimension characterizes the organization as a group of persons interacting on a day to day basis. Obviously, data collection of this dimension is best accomplished through observations of interpersonal interactions. Because of the broad range of such interactions, a major task of the evaluator is to limit such observation to the most relevant interactions, choosing appropriate time periods for observations.

The above five dimensions are not exhaustive of all aspects of the structure and function of the organization and, further, are highly interdependent and thus not mutually exclusive. They do provide, however, options for data collection and analysis, covering the major characteristics of the organization.

The major problem for the evaluator in using the above format, or any other in which the organization is analyzed as a case study, is not to determine what is to be examined, but rather what is to be excluded from examination. Given that there is an infinite amount of data which could be collected, the main task is to limit the inquiry--to establish boundaries based on considerations of costs and relevance to the total evaluation effort.

An emphasis in the evaluation model recommended in the Guidelines is on the decision making function within the program. Therefore, one might examine the relationships or linkages between program decision making and information feedback, values of the individ-

uals, resources available, and internal authority (a formal aspect). Each of these components may be expected to change over time, according to program dynamics.

Additionally, understanding of program processes might be enhanced by knowledge of the record or history of the program. Length of a program's existence, factors which led to its creation, size trends over time, past successes and failures, program directions over time--all of these can illuminate the conditions under which the program operates. Sources for such information include program records, interviews with staff, and perhaps the mass media.

PROGRAM SERVICE DELIVERY

This section focuses on the services to be delivered, in particular the rationale for their delivery and the manner and setting in which they are delivered. The information required for evaluating program service delivery is descriptive and may be available from program records, especially if the ideal model is followed during program development and implementation.

Any evaluation of process should include some discussion of the assumptions, concepts, and theory underlying the strategy or modality in effect. As noted in the introduction to the Guidelines, the various prevention modalities correspond to presumed causal or need patterns and measures that can be taken to meet those needs or circumvent the causal forces.

If program development has been documented, information should be available to indicate the rationale for selecting a specific modality or group of modalities for the program being evaluated. If such documentation is not available, the evaluator should ask program personnel to reconstruct development procedures as they apply to the selection or design of a program strategy.

As noted previously the entire program development exercise may be considered to be a component of process evaluation. Chapter 2 discusses program development phases in terms of their implications for the validity of subsequent outcome evaluation. But each of these phases also has implications for process evaluation.

When evaluating process, evaluators should be concerned not only with the selection and design of a strategy but with the decision process that led to the prevention program. This brings us to another aspect of program development-- needs assessment. The need being addressed by a particular program should be directly related to the design of the program's strategy. Of importance from a process evaluation perspective are factors such as how the need was identified. Was a formal needs assessment conducted? When? How? What data were collected from whom? If no formal assessment was conducted, were other available indicators used to infer need? What were they? Who were the principal actors participating in the assessment of need?

The evaluation of services actually delivered by a prevention program can proceed along several dimensions, including timing, content and service integrity. Both of these dimensions may be assessed in terms of the degree of rigor versus flexibility.

With regard to timing it is important to know intended and actual length of service delivery, that is, is the program designed to be a one time presentation for those participating in it? Is it a sequence of offerings extending a week, a month, or a year? How frequently are services delivered (daily, weekly, monthly)? How long is each session? Are sessions scheduled in advance or on demand?

The content or substance of prevention services may be more difficult to evaluate. Content pertains to the issues that are discussed or experienced and the materials used by staff or participants. It also relates to the degree to which the program encourages or discourages individual innovation or interpretation by staff. Also at issue is the integrity of the services being delivered, that is, whether activities aimed at, say, clarifying values or strengthening interpersonal communications skills are actually adressing that variable. Where process evaluation is included as a supporting component in an outcome or impact evaluation, it is still important to understand how the content of service delivery is expected to relate to program objectives.

Also important from the perspective of service delivery are factors such as the institutional and physical settings in which services are delivered, the procedures undertaken to manage program activities, and the kinds of data collection techniques that are being used in the program.

With the exception of mass media campaigns, most prevention programs involve direct exchanges between the providers and receivers of services. Therefore, as has been urged above for different reasons, an evaluation of program processes should include analyses of the relationships among participants, staff, and the program.

PARTICIPANT-PROGRAM RELATIONSHIPS

Participant-program relationships vary widely with the modality of the program. Fortunately, analysis is aided by the similarities that reach across most of the programs.

One such characteristic pertains to the referral or identification procedures through which individuals come to participate in programs. Some referral of sorts may occur when groups are targeted during the needs assessment process. But actual referral may be a procedure or set of procedures similar to intake in other human services programs. In some cases, referral is nothing more than identification based on membership in a specific age or educational group. This would be the case in which a "drug prevention unit" is incorporated in a school health curriculum. In other cases, a need may become apparent for a special type of service provided by a prevention program. Referral may be made by a teacher or other school employe, a peer, or a social service or other public agency. This type of referral may be associated with early intervention alternatives or affective education programs. And, participants may self-refer. The referral component of process is of interest because of its potential for influencing outcome as well as for its relationships to other components of process.

It is unlikely that participants will be privy to the objectives of prevention programs. Nevertheless, it is important to ascertain their expectations from the program, regardless of the referral procedure. In an evaluation of a school based early intervention program in Atlanta, researchers at Emory University's Center for Research on Social Change investigated expectations and actual experiences by means of a 20-item questionnaire administered after the participants finished the program. The items corresponded to possible needs situations such as peer and family relationships, school work, legal problems, and health related issues including problems with drugs.

The participants were asked to rate each factor twice. The first was to describe self-perceived needs when the program was joined, the second to describe the participant's perception of the help actually given through the program. Each item was rated on a scale of zero to five where zero represented a little need or help and five represented a lot.

Participation varies among the individuals in a program in terms of time and quality of participation. Attendance records may provide information on the extent of participation but not on the reasons for participation. If this aspect of process is considered important, attendance data should be supplemented with interviews. This may yield reasons for attrition or for "poor" or "low" levels of outcome. There are degrees of intensities of participation which range from participating because it is a requirement or the "lesser of two evils," to having a strong degree of attachment to the program. Information useful for assessing degrees of participation can be obtained by means of staff reporting, direct observation, and participant questionnaires.

As has been noted previously, relationships between knowledge, attitudes, and behavior are not clearly understood. Cohen (1967) concluded that participants' use of drugs could be altered, depending upon the information acquired during a program. Fifty percent of Cohen's participants said that they would discontinue using psychedelic drugs if they were aware of scientifically proven harmful effects. In a subsequent study, Swisher and Crawford (1971) found that although participants gained an understanding and knowledge of drugs and their effects, their attitudes toward use of drugs were not altered significantly. These and other studies have led to changes in program design and presentation. And, as is illustrated in the outcome chapter, most prevention programs are operated on the assumption that the services being provided will have a "favorable" impact on knowledge, attitudes, and behavior.

Emrich (1979) has suggested that the link between the information or experiences of prevention programs and the resultant attitudes or behaviors is the impressions formed by the individual participants. Impressions are influenced by prior knowledge and experiences, awareness, and the content and format of presentations. To identify impressions created by prevention programs, Emrich has used a micro-ethnographic measurement technique which is based on participant repsonses to a time related sequence of still photographs of program activity. By asking the participant a series of questions designed to elicit his/her thoughts and feelings about program activities, the observer can establish a subjective assessment of the activities, compare apparent degrees of awareness among participants, and explore whether different outcomes are associated with varying degrees of awareness.

PARTICIPANT-STAFF RELATIONSHIPS

Participant-staff relationships have both qualitative and quantitative dimensions. Quantitatively important indicators include frequency and duration of contacts, and the number and ratio of planned to unplanned contacts and formal to informal contacts. All of these provide vital process descriptors. To these should be added the qualitative dimensions of these relationships. Factors such as purpose of the contact, its content, topic, and dynamics will provide information on what transpired. Below are typical micro-level indicators of participant-staff relationships: the context in which contact was made (small group, large group, or one to one), the location in which the contact occurred, the party initiating the contact, and actions taken by the respective parties. And what was the outcome of the contact? Was the contact cooperative or antagonistic? Did the parties express anxieties that were or were not alleviated by the contact? Or did they express pleasure as a result of the contact? Data for these parameters can be collected relatively routinely by staff reporting or interaction analysis. The structure of the data should be tailored to the program in question, reflecting the strategic assumptions on which it is based.

STAFF-STAFF AND STAFF-PROGRAM RELATIONSHIPS

Staff-staff relationships can be analyzed in terms of communications, working relationships, and balance or agreement on program-related issues, procedures, and objectives. Dimensions of communications appropriate for process analysis include its form (written, spoken), its intended audience (one or several persons, the entire staff), parties to the communication (administrators, managers, office staff, service delivery staff), pattern of communication (chain, circle, wheel, Y), the initiator of communication, intent of communication (transmit information one way, stimulate dialogue, encourage feedback), content of communication (cognitive, normative, or affective information), amount and frequency of communication, and whether it is formal or informal.

Price (1972), in a review of measures used to study organizations, noted that although communication is considered to be one of the most significant elements of organizational interaction, its measurement is relatively poorly developed. Questionnaires, self-reporting forms, and observation--especially by means of interaction analysis--are the principal techniques recommended for collecting data to evaluate communications.

These same techniques are well suited for collecting data to analyze working relationships. In an evaluation of process it is important to know who works with whom to perform what tasks, with whom do the staff consult for advice on program-related problems, and who are the informal leaders among staff. This approach to analyzing staff-staff relationships blends the concepts of group role and functional relationships. As with other dimensions of process, actual relationships should be compared with intended relationships.

There are many concepts that relate to balance at the level of the individual and the organization. The use of balance as a device for gaining perspective on staff-staff relationships within a program involves an assessment of the extent to which there is harmony among the staff. Key points are:

- Attitudes and motivations of the individual
- Their formal status, roles, and objectives

- Program issues, procedures, and objectives.

Scott (1967) suggests that the above elements of balance are interdependent. This concept of balance includes staff-program as well as staff-staff relationships. Other indicators of staff-program relationships include absenteeism, turnover and burnout rates, amount and type of training received as part of the program, and the discrepancies between formal job descriptions and actual roles in the program.

Absenteeism may be measured quantitatively in terms of frequency and length of absence, and qualitatively in terms of reasons for absence. Turnover rates may be indicated by the ratio of changes in staff to total staff for specified periods of time. Equally important are the reasons for turnover. Exit interviews might yield important information for planning strategies to reduce turnover. Turnover because of burnout is not uncommon with intensive prevention programs. One administrator, sensitive to this phenomenon and its potential in her program, schedules one day per week for both on the job training and trouble shooting. The agenda varies from week to week but always staff members are encouraged to share their experiences and problems with other staff members as a way of gaining support and seeking solutions.

As the relationship between actual program roles and intended program roles can be a source of friction, it is worth one's while to study formal job descriptions, organization charts, and interviews with program administrators to see what the intended program roles are. Then, through observation, program records, and interviews find out what the actual roles are, that is, what is really happening with the staff and the program.

Interviews and observation techniques can be used to assess the extent to which there is harmony or conflict among program elements, as well as to assess absenteeism, turnover, burnout, on the job training, professional development, and support services. Apparently no prevention process evaluations have attempted an analysis in this great detail. Consequently, an evaluator contemplating an exercise of this scope is referred to the organization evaluation and analysis literature for instruments and measures that may be appropriate for gathering data for these process elements. Two especially relevant volumes are Measures of Occupational Attitudes and Occupational Characteristics (Robinson, Athanasiou, and Head 1973), and Handbook of Organizational Measurement (Price 1972). These are discussed in the appendix to part II.

ENDNOTES

[1] The above discussion paraphrases the work of Cline and Sinnot (1980), included here with the permission of the authors.

REFERENCES

Abrams, L.A., Garfield, E.F., and Swisher, J.D. (Eds.). Accountability in drug education: A model for evaluation. Washington, D.C.: Drug Abuse Council, Inc., 1973.

Cline, H. and Sinnott, L. What can we learn about organizations? In S. Ball (Ed.), Program evaluation. San Francisco: Jossey-Bass, 1980, in press.

Cohen, A. LSD and the student: Approaches to educational strategies. Paper presented at the National Association of Student Personnel Administrators Drug Education Project Regional IV Conference, February 1967.

Emrich, R.L. Personal communication, May 1979.

Fishbein, M. Consumer beliefs and behavior with respect to cigarette smoking. Report to the Federal Trade Commission, May, 1977.

Fogel, F. R. Language and thought: An investigation of social class differences. Unpublished manuscript, University of Houston, 1975.

Goodstadt, M.S. Myths and methodology in drug education: A critical review of the research evidence. Substudy 588. Toronto: Addiction Research Foundation, 1975.

Goodstadt, M.S. Attitudes toward prevention. Paper presented at Prevention Seminar, School of Addictions, Addictions Research Foundation, Toronto, Canada, April 1979.

Kemnitzer, L. S. Adjustment and value conflict in urbanizing Dakota Indians. American Anthropologist, 1973, 75, 687-707

Meyers, V Survey methods and socially distant respondents. Social Work Research Abstract, 1979 15, 3-9.

Nehemkis, A., Macari, M.A., and Lettieri, D.J. Drug abuse instrument handbook: Selected items for psychosocial drug research (NIDA Research Issues Series No. 12). Washington, D.C.: U.S. Government Printing Office, 1976.

Nielen, G. C. Foundations for a curriculum in "large systems." In R. A. Buckingham (Ed.), Education and large information systems. Amsterdam: North Holland Publishing Company, 1977.

Price, J.L. Handbook of organizational measurement. Lexington, Massachusetts: D.C. Heath and Co., 1972.

Robinson, J., Athanasiou, R., and Head, K. Measures of occupational attitudes and occupational characteristics. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1973.

Rogers, C., On becoming a person. Boston: Houghton-Mifflin, 1961.

Rogers, C. and Dymond, R. (Eds.). An overview of the research and some questions for the future. In Psychotherapy and personality change. Chicago, Illinois: University of Chicago Press, 1954.

Scott, W.G., Organization theory. A behavioral analysis for management. Homewood, Illinois: Richard D. Irwin, 1967.

Swisher, J.D., and Crawford, J. An evaluation of a short-term drug education program. The School Counselor, 1971, 265-272.

CHAPTER 4: OUTCOME EVALUATION - INDICATORS AND MEASURES

INTRODUCTION

This chapter emphasizes indicators of outcomes in relationship to prevention program objectives. These objectives include intermediate objectives, which are those that start the causal process or movement toward ultimate objectives, the desired long-term effects. For instance, a program might have the intermediate objective of improving social skills, which is expected to causally relate to the ultimate objective of preventing drug or alcohol abuse. The chapter begins with a discussion of objectives for prevention, shifts to a discussion of ultimate objectives related to drug and alcohol use and attitudes, and reviews intermediate objectives in terms of available instruments.

OBJECTIVES

SPECIFYING PROGRAM OBJECTIVES

One of the major difficulties encountered in attempting to establish an outcome evaluation is the development of a clear statement of objectives by the program staff. It is axiomatic that an outcome evaluation cannot proceed unless adequate objectives have been developed. Consequently, it will frequently be necessary for an evaluator to spend time with the staff in an attempt to articulate measurable outcomes for a program. The role of the evaluator in this situation is to facilitate the preparation of objectives rather than attempting to interpret program intentions and impose objectives on the program staff. Mager's (1962) text on developing objectives indicates three areas of focus that an objective should address. They are:

- Desired behaviors (for example, a decrease in drug use)
- Specific circumstances (for example, after three months in an alternatives program)
- Level of performance (for example, sustained at a one year followup).

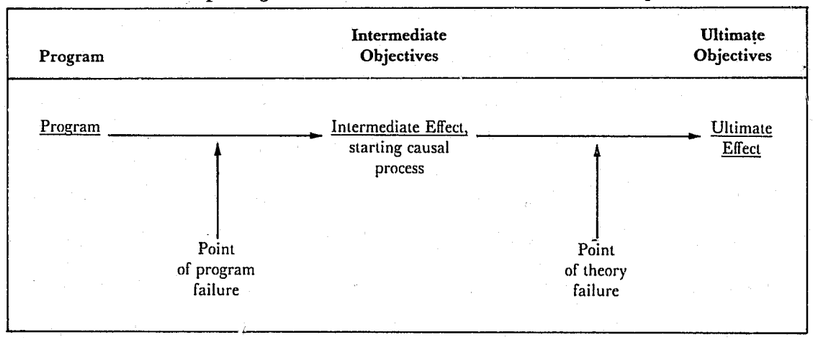
As Suchman (1969) has pointed out, the problem of specifying the objectives of a particular program is more complex than one might think. Most programs are designed to produce "intermediate" effects on the program recipients. The assumption made by the program is that these intermediate effects (for example, improved self-concept) are both necessary and sufficient conditions for achieving the program's ultimate goal of reduction of drug and alcohol abuse.

A program may appear to be ultimately unsuccessful for two reasons:

- The program may fail to produce a unique short-term effect, thus failing to initiate the process leading to the attainment of the long-range goal.
- The program may succeed in bringing about the unique short-term goal, but this proximate effect may not be sufficient to bring about achievement of the long-range objective.

The first state of affairs may be considered to be a true failure of the program, while the second is a failure of the theory of the assumption(s) underlying the program. There is of course a third possibility: the program may fail to bring about the proximate objective but nevertheless succeed in bringing about the desired long-range effect. Table 1 portrays these three possibilities in the context of intermediate and ultimate objectives of drug and alcohol abuse prevention programs.

Table 1. Interpreting Outcomes in Relation to Intermediate Objectives



The point of table 1 is to emphasize the importance of assessing the intermediate objectives as well as the ultimate objectives in any prevention effort. With each type of prevention program, certain specific intermediate objectives are expected to result in the achievement of the ultimate, or long-range objective (reduction of alcohol and drug abuse). Thus, the first and fundamentally most important task of a prevention evaluation is to identify the objectives of the program. In practice, it is rare that any drug or alcohol abuse prevention program has only a single objective. As Weiss (1972) and others have pointed out, some programs have as their objective sheer survival, at least in their initial stages. Further, some programs' public or "manifest" goals may be quite different from their real or "latent" goals. Various program staff members may have different conceptions of their program's goals. Finally, the degree of support of a given goal on the part of staff members varies. Despite these problems, the fact remains that a program can be evaluated only if its objectives are clearly known. It is surprising, then--and obviously very significant--that many programs fail to fully identify their objectives.

OUTCOME INDICATORS

The outcome indicators fall into two broad categories: measures of drug and alcohol use, the primary and virtually essential class of indicators for any program type; and measures of adjustment, including personal, family, and peer adjustment. These reflect most of the personal and social correlates of drug and alcohol abuse, many of which are frequently woven into intermediate program objectives. Also included are indicators of performance and behavior. The emphasis in this category is on activities in school, social/recreational pursuits, and discipline.

ULTIMATE INDICATORS

Ultimate indicators of drug and alcohol use are appropriate across all six program types. Outcome indicators of drug and alcohol use are typically self-report measures. It is unfortunate that drug use, the central outcome criterion--and a sensitive one at that--is measured by self-reporting techniques, which are vulnerable to reactive bias. However, methodological research on the collection of drug use prevalence data (for example, Hurst, Cook, and Ramsay 1975; and Abrams, Garfield, and Swisher 1974) has shown self-report measures of this type to be reliable and valid.

INTERMEDIATE INDICATORS

Intermediate indicators comprise most of the measures of program outcomes, generally thought to be linked in some way to drug abuse. Indeed, most drug and alcohol prevention programs (with the exception of purely informational ones) devote a large proportion of their time and resources to one or more of these areas of adjustment, in the belief that some minimum level of adjustment/coping skills of the target group is required to prevent alcohol and drug abuse. As mentioned above, the literature generally supports this view, although the direction and specific nature of these influences are neither simple nor clear.

Attitudes toward drug and alcohol use are too often substituted as indicators of the ultimate behavior because of the perceived difficulty of measuring use through self-report or other means. This substitution rests on the problematic assumption that attitudes, expressed verbally or as responses to pen-and-paper instruments, are extremely valid indicators of use or intended use. Although attitudes and use have been shown to be correlated (Fejer and Smart 1973) this does not necessarily warrant the substitution of one indicator for the other. In fact, positive attitude change at program completion without concomitant behavioral change can create enough internal conflict to result in reversion to former attitudes or, of course, to later reduction in drug use. So far, no definitive study has adressed this issue in prevention.

The following list of intermediate indicators contains "attitudes," rather than only "attitudes toward drug and alcohol use," to emphasize that the measurement of attitudes toward other indicators is also relevant to an evaluation. For instance, attitudes toward criminal activity or teachers, especially when accompanied by other measures, can aid in increasing the social significance of findings. The major source for the following list of intermediate indicators is Lettieri (1975), which contains comprehensive discussions of various indicators and references a range of studies which demonstrate their relationships to drug and alcohol use.

- Attitudes
- Intentions to use
- Personal development
  - Self-concept
  - Responsibility
  - Moral development
  - Alienation
  - Decision making
  - Locus of control
  - Values clarification
  - Social interaction
  - Achievement motivation
- Family interaction
- Peer interaction
- Knowledge about drugs and alcohol
- School performance
  - Academic
  - Activities
  - Teacher interaction
- Criminal activity
- Social/recreational activities

For most programs beyond the informational level, personal adjustment indicators are quite relevant. The appropriateness of family and peer interaction indicators depends on the thrust of the particular program. Typically, programs strive to improve the quality of family relations. Both peer and family relations are measured, most often through instruments administered to the participant and sometimes, in the case of the family indicators, to a parent or sibling as well.

With the exception of data on social or recreational activities, indicators for schools and community are based on the records of schools and criminal justice agencies. Social and recreational activities usually are assessed by self-report checklists.

The above list is far from exhaustive. The evaluator must be highly attuned to program activities and stated objectives to ferret out other meaningful and relevant indicators. As an example, the "Outward Bound" type of program emphasizes the development of self-reliance, which has been shown to be negatively correlated with drug abuse (Segal 1975). The evaluator must also be sensitive to a host of culturally relevant indicators which may be unique to minority oriented programs.

An attempt has been made to review instruments for each of the program types and for the ultimate and intermediate objectives. These reviews are found in the appendix to part II.

DATA SOURCES

Once program objectives are fully identified, the next critical task facing the evaluator is to link them to observable and measurable criteria. It may be that important objectives are too often overlooked, but at least as often evaluators fail to identify variables or methods by which program objective achievement can be measured. The section below describes some of the major sources of data useful to this aspect of drug and alcohol abuse prevention program evaluation.

A wide variety of sources may be tapped for indicators and measures of program effects. Weiss (1972) lists 15 distinct sources as follow:

- Interviews (with the recipients themselves or with peers, teachers, etc.)
- Questionnaires
- Observation
- Ratings
- Psychometric tests of attitudes, values, personality, preferences, norms, beliefs, and other psychosocial variables
- Institutional records (e.g., grades, discipline records, truancy records)
- Government statistics (e.g., local police file on drug related incidents)
- Tests of information, interpretation, skills, application of knowledge
- Projective tests
- Situational tests presenting program recipients with simulated life situations
- Diary records (of school personnel, parents, etc.)
- Physical evidence
- Clinical examinations
- Financial records (e.g., expenditures of a school district for vandalism repair)
- Documents.

All of these sources represent different methods of assessing a program's attainment of objectives. In practice, the more nonreactive or "unobtrusive" measures are not used as frequently as they might be. (See Webb, Campbell, Schwartz, and Sechrest 1966 for a general discussion of these kinds of measures.) In practice, verbal or questionnaire type self-reports are the kind used most frequently in drug abuse prevention.

ISSUES IN PREVENTION MEASUREMENT

SELF-REPORT, ARCHIVAL, AND OBSERVATIONAL MEASURES

The various indicators available to prevention evaluators may be broadly categorized into self-report and observational measures. The pros and cons of each of these categories are discussed below.

The most commonly used indicators in prevention evaluation are self-report measures. These include written instruments, and the less common face to face or telephone interviews. Such measures are relatively easy to obtain and score, and their generally good reliability is an attractive feature. However, detractors argue that self-reports rely on a presumably weak assumption that respondents are willing or able to accurately report their own feelings, attitudes, past behavior, or behavioral intentions. It is further argued that the problems with self-report are compounded when the issues under consideration are sensitive, such as drug or alcohol use.

In general, the case against using self-report data has been somewhat overstated. While it is true that respondents will distort or even falsify self-reports, a number of techniques exist for improving the general validity of self-report data. These include:

- Impressing respondents with the importance of the evaluation, and especially with the importance of their contribution to it.
- Establishing a trusting and open relationship with respondents before testing begins.
- Clearly establishing that individual responses will be kept confidential.
- The use of consistency checks to detect false responses.

Without such safeguards however, the validity of self-reports, especially of sensitive behaviors, will indeed be suspect.

Observational methods subsume two broad categories of indicators--behavioral observation and archival or records data. The latter include school attendance, grades, criminal records, referral records, and so on. These data are generally collected for some reason other than evaluation and are usually not under the direct control of the evaluator. As such, archival data are subject to numerous biases, errors in reporting, differences in interpretation, and general sloppiness which may not be known or knowable. Eber (1975) provides a delightful example in which patients' stated age and age as calculated from birth date correlated .91.

By contrast, behavioral observation is generally undertaken solely for the purpose of evaluation and is usually under direct evaluator control. Observations may be made of classroom behavior, teacher behavior, communication skills, interpersonal skills, and so on. Problems of observer bias may be adressed through multiple raters and inter-rater reliability coefficients, although the bias introduced by being observed can, in principle, never be adequately adressed.

A third class of measures sometimes used by treatment programs is physiological indicators. These indicators rely on the detection in blood, urine, or saliva, of traces or byproducts of various substances, including nicotine, opiates, alcohol, barbiturates, and amphetamines. Prevention evaluators may be tempted to use such measures, owing to their high reliability. However, the primary criterion of the selection of these measures, as all others, is the relevance to the prevention program in the context of the program's total environment. For example, programs which emphasize the development of trusting relationships between participants and staff might find such measures ill-advised.

MEASURES OF DRUG AND ALCOHOL ABUSE

The purpose of any prevention program is to delay or prevent drug or alcohol abuse. Accordingly, there should be no question that the major outcome indicator is consumption

patterns and the negative consequences of use. However, goals and objectives related to changes in consumption patterns may be different for drug programs than for alcohol programs, and may differ for different drugs within the same program. For example, most alcohol programs aim to prevent use related problems such as alcohol related arrests, health problems, and so on rather than use per se, and drug programs may have different objectives for marijuana than for heroin. The evaluator must take particular care to assure that consumption measures are appropriate to program objectives.

PROTECTION OF HUMAN SUBJECTS

An issue which the evaluator needs to consider when administering interviews is the protection of human subjects. DHEW has regulations (45 CFR, Subtitle A, Part 46) governing the rights of human subjects in activities supported by grants (except formula grants) or contracts. Of particular importance are the procedures to be followed in obtaining subjects' informed consent. Informed consent is defined as:

> The knowing consent of an individual or his/her legally authorized representative, so situated as to be able to exercise free power of choice without undue inducement or any element of force, fraud, deceit, duress or other forms of constraint or coercion (ADAMHA 1975).

Basic elements of information necessary to informed consent are:

- A fair explanation of the prevention program and its objectives, the procedures to be followed and their purpose, including identification of any procedures which are experimental.
- A description of any attendant discomforts and risks reasonably to be expected.
- A description of any benefits reasonably to be expected.
- A description of any appropriate alternative procedures that might be advantageous for the study subject. For example, if study subjects are receiving services in the prevention program, describe appropriate alternative service, if any, to which such subjects could be referred if they choose to discontinue participation in the project.
- An offer to answer any inquiries concerning the procedures.
- Instructions that study subjects are free to withdraw consent and to discontinue participation in the project at any time without prejudice.

Another reference on the issue of human subjects is Ethical Principles in the Conduct of Research with Human Participants (APA 1973) which includes principles of ethical responsibilities, examples of situations where principles might apply, and discussion of application of the principles.

SAFEGUARDING ANONYMITY

Because of moral, legal, and reliability considerations, it is absolutely necessary to guarantee the anonymity of subjects and to spare them self-incrimination. Thus, where a pre and posttest design is to be done, various code systems may have to be established to facilitate anonymity. One useful, but unreliable, system involves assigning a number generated to a subject as a function of given characteristics (such as birthdays, parts of social security numbers, nicknames converted to number codes, street numbers, and so on). Though the code sounds simple, it tends to cause data losses when subjects are asked to reproduce it on subsequent occasions.

Another system involves developing a single master list that has code numbers and names. Theoretically, such a list is only accessible to the evaluator, but the subjects may not be convinced; furthermore, such records could be subpoenaed as part of a legal proceeding. Fearing trouble, the subjects may totally distort their involvement with drugs, and thus subvert the evaluation. So, in order to both safeguard privacy and to improve the quality of an evaluation, assurances must be made to the subjects that adequate safeguards have been taken. Efforts are being made at the Federal level to provide some legal immunity to researchers and evaluators in this field.

One way to protect anonymity and possibly reduce response bias when asking sensitive questions is to use random response techniques on questionnaires. This approach protects the anonymity of the question rather than the respondent. In one of the simplest models, two questions are presented--the sensitive question and an innocuous question for which the probability of response is already known. Respondents are asked to choose a question by, say, flipping a coin, and then to respond without letting the interviewer know which question is being answered. Given prior knowledge of the probabilities of question selection and responses to the innocuous question, estimating the proportions of group responses to the sensitive question is straightforward.

Many other models have been developed, including some which allow responses to frequency of sensitive behaviors. The major drawback to all methods is that only aggregate data are obtainable, preventing further analysis at an individual level. Further, larger sample sizes are required because the obtained variance is a function of the proportion of the sample responding to the sensitive question rather than of the entire sample. (French 1979; Fox and Tracy 1980).

EXAMINER-RESPONDENT INTERACTION

Precautions should be taken to minimize examiner influence in data collection. The relationship between the examiner and examinees has always been considered critical in the administration of any type scale (for example, see Anastasi 1976). In soliciting answers to substance use scales, Horan, Westcott, Vetovich, and Swisher (1974) randomly assigned subjects to assessment conditions in which subjects were either anonymous or identified by name by the examiner. Subjects identified by name claimed significantly less drug use than subjects who remained anonymous.

Whereas it is important for the examiner to have the trust of the subjects, it is equally important that s/he not be personally acquainted with them. Aspects of the experimental situation (the experimenter's appearance and behavior, the content of instructions given to subjects, features of the physical environment) can systematically bias subjects' behavior. Especially when the behavior in question is controversial, subjects may be sensitive to situational cues that label the behavior acceptable or unacceptable and vary their answers accordingly. Cues having such effects are referred to as demand characteristics. Obviously, when such influences occur, findings may be accurate for the specific situation but unrepresentative (that is, lack external validity). Demand characteristics can be lessened by using appropriate instructions to subjects and employing unobtrusive measures whenever possible. Alterations in test administration can also influence subjects' performances, so these too should be guarded against by preparing examiners and selecting adequate testing conditions.

Advance preparation of examiners is necessary in order to ensure uniformity of procedure. Examiners should memorize verbal instructions exactly so that they can present the test in a natural, informed manner without hesitation or misreading. Likewise, test materials should be available to the examiners prior to administration of the test.

Attention to details of testing conditions is important, as minor aspects of the test environment can alter subjects' performances. This requires selection of a suitable room free from distractions, with appropriate lighting, ventilation, seating, and working space. A procedure to prevent interruptions should be implemented. Any unusual testing conditions should be recorded and taken into account when interpreting test results.

CONSISTENCY SCORES

Self-report instruments should contain consistency scales. Some individuals, due to peer pressure (see Shute 1976), will either amplify or underestimate their levels of use. It is therefore recommended that all scales contain dummy drugs which are never used or if

used would be deadly (for example, curare--a paralytic drug sometimes used in surgery where artificial life support systems are available). Positive responses to dummy drugs allow for the elimination of these subjects. However, there is no known method for determining whether subjects have underestimated. One solution is to ask subjects how honestly they completed each questionnaire, but then they may lie about having lied in the first place (Swisher and Crawford 1971).

WEIGHTED SCORING

Composite scores can be developed for data reduction of self-report drug and alcohol use scales. One developement in the realm of scaling measurement of use has been the assignment of weights to the various products consumed. It is relatively easy to obtain agreement among experts that once a day use of marijuana is less harmful than the same use of a barbiturate. However, it may be impossible to obtain a consensus, say, that three drinks of whiskey a day is more or less harmful than a comparable use of barbiturates.

A simple scheme such as assigning a proportionate weight to various substances has been used in prevention evaluations (for example, Swisher, Warner, Upcraft, and Spence 1973). In this instance heroin was given an arbitrary weight of two, barbiturates one and a half, marijuana one, and so on. The derived score is the sum of the weights times the extent-of-use index. The weights assigned can vary according to the unique concerns and consequences for a particular client population. However, it is important to establish some system of weighting as part of the evaluation process.

Another, more complicated technique was developed by Gunderson, Russell, and Nail (1973) who rank ordered eight classes of drugs according to perceived severity of psychological and physiological effects, and then modified these scores by considering individual usage rates, methods of use, age of onset of use, and, finally, duration of use. By summing scores, composite scores reflecting total drug involvement were obtained. Their findings suggest that the involvement scale related to various aspects of drug rehabilitation in a meaningful way and may have potential value for various clinical and administrative purposes.

Lu (1974) developed a composite score by deriving sets of appropriate index weights for each state of use (that is, extent of involvement) for each drug type based on the distribution of data from a sample. A user has an assigned weight for each drug used, considering extent of use, and the user's drug-use index is the sum of the weights for the individual drugs. This index inherently places more weight on drugs which are used less frequently in the sample.

Pandina, White and Yorke (1979) have extended Lu's work by including (1) extent of use, (2) frequency of use, and (3) recency of use, and combining these factors into a composite score. Again, the user's drug-use index is the sum of weighted scores. The exact procedure is detailed in their study.

All of the above weighting systems can be easily modified to incorporate alcohol or tobacco consumption. Marlatt (1978) reviews a variety of issues regarding self-reporting of alcohol consumption and suggests several unique approaches to overcoming some of the problems in this area.

As in all other cases, reduction of drug and alcohol use data to composite scores loses information. Obviously, such techniques are inappropriate if the evaluation seeks to examine interactions between specific kinds of drug use and program interventions. However, as general measures which can be used as dependent variables in outcome studies, rationally derived composite scores are useful, particularly in the common situation where both low proportions of subjects are actually users and where a wide variety of use patterns exist, such that analyses which attempt to distinguish or classify on these bases would require extremely large sample sizes.

DRUG ABUSE

One of the pitfalls in assessing extent of drug use is to equate use with abuse. Unfortunately there is no standard definition of abuse, and various positions seem to reflect

abuse is use other than as prescribed. It behooves the evaluator and the program personnel to define abuse. The most acceptable concept of abuse today appears to be that use of drugs which interferes with physical health, psychological functioning, social adaptation, educational performance, or occupational functioning. No really adequate definition of abuse is available at this time, but it is imperative that consideration be given to this concept and that an operational definition be stated as part of the evaluation.

Some of the numerous intermediate outcome indicators which are valuable measures of a program's effectiveness have been listed, and there are several reasons for employing them in addition to drug use indicators. The legitimacy of these indicators (for example, self-esteem) does not hinge solely on the existence of causal connections between the indirect indicator and drug use or alcohol misuse. Their legitimacy can also be established by (1) the existence of program goals and objectives which seek to change such intermediate attitudes or behaviors, and (2) credible evidence of an association between drug use and the particular indicator. For example, if the program has as an intermediate objective the improvement of a client's self-esteem, and if some program activities are directed toward that objective, the use of a measure of self-esteem as an outcome indicator is entirely appropriate. (With regard to the latter point, it is necessary to demonstrate the existence only of consistent, predictable associations between an indirect measure and drug use, not the presence of a causal relationship between the two.) If the literature provides evidence of a linkage which is sufficiently significant, then the intermediate measure is an appropriate indicator. It is likely that constructs such as self-esteem, attitudes toward deviance, alienation, attitudes toward family or personal responsibility are not linked to drug use in any neat, predictable chain but rather change in concert with drug use (Gorsuch and Butler 1976). Partly because attention and concern is focused on drug use, there is a tendency to view it as the terminal event in a sequential chain rather than simply as another behavior occurring as part of a broad set of changes.

TARGET GROUP CHARACTERISTICS

There are certain characteristics of a client population or target audience that can place several constraints on the methods of measurement. In particular, evaluators must exercise care in the use of paper and pencil instruments. For example, the reading level of an instrument may be too high for a particular client population. If a portion of the population is functionally illiterate or just marginally literate, no instrument should be used. More likely, the client population will consist of adolescents (ages 10-18) who can read, but whose reading abilities may not be very high.

A related problem, often overlooked, concerns the conceptual level of an instrument. A client may be able to read and understand individual words and terms of an instrument but may be unable to comprehend or relate to the concept that is being addressed. Instruments that contain a number of sophisticated abstractions ("spiritual meaning of life," "political justice," "world order") may not translate well to many client populations. On the other hand, efforts to reduce the reading and conceptual levels of an instrument can go too far. The instrument must be appropriate for the social and maturational level of the clients.

If published instruments are being considered, make sure that the manual describes the characteristics of the population with which the instruments have been used. To be safe it is advisable to pretest any instruments (whether published or developed inhouse) on a small sample (say, five clients) to determine whether they are understood and viewed as relevant.

Another obstacle to the use of pencil and paper instruments concerns general feelings about "taking tests." Many young people and certain ethnic groups have developed negative attitudes toward instruments, viewing them as alien tools designed to label them as failures or deviants. Depending on the degree of these feelings, the climate in which the instruments are administered can help to overcome this problem. Instead of asking clients to "take a test," the evaluator should explain carefully why the information is necessary ("We want to find out how we are or are not helping you and what your needs are so that we can improve our program") and ask them to help by "giving some information." An instrument, just like a counseling session, is a means of obtaining information about the participant. A cooperation-seeking posture on the part of the staff often can overcome initial negative feelings about instruments.

When selecting specific outcome indicators, one must be fairly certain that the indicator is capable of reflecting change. In particular, it must be remembered that the sine qua non of instrument development, especially personality instruments, is reliability--stability of measurement over time (as well as internal consistency). Thus, items which are especially stable over time are more likely to be retained as the instrument is developed. Often the more carefully developed the instrument is, the more likely it is to contain items which are relatively impervious to change. This item stability is especially characteristic of personality tests, which are typically designed to tap basic, stable personality traits. A review of a sample of personality inventories, for example, reveals several items containing such phrases as "I have always been...", "As a child I felt...", or "In the past I have usually...". Although these are basically perceptions which can possibly change over time, one is loading the deck against the detection of change. The same problem exists with tests measuring values.

Therefore, when selecting outcome indicators, the evaluator must not only understand the phenomenon but scrutinize the measuring device, especially personal adjustment instruments, to be certain that it is capable of detecting change.

INSTRUMENTATION FOR CULTURAL AND ETHNIC MINORITIES

The last decade or so has seen an increasing awareness of the difficulties of conducting social research in minority communities (see, for example, Montero 1977). These difficulties are brought into sharp focus when the prevention evaluator is faced with the task of helping minority programs choose evaluation instruments.

On the simplest level, one must face the reality that instruments specifically designed for minority populations are extremely rare. This does not mean that instruments designed by or for the dominant culture are necessarily inappropriate for minority group members. However, attempts to use such instruments to evaluate minority prevention programs will compound enormously any problems in assessing reliability, validity, and relevance. When "middle class white English" is not the dominant language of the population under consideration, these problems are often insurmountable, and simply translating existing instruments is definitely not a solution.

In a more general sense, evaluation instruments for minority programs should reflect the social reality of the program staff and participants, a difficult proposition at best. Clearly, cultural and ethnic groups vary along such obvious dimensions as dominant values, taboos, definitions of deviance, and the meanings attributed to certain behaviors. Important differences also exist between various groups' understanding and ownership of science and the scientific method (see Trimble 1977) and in their perceptions of the value of cooperating with social research (Weiss 1977). An instrument which does not reflect the social reality of respondents will at best be difficult to complete, and at worst, will yield data which are grossly misleading.

It is unlikely that an evaluator from one cultural or ethnic group can accurately assess the appropriateness of a given instrument for respondents of another cultural and ethnic group without intensive and costly ethnographic study of the group. Even when the evaluator and the respondents are from the same cultural or ethnic background, there is no guarantee that a similar social reality will be shared. For example, Mayovich, a Japanese born in Japan, reports that she was perceived by activist Japanese Americans as "a 'white washed' racist, married to a white American, who was doing research based on white values" (1977, p. 115). For these reasons, we urge that community members be involved early on in the selection or development of instruments when minority programs are being evaluated, a policy advocated elsewhere in the Guidelines for evaluation of all programs.

A final issue surrounding the selection of instruments for cultural and ethnic minorities is the tendency among some evaluators to paint with too broad a brush when defining the "community" under study. For example, one evaluation purported to concern "Hispanic" adolescents when in reality, the study population was comprised of Mexicans, Mexican Americans, Central Americans, South Americans, Puerto Ricans, and Cubans. It is likely that the program participants differed on a number of dimensions, a fact concealed by the singular label, "Hispanic." As Weiss (1977) has noted,

> The idea that there is a "community," and a single community with which the researcher can communicate and which he/she can satisfy is a misunderstanding of reality in most situations. Different groups exist, compete, conflict; cleavages of ethnicity, religiosity, life goals, militancy, and interest make the search for "the community" an elusive quest (p. 34).

The prevention evaluator is urged to consider carefully the community with which s/he is dealing and to be sensitive to its diverse elements in aiding programs in selecting instruments.

It seems worth noting in closing one hypothesis as to why issues surrounding instrumentation become highly complex when ethnic and cultural boundaries are crossed. People are multidimensional, integrated, dynamic wholes (Gestalt), while instruments are most often still frame and unidimensional. Depending upon the dimension sampled, two individuals may seem very similar or greatly different. The same holds true for different ethnic and cultural groups.

This section began by asserting that issues of instrumentation bring into sharp focus the difficulty of conducting social research in minority communities. Perhaps the reverse is also true: considering multicultural issues brings into sharp focus the difficulty in choosing and using prevention evaluation instruments.

MULTIPLE INDICATORS

There are usually several ways in which a desired outcome can be manifested. For example, a participant's adjustment to school may be measured by attendance and punctuality, grades, disciplinary infractions, teacher ratings, a score on a school adjustment instrument, or even his/her response to the questions, "Do you mind going to school?" A program preferably should not rely on a single measure of criteria; it should gather as much of the relevant data as is feasible.

Several reasons exist for this multiple measures approach. First, most programs with a prevention orientation typically cannot expect massive changes in participant behavior or attitude. It is more likely that some particular aspect of a client's attitude or behavior will be altered by the program experience; other aspects will not change.

Similarly, change along certain criterion dimensions will vary according to the individual. One person may improve school performance through better grades, whereas another improves through increased participation in extracurricular activities.

A way to reduce the limitations imposed by single measures is to use multiple measures obtained by different methods, a strategy known as triangulation. If findings from several methods are congruent, then policy makers are entitled to greater faith in the validity of the results. If, however, the results are incongruent, then there is justification to explore the reasons for these differences. Triangulation is extremely helpful when using measures with questionable reliability or validity. (See Campbell and Fiske 1959 for a discussion of convergent and discriminant validity.)

Multiple measures also provide for the opportunity to detect latent changes in participants. Changes that have not been manifested in behavior may be detected at the attitudinal-value level. This advantage is especially important when trying to evaluate program effectiveness over a brief period of time. For example, when attempting to evaluate the impact of a values clarification session, one often cannot afford to wait several months or years to determine the full impact of the sessions. Therefore, determining whether the client has begun to reassess, if only mentally, his or her system of values is desirable. Instrument-based measures can be useful for such purposes.

REFERENCES

Abrams, L. A., Garfield, E. F., and Swisher, J. D. (Eds.). Accountability in drug education: A model for evaluation. Washington, D.C.: Drug Abuse Council, Inc., 1973.

Ad Hoc Committee on Ethical Standards in Psychological Research. Ethical principles in the conduct of research with human participants. Washington, D.C.: American Psychological Association, 1973.

Alcohol, Drug Abuse, and Mental Health Administration. The ADAMHA guide for the protection of the human subjects (Rev. ed.). Washington, D.C.: 1975

Anastasi, A. Psychological testing. (4th ed.). New York: MacMillan, 1976.

Campbell, D. T. and Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Fejer, D. and Smart, R. The knowledge about drugs, attitudes toward them, and drug use rates of high school students. Journal of Drug Education, 1973, 3 (4), 377-389.

Fox, J. A. and Tracy, P. E. The randomized response approach and its applicability to criminal justice research and evaluation. Evaluation Review, 1980, Forthcoming.

French, J. F. Randomized response: A method for increasing the privacy of individual responses to surveys. Current trends in drug abuse, 1. Rockville, Maryland: National Institute on Drug Abuse, June, 1979.

Gorsuch, R. E. and Butler, M. C. Initial drug abuse: A review of predisposing social psychological factors. Psychological Bulletin, 1976, 83, 120-137.

Gunderson, E. K., Russell, J. W., and Nail, R. L. A drug involvement scale for classification of drug abusers. Journal of Community Psychology, 1, 399-403.

Horan, J. J., Wescott, T. B., Vetovich, C., and Swisher, J. D. Drug usage: An experimental comparison of three assessment conditions. Psychological Reports, 1974, 35, 211-215.

Hurst, P., Cook, R. F. and Ramsay, D. A. Assessing the prevalence of illicit drug use in the army. Alexandria, Va.: Army Research Institute for the Behavioral and Social Sciences, 1975.

Lettieri, D. (Ed.). Predicting adolescent drug abuse: A review of issues, methods and correlates. Rockville, Maryland: National Institute on Drug Abuse, 1975.

Lu, K. H. The indexing and analysis of drug indulgence. International Journal of the Addictions, 1974, 9 (6), 785-804.

Mager, R. F. Preparing instructional objectives. Palo Alto, California: Fearon, 1962.

Marlatt, G. A. Behavioral assessment of social drinking and alcoholism. In G. A. Marlatt and P. E. Nathan (eds.), Behavioral approaches to alcoholism. New Brunswick, New Jersey: Center of Alcohol Studies, Rutgers University, 1978.

Mayovich, M. K. The difficulties of a minority researcher in minority communities. Journal of Social Issues, 33 (4), 1977, 108-119.

Montero, D. Research among racial and cultural minorities: An overview. Journal of Social Issues, 33 (4), 1977, 1-10.

Pandina, R. J., White, H. R., and Yorke, J. Estimation of substance use involvement: Theoretical considerations and empirical findings. International Journal of the Addictions, (Forthcoming, 1980)

Schute, R. The effect of peer influence on verbally expressed attitudes of male college students. State College, Pennsylvania: Addiction Prevention Laboratory, The Pennsylvania State University, 1976.

Segal, B. Personality factors related to drug and acohol use. In D. Lettieri (Ed.) Predicting adolescent drug abuse: A review of issues, methods and correlates. Rockville, Maryland: National Institute on Drug Abuse, 1975.

Suchman, E. A. Evaluating educational programs: A symposium. Urban Review, 1969, 3, 15-17.

Swisher, J. Program planning dimensions. In A. Abrams, E. Garfield, and J. Swisher (Eds.), Accountability in drug education: A model for evaluation. Washington, D.C.: Drug Abuse Council, 1973.

Swisher, J. and Crawford, J. Evaluation of a short term drug education program. The School Counselor, 1971, 18, 265-272.

Swisher, J., Warner, R., Spence, C., and Upcraft, L. Four approaches to drug abuse prevention among college students. Journal of College Student Personnel, 1973, 14, 231-235.

Trimble, J. E. The sojourner in the American Indian community: Methodological issues and concerns. Journal of Social Issues, 33 (4), 1977, 159-174.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. G. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Weiss, C. H. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Weiss, C. H. Survey reseachers and minority communities. Jcurnal of Social Issues, 33 (4), 1977, 20-35.

CHAPTER 5: IMPACT EVALUATION - INDICATORS AND MEASURES

INTRODUCTION

Social programs are designed to impact a problem or need that exists within the population. The type and magnitude of impact are used as the basis for making decisions about the effectiveness of prevention modalities and the need for program expansion. Thus the identification and estimation of impact is an important aspect of prevention evaluations.

The concept of impact as defined in chapter 2 distinguishes impact from and relates it to program outcomes. These relationships also are relevant for impact indicators, in that the types of effects that are appropriately considered as outcomes have impact counterparts. The primary difference between estimating and defining indicators of outcome and indicators of impact is the population for whom effects are measured. To reiterate a point made in the preceding chapter, outcome indicators measure changes in program participants, while impact indicators measure changes in the entire population for whom generalized effects are expected. Therefore, data collection for impact indicators includes both program participants and the general community. Impact indicators are used to define or describe program related and, hopefully, program induced changes in drug or alcohol abuse and related problems within a community.

Impact effects are generated from prevention program outcomes that are generalized above and beyond the specific effects on program participants. For example, a desired prevention program outcome is the reduction of drug or alcohol abuse by program participants. Assume that as a result of a program's operation over several years, there has been a favorable change in self-reported drug or alcohol consumption patterns in school program participants, drug traffic has been reduced, teacher-student relationships have improved, school vandalism has decreased, neighborhood disturbances by youth have become less violent, and student-parent relationships have improved. In other words, improvements in the school environment, in family relationships and neighborhood safety, and a decrease in school maintenance costs may be brought about by changes in consumption patterns. These are potential generalized impact effects, while the changed consumption patterns of program participants is an outcome effect.

Since these generalized effects occur throughout the community and across prevention programs within that community, they may be measured in aggregate or cumulative form. Impact effects are measured by indicators such as prevalence and incidence levels, rates of drug or alcohol related arrests, the incidence of drug or alcohol related hospitalizations, and so on.

Impact evaluation may have the same type of indicator as an outcome study, however, the data base of impact indicators is broader. Again, in our example, assume the high school has 2,000 students who are relatively homogeneous in terms of marijuana use. Of this population 1,000 have thus far participated in the prevention program. As a result of their participation, reported daily use of marijuana for those students has dropped from thirty-five percent to twenty-five percent. This effect is a measure of program outcome. Over the same time period, the self-reported daily use of marijuana for the entire school has likewise decreased from thirty-five percent to twenty-nine percent. In essence, the direct outcome of the program has been to change the daily marijuana use of 100 children. The generalized effect of the program, however, has potentially changed the marijuana use of an additional 20 students within the same school who did not directly participate in the prevention program.

Further, assume that the marijuana use of students in a comparable high school nearby decreased in a similar way, although the prevention program had not been introduced into the curriculum. These effects, in addition to the generalized effects at the original high school, are indications of prevention impact and would be included with other relevant indicators in an impact evaluation of prevention programing within that school system.

The distinction, then, between outcome and impact is at the program versus community level, affecting program participants only versus participants plus third parties, and in producing direct effects versus generalized effects. It should be noted, however, that (1) not all generalized effects are positive, and (2) these effects may overlap, making assessment difficult.

Impact indicators collectively represent those specific risk states and consequences that have been found or assumed to be associated with drug or alcohol abuse and related behaviors. From another perspective, impact indicators are the measurable aspects of quality of life in a community that may be related to drug and alcohol abuse. In response, prevention programs are instituted to adress the community problems by moderating, changing, and altering those risk states which contribute most to drug and alcohol abuse and are most amenable to program intervention.

The need for prevention programs, implicitly or explicitly, is based on an assessment of the risk status of the community. Both drug and alcohol specific and nonspecific characteristics may be used to define a risk state. Crime rates, educational levels, literacy rates, general health, and socioeconomic levels are but a few of the related indicators of risk. Some of these indicators may be affected by prevention programs, depending on the strategies, objectives, and activities of the programs. Changes in these indicators represent alterations in the risk status of the community and are the basis for testing whether the change was related to prevention programing.

In addition to the concept of a risk status of the community as a whole, there will be groups of individuals within a community who are relatively more or less at risk for drug and alcohol abuse. A change in the proportion of the population at risk is another possible indicator of program impact. The determination of risk status for subgroups of the population is important for targeting services as well as evaluating their impact. Outcomes may indicate that a program has been effective in changing knowledge, attitudes, and consumption patterns of participants, but if the program has been targeted toward low risk groups, or if only a small proportion of the high risk group has been reached by the program, the magnitude of impact will be diminished.

One indicator of drug problems in a community might be the number of individuals who have their first experience with illegal drugs during a given time period--incidence data. Assuming a prevention program has been operational in the interval, the change in the number of first users from one measurement period to the next provides an indicator of possible impact. The indicators can be made more meaningful by expressing the data as rates or ratios. For example, the number of the first users per 1,000 population in each period provides a better estimate of how common the problem is, and provides a basis for expressing the problem in terms of the percentage of the population that is affected. Variations in the age distributions of first users of different drugs at two periods in time may indicate the extent to which a prevention program has had an effect in delaying use. The numbers of first users can be related to the size of that component of the population during two periods in time, thereby controlling for changes in the population base and providing more precise indicators with which to estimate impact. Other adjustments that might be made to crude rates include controlling for socioeconomic level, race, sex, level of education, supply of illegal drugs, variations in community crime, and so forth. The more clearly the impact indicators are defined, the better are the inferences that can be drawn about the program and the problem.

Prevalence data provide another indicator of the drug and alcohol abuse problem. Prevalence rates indicate for example, the number of individuals who have had an identifiable alcohol problem in a specified time period. The time period may be a month, a year, or even a lifetime, and adjustments similar to those that are possible with crude incidence rates, can be made to crude prevalence rates. Incidence and prevalence rates or ratios, and at-risk levels actually are summary measures that can be used to define the status of a problem in a community and the impact a prevention program has had on the problem.

Since impact indicators are broad, covering individual participants as well as the community, it would be logical to assess impact throughout the entire social system. But philosophers and sociologists have had mixed success breaking down the social system into its many parts. So, for the purposes of the Guidelines, the social system will be considered to have three components: societal structure (individual, family, neighborhood, and community); institutions (school, hospital, police, courts, and other governmental institutions); and economy (physical or monetary resources). Impact indicators will be discussed within this framework.

The next section of this chapter takes a closer look at prevention program impact indicators. Many of these indicators, while appropriately considered to be possible indicators of prevention program impacts, are not limited to drug or alcohol prevention activities and objectives, but may be indicators of quality of health and life. Drug and alcohol abuse or their prevention are but one determinant of community and individual health.

CONSIDERATIONS FOR IMPACT EVALUATION

Program administrators as well as funding agencies often would like to enumerate all possible impacts of their program in their community. However, before a meaningful impact evaluation can be made, a number of factors must be considered carefully. These are a definition of community; the relationship between a program's size, input, and impact; intended and unintended impact; delay and durability of impact; and net effects, with a consideration of possible double counting.

ALTERNATIVE DEFINITIONS OF COMMUNITY

In addition to its importance in estimating program impact, the definition of community has implications for describing the pervasiveness of a program and estimating the probability of contact with a program by members of a target group. The pervasiveness of a program depends upon its size and scope and the definition of community. The probability of contact with a prevention program by members of a target group is directly related to the size of the group. Drug and alcohol prevention programs may be targeted at the elderly, parents, community groups, or the general public. However, much of the discussion on impact is phrased in terms of programs directed toward youths as this is the largest single target group for prevention programs.

Definition of community, then, varies with the program size and target group. A community can be a family, school, school district, neighborhood, police precinct, township, borough, city, or state. If a program is limited to a sixth grade class within a school, the community to be effected will probably not be the entire school district, nor will it be the entire city. The impact of this program will be limited to a few families and a neighborhood. Alternatively, in the case of a TV program, the impact will not be limited to only a particular school and neighborhood, but will include a much larger community, perhaps even the nation. The size of a program is a critical factor in estimating the relative impact on the community--if a program is "small," it is unrealistic to expect a sizable impact in a large community. Therefore, a proper definition of community is one that relates to the scope and intentions of a program. A reasonable way to define a community is the area in which detectable impacts may result.

INTENDED AND UNINTENDED EFFECTS

Programs may have both intended and unintended effects. Intended effects relate to program objectives, while unintended effects do not. For example, in the case of a public information campaign or program, a reduction of drug use may result. This is an intended effect. But perhaps in reducing one type of drug use, individuals may relieve their anxiety by increasing their consumption of another, such as tobacco or alcohol. This increase is an unintended effect. This example is not meant to imply that unintended effects are neces-

sarily negative or detrimental to overall program goals. Both intended and unintended effects are impacts of the program. Differentiating between them is useful for program decision makers in modifying programs.

DELAY AND DURABILITY OF IMPACT

Time is always required for a program to produce outcomes and impact on a community. Different types of programs take different lengths of time. An alternatives program such as a baseball team intended to divert students' interests in drugs will take a different amount of time to produce outcomes or impact than an information program that directly informs students of the consequences of drug use. If a program is evaluated too soon, or too late, the findings may show little or no impact on the community. But this would be due to faulty evaluation rather than to the program's efficacy.

In addition to the timing of evaluation, one must consider the duration of the impact of a program. The impact may not last very long, or be constant throughout a program's life cycle. Depending upon the nature of the prevention program, its impact may be manifested differently depending on when one looks at it. To be confident of the time dimension, then, one must either conduct a followup study or check references to similar studies.

IDENTIFICATION OF NET IMPACTS

Proving that a given impact is a result of a given program is not easy. It is especially difficult to isolate impacts of a specific prevention program if several programs have been implemented in the area. Ideally, the impact of a drug prevention program is measured by examining indicators in a community before and after the program, so that changes in the magnitude of the indicators can be compared. With before and after comparisons, however, one must always be concerned that factors other than the prevention program caused the change. By their very nature, it is difficult to measure and control for these external factors (see chapter 6 for a discussion of these issues).

For example, assume that a drug and alcohol education program is implemented in a given school. Assume further that toward the end of the program, a local law enforcement agency conducted a drug search in another school, making many arrests. The task for the evaluator is to find out the extent to which any community impact is due to the drug or alcohol prevention program in the absence of the arrests. In this case the evaluators could exclude the particular school having the law enforcement action and reserve comparisons for other schools in the district, or the possibility exists that the other school could be viewed as a separate comparison group.

In addition to these kinds of considerations, there are statistical techniques such as regression analysis and multivariate analysis that can be used to analyze the net impact of a program. These analyses require that there be variations in impacts and processes among the competing programs and that there is full information about changes of events that are relevant to the fluctuation of impact indicators in the community.

DOUBLE COUNTING AND THE LIMITS OF COUNTING

A prevention program eager to show its achievement or to justify its existence may report as much as possible about the various components and aspects of program impact within a community. For instance, drug prevention may reduce drug use, which in turn may reduce property crime. If in a report one includes the theoretical savings from the reduction in property crime, as well as the reduction of property crime incidence, this is double counting. The reduction in property crime and the cost savings thereof are alternative measures of the same impact indicator. The former is put in terms of a criminal index whereas the latter is phrased in monetary terms.

The above example can be extended further. One may argue that a sufficient reduction in drug use and property crime may enable the enforcement agency to divert its manpower from drug enforcement to other areas of crime prevention or law enforcement. The resources saved within the enforcement agency may be counted as additional impact due to the drug prevention programs. However, if one tries to include savings stemming from the other crime prevention programs (now improved via resources switched over from drug prevention), then the impact evaluation is extended beyond the reasonable limit. When estimating program impact one may include secondary effects but no so-called "third-stage effects."

IMPACT CANNOT BE ASSUMED

Many prevention programs are limited to given age, racial or ethnic groups, or a given geographic area. With these focuses, one may think it would be difficult to extend the program outcomes beyond their boundaries. This is indeed the case. Therefore, evaluators should be realistic in examining possible program impacts. They must understand that there are different kinds of barriers in our society, any of which may prevent the extension of a program outcome.

PRIMARY INDICATORS OF DRUG AND ALCOHOL PREVENTION PROGRAMS CHANGE IN USE AND RELATED ATTITUDES

Impact can be presented via three different kinds of parameters: by prevention modality, social structure (institutions or community components under which to seek indicators), or by specific indicators. To facilitate discussion, this section will consider the primary indicators of drug and alcohol prevention across the social structure. Since primary indicators apply to all types of prevention modalities, a separate identification is not required for each program.

Depending upon the focus of the impact evaluation and the goals of the prevention programing being assessed, a variety of specific impact indicators could be selected, to include: self-reported drug and alcohol consumption patterns; attitudes toward drug and alcohol use; community awareness levels; changes in drug and alcohol related arrests; reductions in drug and alcohol admissions in emergency rooms; changes in drug and alcohol related mortality/morbidity estimates or economic measures of effect, for example, drug or alcohol prices; reductions of treatment demand; and cost savings in law enforcement. In most impact evaluations, however, changes in consumption patterns and related attitudes will be essential variables of interest and will be the focus of this section of the Guidelines.

DECREASE IN DRUG USE AND ALCOHOL RELATED PROBLEMS

The most important impact indicator is the change of actual drug use. Three possible impact indicators that can be employed at the family, community, or school level are the total number of people abstaining from drugs, a reduction in drug doses, a change from hard drugs to milder types of drugs, and the quantity frequency, and variability of alcohol use and associated problems. Drug use information may be obtained from either survey data or institutional records. Surveys may be conducted at any level. For instance, the State of Pennsylvania in 1976 commissioned a statewide Prevalence and Intensity Survey about alcohol and illicit drug use. The survey included households and schools but excluded military barracks, state hospitals, and prisons.

Drug use rates and the quantity, frequency, and variability of alcohol use over populations generally are considered to be dependent variables. Other characteristics of the populations at hand (independent variables) may vary, and thus significantly influence drug or alcohol consumption. For example, knowledge of the detrimental effects of addiction, high levels of unemployment, high frequency of alienation, or widespread family breakdown all may be related either negatively or positively to rates of drug use and alcohol related problems. If a weighted linear combination of the independent variables accounts for a considerable amount of the variation in use rates, then there is some basis for making inferences

about population characteristics associated with high rates of drug use and alcohol related problems, thus helping to identify specific target populations for various prevention strategies.

Properties of populations other than rates can be described and correlated. For example, the teenage populations of 100 census tracts can be surveyed with scales measuring "alienation and neighborhood cohesion." If 50 randomly selected teenagers from each tract are interviewed, the mean or median scores on the two scales for each of the 100 census tracts can be computed to estimate variation in alienation and neighborhood cohesion among teenage populations. Relationships of these two variables with factors such as rates of drug and alcohol knowledge and drug and alcohol dependence may provide additional insight into the determinants of the rates of drug use and alcohol related problems.

Another indicator which may be used to describe community populations is the ratio or proportion of a population with a given attribute. Ratios are usually expressed as percents, such as percent of a census tract population living in overcrowded housing (according to some standard), percent women aged 45-54 who are widowed, percent of the population living alone, and so forth. Ratios, expressed as percents, are especially useful in describing geographically based populations in terms of the vast amount of data available in the various aspects of the census. Attributes such as occupation, income, education, family structure, age distribution, quality of housing units, living arrangements, migration, ethnic distributions, and others serve as important predictors of rates of drug and alcohol abuse and related behavior.

Finally, the rate of change in selected factors hypothetically linked to some form of drug use or alcohol related problems may provide an important indicator of program impact. Changes in socioeconomic status, ethnic distributions, and family structure may be linked to rates of addiction, mortality, mental illness, and other stress related sociomedical problems.

In summary, rates, ratios, measures of central tendency, and measures of change may be used to describe communities and program impact in communities. These indicators provide important characteristics of area populations and may serve as either independent or dependent variables in the estimation of impacts of a prevention program.

In addition to survey data at the community or school level, statistics recorded by law enforcement agencies and health institutions also are available for impact evaluation. A legal impact indicator that directly relates to drug and alcohol use is a change in the rate or number of drug and alcohol related arrests before and after a prevention program. These arrests include those for possession, sale, and manufacture of different kinds of drugs, and for public drunkness, driving while intoxicated, and sale to or possession by a minor. These may be expressed in numbers of offenses for a given period or rates in terms of arrests per 100,000 population in a certain area and period. The distribution of different kinds of arrests (for heroin sale versus marijuana possession) indicates the degree of seriousness of drug and alcohol problems in an area. The distribution of age groups of arrests may help to indicate whether a prevention target group has been influenced or not. Depending upon the scope of impact, these data can be identified at city, county, state, and national levels. In a larger city, police departments often can produce data for a given precinct.

Particularly important sources of information for some types of drug usage are hospitals, especially emergency rooms, and drug crisis centers. Impact indicators include various kinds of drug episodes (heroin, marijuana, LSD, PCP, and so on), and the frequency of various reasons for seeking help (for example, to quell disturbing psychic effects or overcome dependence, or because of severe depression, suicide attempts or gestures, and so on.)

Similar data are available for alcohol related problems. Emergency room visits resulting from accidental injury, cuts and bruises incurred while fighting, and complications of concurrent medical problems (for example, diabetes) can all provide indicators of impact.

There are also indicators of drug and alcohol related illness and mortality. Certain kinds of hepatitis and other illnesses are related to some drug use. Cirrhosis, certain kinds of gastro intestinal disorders, and cancer are all associated with heavy alcohol consumption. These illnesses will generate data regarding length of hospital stay or mortality. It should be noted, however, that drug related illness or mortality is not always apparent. Confusion results if diagnoses are listed in multiple categories (primary and secondary diagnoses) or if different physicians and nurses classify the same patient under different categories.

In addition to hospital based information, various local social service agencies such as drug crisis centers, alcoholism treatment, or information and referral centers may provide data on drug and alcohol abuse and their effects. These agencies serve as referral, treatment, and educational programs. They usually compile information about the number of treatments, kind of treatment, number of requests for information, and type of referral. Self-referral indicators are important when comparing before and after effects of a prevention program.

Health related data may be obtained from the Drug Abuse Warning Network (DAWN), Client Oriented Data Acquisition Process (CODAP), National Drug and Alcoholism Treatment Utilization Survey (NDATUS), Hospital Utilization Projects Records (HUP), PAS hospital records, or Alcohol Epidemiology Data System (AEDS).

CHANGE OF ATTITUDE TOWARD DRUG AND ALCOHOL USAGE

It is assumed that all six prevention modalities aim to change attitudes toward drug and alcohol abuse and inappropriate use. It is also possible that a change of attitude toward related behavior such as smoking will result. These changes begin at the individual participant level, and extend to the participant's peers, school, family, neighborhood, and community, depending upon the size of the program and barriers to the spread of effects. However, these indicators will not be applicable to other social and political institutions or to economic sectors. There are no archival data for attitude indicators. Therefore, the evaluator must rely on survey techniques, perhaps questionnaires distributed among families, schools, neighborhoods, and communities. These instruments are discussed in the appendix to part II.

The "dual" use of survey instruments is illustrated by the following--an evaluation of a mass media information program and a statewide audit of drug use and attitudes toward drugs.

The program was a TV drug abuse prevention campaign that was aired from December 1976 until May 1977 (Wotring, Heald, and Carpenter 1977). Five surveys of approximately 220 persons in the targeted audience (middle income heads of households) were conducted. The team responsible designed several criteria for use in evaluating the effect of the campaign on drug abuse attitudes. The team felt that the effectiveness of the campaign could be measured in terms of modifying the drug-related attitudes and perceptions of the audience. The following statements were included in the surveys:

- Illegal drugs are not the only problem; legal drugs are being abused.
- Drug abuse is everyone's problem.
- It has been said that people in this country regularly use drugs to solve all kinds of problems.

The research team found that among people who had seen the program, 56 percent agreed that "legal drugs are being abused" and 58 percent concurred that "people use drugs to solve all problems." Among nonviewers 44 percent and 42 percent, respectively, agreed with these statements. The research team concluded that the greatest impact of the TV campaign was its effect on the audience's perception of the drug abuse problem. However, due to differences in socio-demographic characteristics of viewers versus nonviewers the extent to which these results can be attributed directly to the campaign was not clear. A more careful research design and statistical analysis of the two groups might provide better information.

In New York, a statewide survey of drug use and attitudes toward drug use was carried out in 1975-76. Some 11,400 interviews were conducted in an attempt to estimate the use of illegal drugs and the nonmedical use of psychotherapeutic drugs among the population 14 years and older. The survey obtained information about the type of substance used, such as opiates, sedatives, stimulants, composite pills, marijuana and hashish, inhalants, psychedelics, cocaine, and alcohol, as well as information on the respondents' age, location, race, and sex.

Consistent with the theoretical research of Ajzen and Fishbein (1977), attitudes towards drug or alcohol use and subjective norms evoked by peers or influential others in a person's life appear to influence an individual's intention to use drugs or alcohol. These relationships can be assessed further during an impact evaluation by using multivariate analysis techniques. In addition, prior drug or alcohol related behaviors and attitudes appear to be important direct influences on use behavior independent of a person's intentions. Bentler and Speckart (1979), using causal modeling techniques developed by Jöreskog (1977) empirically established the relationship between attitudes, prior behaviors, and subjective norms to either influence an individual's intentions to use drugs and alcohol more importantly, to serve as causal events that are directly predictive of use behavior. Impact evaluation needs to reflect both the complexity of the relationship between attitudes and drug or alcohol related behavior and recent developments in the theoretical modeling of this relationship.

SECONDARY INDICATORS OF DRUG AND ALCOHOL PREVENTION PROGRAMS - CHANGE IN SOCIAL BEHAVIOR, SCHOOL, AND WORK PERFORMANCE

Many prevention programs, such as intervention, alternatives, and educational programs, endeavor to prevent or decrease drug or alcohol abuse on the individual level. These services may, at the same time, change the individual's interpersonal behavior (peer and family relationships) and personal behavior (school and work performance). As indicated earlier, due to the multiple objectives of these social programs, changes of behavior may result directly from the prevention program or indirectly from changes in consumption patterns.

A variety of indicators may be relevant depending upon the focus of the impact study (community, school, law enforcement, health institutions, or economic) and the prevention programing being assessed, to include: improvement in school and work performance; reduction of socially undesirable behavior; improvement of interpersonal relationships; a decrease in family court cases; reduction of undesirable conduct (quarrels, fights) and of criminal convictions or citations; less school vandalism; increases in school attendance; improvement in scholastic record; increases in school graduations; reductions in property damage; reductions in law enforcement cost; improvement in job skills; and increases in employment wage rates. Several of the more common indicators will be discussed in this section of the Guidelines.

REDUCTION OF SOCIALLY UNDESIREABLE BEHAVIOR

Impacts on socially undesirable behaviors can be found in all target areas--the individual, the family, schools, and the community. The instruments and data sources for these impact indicators are very similar to those discussed in the outcome indicators chapter. These and other scales are summarized in the appendix to part II.

In addition to the above indicators, it is possible to measure such things as a reduction of family court cases and in drug related criminal convictions such as property damage and theft. However, in order for these aspects of impact to be assessed, the magnitude of the change must be great enough to be revealed in data reporting systems. At the school level, reduction of violence and school vandalism are positive impact indicators. At the level of the community, these reductions may improve the general sense of security. Data for the indicators are available primarily from Uniform Crime Reports, family court records, juvenile probation files, and school discipline records.

There are certain problems which are inherent in drug and alcohol related arrest data. Police records vary considerably in terms of details, and have been known to show biases. Also, arrest records may be completed by officers unconcerned about details which may have longterm effects. Police can choose convenient charges rather than technically correct charges. Finally, police data sometimes omit demographic, social, economic, or work history information which would be important in an evaluation.

One needs a strict definition of a drug or alcohol violation to analyze arrest and conviction information. For instance, individuals arrested for felony charges are sometimes subsequently convicted for lessor offenses, such as disorderly persons charges, or are released without conviction through pretrial intervention. These processes serve to obscure the relationship between criminal justice records and actual drug or alcohol problems.

The extended effects of a reduction in criminal justice activity relating to drug and alcohol cases may reduce the costs of law enforcement for both the police and the court. These data have to be collected by survey or interview and must be based on current financial information. It is obvious that in order to be counted, the reduction must be sufficiently large to alter the allocation within law enforcement agencies.

In addition, dollar values can be imputed for reduction of school vandalism and criminal, drug-related property damage. These cost savings have an impact on the economy; that is, the resources saved through lack of damage may be used for other purposes. Civic, court, or police records may reveal the past valuation of property damage. One may use the information as a reference for imputing current dollar value, or one may have to assign a dollar value based on market information.

IMPROVEMENT IN SCHOOL AND WORK PERFORMANCES

Many prevention programs provide educational services or work experiences to program participants. These may not only impact the family or neighborhood, but may affect either the school or the labor market. At the school level, appropriate indicators are an increase in attendance, improvement in course work (grades), and an increase in the number of graduates. These data are available from school records.

At the labor market level, work experience provided by the program may improve a participant's job skills, which in turn may increase wage rates and employment opportunities (less job search time, less unemployment). Or, the reduction in drug or alcohol use may improve a participant's work productivity, decrease absenteeism, and lengthen employment duration. These indicators can be found at local employment agencies; the Bureau of Labor Statistics may furnish such data as wage rates, productivity index, earnings, and unemployment rates. These may also be obtained from State agencies or by survey methods, if resources permit.

Use of sedative and tranquilizer drugs available by prescription are also relevant indicators. It may be that a prevention program can influence the quantity and quality of use of certain prescription drugs. Drug prescription data might be available from the National Prescription Audit, Medicaid files in each State, and the Blue Cross-Blue Shield data file. The Medicaid file is limited to low income individuals while the Blue Cross-Blue Shield data pertain only to insured individuals (mostly middle income). The combination of these two sources provides fairly complete coverage.

Alcohol consumption information is available through drug and alcohol surveys conducted by many States, beverage sales, records, and consumer expenditure information obtained through AEDS. Physiological impact indicators center around the incidence and prevalence of alcohol-related illnesses such as cirrhosis, colitis, pancreatitis, and gastritis. Data on these illnesses can be obtained from hospital records (Hospital Utilization Projects) and Blue Cross-Blue Shield data files, and from the Alcohol Epidimiology Data System of NIAAA.

Drug and alcohol use indicators vary in importance in terms of prevention program objectives. Multiple indicators may provide a more comprehensive estimate of impact than a single indicator, but they also pose problems in terms of establishing their relative importance. Several approaches may be taken to resolve this situation. One is to rely on a

priori policy or program objectives as a basis for making an ordinal ranking of impact indicators. Another is to assign different weights to each impact indicator or construct a multidimensional scale of these weighted indicators. Such scales can be constructed by means of multivariate analysis (regression, canonical correlation, factor analysis, cluster analysis, and so on) or by assigning weights based on sample observations within a community (Pandina, White, and Yorke 1980).

ECONOMIC INDICATORS

The impact indicators measured at the community level are mainly psychosocial changes stemming from drug use. However, these changes can also produce economic impact among various institutions within the society. For instance, if usage drops while the supply of drugs stays the same, the prices of drugs may be lowered. This could also happen if there were a reduction in drug purity. Drug price and purity information are available at regional and local levels under the DEA-Heroin and Cocaine Retail Price/Purity Index reports.

A reduction in drug or alcohol use may also lower the number of people requiring treatment, which in turn may reduce the cost of treatment. The cost of treatment per case depends upon the nature and duration of treatment. A number of studies (Levenson 1973; A. D. Little 1974; Lemkau 1974; and Goldschmidt 1976) may be used as a reference for calculating the cost savings from reducing treatment.

Reduction of drug offenses, arrests, and convictions may release resources of law enforcement agencies from drug enforcement to other activities. Obviously, a small change of drug offense for a short period of time will not lead the law enforcement agency to reallocate resources away from drug enforcement--it takes a significant amount of reduction of drug offenses or arrests in order to induce a law enforcement agency to do that. Cost savings are possible in the areas of law enforcement, corrections, court systems, and drug traffic control. These savings can be estimated either from agency budget or expenditure records.

Cost savings are also possible at health institutions (emergency room and inpatient utilization) and in future prevention programs. These economic indicators are useful since they are measured in the simple terms of monetary value, and permit the costs and benefits of a prevention program to be weighed. On the other hand, these figures require numerous assumptions about the legitimacy of converting physiological changes into dollar value.

A REVIEW OF PRIMARY AND SECONDARY IMPACT INDICATORS BY SOCIAL STRUCTURE

Instead of presenting indicators across each component of social structure, an alternative way to examine impact indicators is to review them according to components effected--that is, the community at large, school, law enforcement, health institutions, or the economy. This review may repeat (double count) an indicator related to different social components, but it may help program evaluators to identify these indicators and to expedite data collection. Assuming prevention programs have positive effects, one may find the following impact indicators:

Community at Large (Family, Neighborhood, and Community)

- Change in attitude toward drug/alcohol usage
- Change in attitude toward related behavior (smoking)
- Increase in the total number of drug/alcohol abstainers

- Reduction in consumption
- Change in drug/alcohol use patterns
- Improvement of interpersonal relations
- Reduction in undesirable conduct within family or neighborhood (fights, quarrels, and theft)
- Improvement of the sense of community security
- Cost savings on drug/alcohol related property damage.

School

- Change in attitude toward drug/alcohol usage
- Change in attitude toward related behavior (smoking)
- Increase in total number of drug/alcohol abstainers
- Reduction in consumption of drug/alcohol
- Change in consumption drug/alcohol use patterns
- Change in interpersonal relations
- Reduction in undesirable conduct (fights, quarrels, and school violence)
- Reduction in school vandalism
- Increase in student school attendance
- Improvement in student scholastic records
- Reduction in student dropout rate/increase of probability of graduation.

Law Enforcement

- Reduction of drug/alcohol offenses and arrests (possession, manufacturing, transaction)
- Reduction of drug/alcohol convictions (type of conviction)
- Reduction of family court cases (juvenile case)
- Reduction of criminal convictions or citations (drug/alcohol related behavior)
- Cost savings of law enforcement (manpower as well as other nonhuman resources).

Health Institutions

- Reduction of drug/alcohol related emergency room incidents

- Fewer drug/alcohol crisis center referrals
- Reduction in drug or alcohol related mortality or morbidity
- Cost savings due to cutbacks in drug/alcohol treatment

Economy

- Improvement in job skill
- Increase in job employment
- Increase in wage rate
- Increase in earnings
- Cost savings on property damage
- Cost savings on law enforcement
- Reduction of drug prices
- Cost savings on drug/alcohol treatment.

It is obvious that not all prevention programs can generate all of these positive impacts, nor should they be measured in terms of specific areas of impact not relevant to their program activities. It takes a great deal of resources and a well designed program to achieve just some of these impacts. In most cases, depending on the nature of a prevention program, only a few impact indicators can be attributable to a given prevention program. Furthermore, the linkages between primary indicators and secondary indicators, between health effect indicators and economic impact indicators, require strong assumptions and careful research in order to yield meaningful evaluation.

REFERENCES

Ajzen I. and Fishbein, M. Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychological Bulletin, 1977, 84, 800-910.

A. D. Little Co. Social cost of drug abuse. Cambridge, Massachusetts: A. D. Little Co., 1976 (mimeo).

Bentler, P. and Speckart, G. Models of Attitude-Behavior Relations. Psychological Review, 1979, 86 (5), 452-464.

Goldschmidt, P. G. A Cost-Effectiveness model for evaluating health care programs: Application to drug abuse treatment. Inquiry, 1976, 13(1), 29-47.

Jöreskog, K. Structural equation models in the social sciences. In P. R. Krishnaish (Ed.), Applications of statistics. Amsterdam: North Holland, 1977.

Lemkau, P. V., Amsel, Z., Sanders, B., Amsel, J., and Seif, T.F. Social and economic costs of drug abuse. Baltimore, Maryland: The Johns Hopkins University, May 1974 (mimeo).

Levenson, I. Cost/Benefit analysis and program target populations: The narcotics addiction treatment case, The American Journal of Economics and Sociology, 1973, 32(2), 129-42.

Pandina, R. J., White, H. R., and Yorke, J. Estimation of substance use involvement: Theoretical considerations and empirical findings. The International Journal of the Addictions (Forthcoming, 1980).

Wotring, C. E., Heald, G. R., and Carpenter, C. T. Evaluation of the Florida drug abuse campaign (1976-77), Final Report. Tallahassee, Florida: Communication Research Center, Florida State University.

PART III

INTRODUCTION

In keeping with the model of evaluation set forth in chapter 2, part III deals with research designs and methodologies appropriate for process, outcome, and impact evaluations. Again, we emphasize that although these categories provide useful distinctions, there are overlaps and interrelationships between the three which prohibit clear separations in a review of designs.

As the evaluator collects information during the first stages of planning for the evaluation, constraints are developed which will later help to dictate the choice of design. The description of the evaluation process in chapter 2 delineates the key issues about which information must be collected. In essence, major emphasis must be placed on the form of information required, the criteria by which the expected value of the design will be specified, and the expected value of the information required by users. Further, measures and comparisons must be specified which will adequately adress the hypotheses available for testing. All of this must be done within constraints imposed by the program structure, that is, cost, data quality, time, and available services.

In addition to these considerations, the evaluator is further limited in choice of design by such technical issues as whether units of analysis (for example, subjects, classes) can be selected randomly, whether pretesting can be done, and so forth.

The field of evaluation research is expanding rapidly, as policy makers attempt to determine, support, or justify decisions based on the "best" available evidence. Scientific methodologies for determining the merit of various activities have become more sophisticated. The hypothetico-deductive paradigm represented by quantitative, experimental approaches dominates methodology, although the alternative holistic-inductive paradigm, derived mainly from anthropological field methods, is gaining recognition in evaluation research. Patton (1978) discusses the distinctions between these paradigms by considering a set of dichotomies--"qualitative versus quantitative methodology, subjectivity versus objectivity, closeness to versus distance from the data, holistic versus component analysis, fixed versus dynamic system perspectives, and uniformity versus diversity" (p. 236). He then goes on to point out that in fact these are not dichotomies but continua, which provide dimensions of choice for the evaluator.

Recently, the very foundations of arguments regarding the distinction between objectivity and subjectivity have changed. This was fostered in mathematics by Godel (1931), who proved that all closed formal logical systems contain at least one contradiction which can be resolved only by reference to another, or larger system. Brown (1972) presented a calculus of self reference, in which he first had to confront the fact that the world, in order to be viewed, must separate itself into at least one state which sees, and at least one other which is seen. Thus, in any attempt to see itself, it must make itself distict from, and thereby false to, itself. Varela (1976) claims that the capacity to make such distinctions becomes much more interesting than the content of the reality being distinguished. Thus, our intent, energy, or investment must focus not only on the product of our investigation, but also on the process of what we do to arrive at that point. Berger and Luckmann (1966) address the same issue from the perspective of the sociology of knowledge when they comment

that, "despite the objectivity that marks the social world in human experience, it does not thereby acquire an ontological status apart from the human activity that produces it (p. 60)."

Philosophers of science, as well as evaluators, continue to argue the distinctions between objectivity and subjectivity, often claiming that experimental, quantitative approaches are more objective than "soft" approaches such as ethnography. The argument becomes moot if one holds the view that objectivity is nothing more than consensual subjectivity.

This issue has been raised in these Guidelines to alert the evaluator to the potential for developing comprehensive examinations of program issues by melding the two paradigms in such a way that the evaluation plan is truly reflective of the purpose for the evaluation. It is recognized that many decision makers and research scientists give less credence to findings which are not supported by or derived from statistical analysis. A prime role of the evaluator, then, is to ensure that not only the strength of the findings, but also the methods by which they are obtained, are acceptable to decision makers. In some cases this will require an educational process during the developmental phases of the evaluation.

Chapters 6, 7, and 8 provide the reader with a discussion of the most common and most frequently used methods and designs for process, outcome, and impact evaluation, respectively. Chapter 9 gives an overview of data analysis issues and computer analysis techniques. In no way are these presentations intended to limit the evaluator. Rather, they form a framework from which the evaluator can draw in developing a comprehensive examination and analysis which is fitted to the needs of the program being evaluated. The reader's attention is drawn to the assignment of methodologies by process, outcome, and impact, and the fact that some of these methodologies are appropriate for more than the one section in which they are discussed.

REFERENCES

Berger, P., and Luckmann, T. L. The social construction of reality: A treatise in the sociology of knowledge. Garden City, New York: Doubleday, 1966.

Brown, G. S. Laws of form. New York: Bantam Books, 1973.

Gödel, K. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme I. Monatshefte für Mathematick und Physik, 1931, 38, 173-198.

Patton, M. Q. Utilization-focused evaluation. Beverly Hills, California: Sage Publications, 1978.

Varela, F. On observing natural systems. The CoEvaluation Quarterly, Summer, 1976, 26-31.

CHAPTER 6: PROCESS METHODOLOGY

The process of a prevention program refers to the manner in which services are delivered. To fully understand process requires prior analysis of inputs, especially the needs assessment that led to initiation of the program.

NEEDS ASSESSMENT

The information derived from needs assessments is useful in evaluating the extent to which program activities are responding to measured needs and to help identify differential program outcomes in relation to those needs. In addition, if the needs assessment is carried out with methodological rigor, it may serve as a baseline for future measures of impact. A properly conducted assessment of needs becomes, then, an important input to both the program and its evaluation.

There are several ways of analyzing needs. There are, first, survey methods. These include incidence and prevalence surveys, and ecological analyses that depend on both social indicators and direct archival records such as hospital admissions for alcohol and drug related emergencies, medical examiner reports, and the like.

Second, there are key informant surveys, in which individuals within the program's target population who are known to have knowledge about the community and its problems are interviewed concerning both the problems and needed services. With this method it is useful to solicit names (in a "snowball" fashion) of those whose views differ from the initial respondents, and then conduct similar interviews with them.

Third, are community forums in which local citizens are encouraged to give their impressions of the problem and services needed.

Last, there is programatic data analysis, which consists of inferring needs based on service and resource use as indicated by information obtained from both the program's own files and from those of similar programs. There is, however, an inherent weakness in this method which precludes its use as the sole source of needs determination. This is that data analysis of this sort assumes that the present program is meeting the demands of the community in some sort of proportional way, and that services need be adjusted upward or downward only according to changes in demographics. But the analysis is useless if the original input was misconstrued. For instance, while a program's records may indicate that a particular minority group is not using its services, it may be incorrect to infer that there is no need for prevention within that group. In fact, this may well indicate that the program is not adequately addressing the needs of a segment of the population.

A variety of techniques have been developed for needs assessment in order to improve the quality of information. An excellent reference in this regard is A Needs Assessment Workbook for Prevention Planning (NIDA 1979). Another is Needs Assessment Approaches: Concepts and Methods (Warheit, Bell, and Schwab 1977).

IDENTIFICATION OF OBJECTIVES

An adequate assessment of needs is a necessary prerequisite to the development of rational goals and objectives. The more complex an organization, the more goals and objectives it is likely to have. The evaluator who relies solely on original grant proposals or other such documents to identify objectives might end up in serious trouble. Too often such statements reflect the program administrator's perception of what the funding source wants to hear, rather than what is actually intended. Further, as programs grow and the environment changes, shifts in the direction of program activities and objectives can occur. Thus, a prime responsibility of the evaluator is to clarify and confirm stated objectives. The evaluator must take steps to ensure that what is being stated is a reflection of what the program truly expects to accomplish.

Several formal methods exist for clarifying and prioritizing objectives. Among these are multi-attribute utility analysis (Edwards, Guttentag, and Snapper 1975), based on a decision theoretic approach to evaluation, and the Nominal Group Technique (Delbecq and Van de Ven 1971). Such techniques are most appropriately used during the planning stages of the program, but very often the evaluator will find that an ongoing program has not sufficiently clarified objectives.

Even when objectives are clearly stated, reasonable, quantifiable, measurable, and prioritized, there still is not an answer to which objectives should be the focus of an evaluation study. This question must be addressed by the evaluator and program. As Patton (1978) points out, the evaluator can easily fall into the trap of defining objectives for the program or, just as bad, choosing which objectives to examine. Rather, these issues are the responsibility of the program working in close cooperation with, and assisted by, the evaluator.

ANALYSIS OF INPUTS

Very often there are strong interrelations between inputs and processes of a program. Even so, two programs may have the same inputs, and yet one is more effective than the other because of differences in the manner in which resources are combined. Indeed, a program may fail to produce favorable outcomes not because of a lack of resources but because of poor program process.

The resources that go into prevention programs may be categorized as human or physical. It is the interplay among the human and nonhuman inputs that forms the basis for evaluating process. There are two approaches to analyzing both types of inputs, one monetary and the other nonmonetary. The choice of which to follow depends upon the objective of the evaluation, the resources available, and the user of the evaluation. It is possible, and sometimes desirable, to combine elements of both the monetary and nonmonetary approaches.

NONMONETARY ASSESSMENT

The staff and participants--the so-called "human inputs"--may be analyzed in terms of numbers and qualitative characteristics. The number of employees and participants may appear easy to estimate, but procedures should be followed to ensure that full and parttime involvement is differentiated. In general, figures for staff should be converted into fulltime equivalents (FTE). Quantification of participant involvement is more complicated, depending usually on some arbitrary determination of what constitutes full or less than full participation.

The time period during which human inputs are quantified also is important. If data on inputs are collected at the beginning of the program, the actual program size (staff as well as participants) may be overestimated, whereas gathering data at the end of the program may lead to an underestimation of the size of the program. Therefore, a "stable period" must be used when measuring program size, normally a few months after the program has begun. An alternative is to assess the program at several different times during its life and then take the average of these estimates.

Important qualitative characteristics of staff and participants include the qualifications, experiences, motivations, attitudes, and behaviors they bring to the program. These factors are important in affecting not only program process but outcomes and impacts as well. They might be measured in terms of years of education, years of program related experience, academic degrees, numbers and frequencies of behavioral indicators, or more subjective assessments of motivations and attitudes.

Physical program inputs include physical plant space, equipment, facilities, supplies, transportation, and other similar resources. The kinds of physical inputs may vary among different types of prevention programs. For example, an information program may have extensive communications equipment whereas an alternatives program may use a number of transportation vehicles. In general, all inputs that are assumed to affect the outcome of the program should be assessed.

MONETARY EVALUATION - COSTS

Analysis of prevention program costs is useful and important because it can provide program managers and policymakers with data to account for use of public funds, to compare the costs of alternative prevention services, and to identify the efficiency of the operation.

Costs of prevention are defined as the value of inputs (resources) used for prevention activities. If the market for resources is competitive, and if the operators in these markets are motivated by economic forces, it can be shown that the observed market price per unit of resource is a true measure of the cost of the resource.

However, the private market may fail to accurately value inputs because markets may be nonexistent for inputs or outputs of prevention programs, because the production of these inputs or outputs in the private market may not be efficient, or because the drug prevention services may be directed toward a special group of people with different private market valuations of the services compared to other groups of people. When the private market fails to yield a price for inputs, "shadow prices" are used to estimate the societal value of these resources. A shadow price is an imputed market price. The procedures for imputing market prices are not straightforward; some rely on the willingness-to-pay approach, while others use the cost-savings method approach (the former refers to how much one is willing to pay to avoid alcoholism or drug addiction, the latter to the cost of treatment versus the savings resulting from the program).

Estimated costs may be classified as social, public, and private. Social costs refer to the value of resources incurred for a program by society as a whole. These include both public and private costs. Public costs include money expended by governmental agencies (Federal, State, and local) whereas the latter include the value of resources incurred by individual program participants (incidental costs to participants and earnings foregone while participating in the program) plus donations from private organizations.

It is convenient to collect cost data at a program level. Within a drug prevention program, costs can be divided into operating and capital costs. Operating expenses include those for personnel, transportation and communication, maintenance, and other items. Capital costs encompass building and equipment outlays.

SPECIAL CONSIDERATIONS FOR COST ESTIMATION

A careful study of drug prevention programs, especially regarding the availability of data, reveals that a number of issues must be discussed and resolved before costs can be estimated. These issues are related more to the calculation of program costs than private costs. They include (1) budgetary expenditures versus economic costs, (2) allocation of shared costs, (3) treatment of capital costs, and (4) imputation of opportunity costs. A brief discussion of these four considerations follows. (For a detailed discussion, see Hu, Swisher, and McDonnell 1978.)

Budgetary Expenditure vs. Economic Costs

It is customary to think of the terms "costs" and "expenditures" as interchangeable. From the economist's point of view, however, these are not the same. Costs are related to a specific output, whereas expenditures are often stated without relation to the output-time dimension. In drug prevention programs, some inputs are not consumed during the accounting period in which they were purchased (for example buildings, equipment, books). Assets of this type provide a stream of services over a number of accounting periods before they are exhausted. In such cases it is necessary to employ depreciation allowance estimates in order to convert expenditures to costs.

Finally, a third party may pay for a basic service or incur expenses on behalf of the drug prevention program. These expenditures should be treated as costs of the program. Full and accurate estimation of outlays (not merely expenditures) is essential in determining the cost of drug prevention programs.

Shared Costs

Shared costs occur within two contexts. First, at a given point in time, a specific input or facility may produce two or more distinct program outputs. Second, a facility or input may be used over time by successive cohorts of participants representing either the same or a different type of output.

In actual practice, shared costs frequently are averaged among different programs. A convenient approach is to allocate costs proportionally to the various outputs of a program. Such allocation is always arbitrary in nature, but the fact is shared average costs of prevention programs simply cannot be measured accurately in any economic sense.

Capital Costs

Capital costs may be listed for four different elements: site acquisition, capital improvements, physical plant construction and maintenance, and equipment.

Serious measurement problems stem from several physical and institutional factors. For example, the physical plant of the prevention institution usually has an economic life longer than the period of a particular program provided for any given cohort. Also, the services of this capital stock are not easily valued in market terms. Various approaches to capital depreciation have been used in the past, such as the straight-line depreciation based on historical costs or replacement costs, or the assessed valuation. Other approaches just ignore the costs since, in the shortrun, the capital has no alternative use. Different approaches will result in different cost estimates. A study by Hu, Lee, Stromsdorfer, and Kaufman (1969) has a more thorough discussion of these approaches.

For simplicity, one may use the straight-line depreciation method to impute the cost of capital (based on the purchase price and the number of useful years obtained from a piece of equipment or a property).

Opportunity Costs

Opportunity costs are the foregone value of resources devoted to a given program. For a prevention program, a church may donate space or an individual may donate time to provide services. These resources might be used at no cost to the program, but they could be used for other, revenue producing activities. Thus, from society's point of view, these values should be added to the program costs. The problem is to determine the basis for assigning monetary values to these foregone values. One could apply, say, the current market rent for a similar space or the wage rate for a similar work task in order to estimate opportunity costs. But this approach may overestimate these costs, since alternative uses of these particular resources may well be idle space and leisure hours.

Opportunity costs also apply to program participants. Some prevention programs require extensive participation. Students may have to give up parttime jobs. Loss of such income should be included as part of the private cost component. It is customary to obtain wage rates for similar types of students (grade, race, sex) and multiply by the number of hours of participation to estimate foregone earnings. Again, this approach will overestimate the participants' foregone earnings to the extent that they have no parttime jobs. The probability of employment of these cohort groups may be estimated in order to make appropriate adjustments in foregone earnings.

It should be noted that the above discussions of costs estimation and conceptual cost problems are applicable to all six prevention modalities. Among these some (education and alternatives) have higher personal costs while others (information) spend more on communication. These variations place emphasis on different elements of program costs, but the approach to estimating each of these costs should be the same.

Average Costs vs. Marginal Costs

Average costs are defined as the costs per unit of output of the program, whereas marginal (incremental or added) costs are those resulting from an additional unit of program output. These cost estimates are important in examining the efficiency and effectiveness of a program. When a policymaker is reallocating resources among various types of prevention programs, the critical criterion is frequently that of incremental costs rather than average costs. Therefore, it is useful to estimate marginal costs as well as average costs for each program.

The statistical technique of regression analysis may be used to examine the relationship between average costs and program size (number of participants). The sign of the coefficient of program size in relation to average cost can indicate whether the program is operating on a constant cost, decreasing cost, or increasing cost condition. As a result, the optimal size of a program (in terms of minimum average cost condition) may be obtained. Total cost and program size can be related to the magnitude of marginal cost of a program. Hu, Booms, and Kaltreider (1975) provide empirical formulation and estimation of these two cost concepts. An introductory presentation of regression analysis can be found in Kmenta (1971) and Hu (1973).

ANALYSIS OF PROCESS

Process evaluation is in part administrative monitoring. Every program should be monitored regarding the integrity and efficiency of the delivery of its services. But process evaluation is also partiy an evaluation of the ingredients between program inputs and program outputs. The process of a program may be described in terms of program participant/staff ratios, frequency and duration of program contacts, contents of program contact (type of services delivered), staff to participant interaction patterns, staff to staff interaction patterns, organizational functions, program management patterns, participant referral and identification process, and so on.

QUALITATIVE ASSESSMENT

Evaluation of program process can be conducted either quantitatively or qualitatively. Qualitative strategies are well known in the analysis of organizational structure and process. They also have a place in other aspects of evaluation, and will be discussed here as they relate to process evaluation. Further discussion of qualitative techniques is included in chapter 7.

The Ethnographic Method

Traditionally, anthropology has been used to study mostly nonwestern cultures and societies. Recently it has been directed at western cities and institutions such as schools, hospitals, and military units. More than any other approach discussed in this chapter,

ethnography is characterized by its ability to enter a field setting (school, government agency) without a prior conceptual structure and to develop a classification of cultural and social elements based on those factors which the researcher discovers to be significant for that setting (staff, clients, administrators, community). This approach is particularly recommended when one is asked to study a process that has not been well documented in the past, or where there is reason to believe that existing documentation misses some critical dimensions of the process. It is therefore one of the most important exploratory methods available to the process evaluator. The principal tools of ethnography are the interview and observation, although the ethnographer is also likely to examine archival data when it is available.

Garfinkel (1967) has named an approach to the study of human interactions, "ethnomethodology." This approach is derived from descriptive anthropology and the sociology of knowledge. It refers to "the methods or practical understandings people acquire and use in the course of action and interaction" (Broom and Selznick 1973, p. 27). Several principles of behavior and thought underlie this work: (1) we make sense of the world by calling upon a set of tacit assumptions, (2) when we act and respond, we sort out and classify our experiences, (3) we negotiate meanings to fit particular circumstances, and (4) we negotiate social order (Broom and Selznick 1973). This school of thought provides a framework for the examination of interaction in a microsystem, but has the disadvantages of not sufficiently taking into account the effect on that system of impersonal, external forces.

Interaction Analysis

This approach arose out of a fundamental concept in social psychology--the small group as a primary social phenomenon. Probably the most influential of early research using this approach was the study of the Hawthorne wiring plant by Elton Mayo and his colleagues. A number of those who participated in this protracted analysis, among them Robert Bales, Elliot Chapple, George Homans, and F. L. W. Richardson, subsequently developed variations on the original approach.

The Chapple (1949) approach to the study of interaction has been employed to test the suitability of workers for various settings that require different patterns of interaction. The Bales (1950) technique for studying interaction allows one to categorize and count the interactions that spring from group sessions and to observe the development of groups over time. Homans' (1950) work attempts to develop a broader conceptual framework of the human group. However, it is Richardson's work (1978) that is probably the most relevant for process evaluation of prevention programs. He classified patterns of interaction in a variety of human service delivery, industrial, and research organizations. He developed an approach for creating a profile of an organization and then assessing this pattern in terms of deficiencies that often lead to poor morale, weak leadership, confusion as to objectives, and other types of organizational malfunction.

An approach which combines some of the elements of interaction analysis with ethnography is discussed by Stake (1976). In transaction-observation approaches, writes the author, "the activities of the programme are studied...with special attention to settings or milieu. Issues are often drawn from the proceedings rather than from theory or from goal statements."

Sociometric Analysis

This is a family of techniques for providing sensitive and objective views of the web of interactions that characterize daily social settings. Sociometric approaches are of particular interest regarding the study of drug abuse prevention programs because, first, they are uniquely able to display the character of the social interactions that comprise these programs, and second, they are particularly useful in the study of educational settings, where the largest number of drug abuse prevention programs are currently being carried out.

A sociometric measure is a means of assessing interactions of interest within a group. It can be used to examine such phenomenon as attractions, repulsions, leadership, communications, power, influence, and so forth. Development of the measure often involves each member of the group privately producing a list of other persons in the group, sometimes by rank or grading in terms of the strength of the interaction between the two vis-a-vis the characteristic of interest.

Sociometry can be applied to various subgroupings of both staff and participants in a program. It can be applied at various points in time, and can be used to identify problems of morale, communication, and group cohesiveness. There is an extensive body of analytic procedures, both quantitative and qualitative, for interpreting sociometric data. As with any method, the analysis depends in part on the care with which one plans its use in the evaluation.

Typically, the researcher specifies the attributes to be studied in a sociometric analysis. Another approach appropriate for small groups is to ask members to rate each possible pair of group members based on their similarity to each other, and then submit these similarity ratings to multidimensional scaling analysis. Interpretation of the extracted dimensions provides insight into those qualities attended to by members of an organization in their interactions with each other. (See French 1977).

ORGANIZATION ASSESSMENT

This kind of assessment deals with the organizational structure of a program in delivering its services. It focuses on who determines the allocation of resources, who decides the program content, who conducts participant recruiting, and who evaluates the feedback to program administrators. The lines of responsibility and command as well as decision making processes are all important elements in the success or failure of a program. It follows then that an organization flow chart showing resources available at each level, together with authority and responsibility, is indispensible in evaluating process. Defining formal authority and responsibility this way is particularly important, since it brings into light the many nuances of informal authority and informal structure. The full picture thus obtainable will more clearly identify existing gaps, conflicts, and shortcomings of the operation of the program, possibly leading to improvement in the future. Likewise, a positive evaluation of a program's organization may provide a successful example of prevention services for implementation in other settings.

Dornbusch and Scott (1975) provide a framework for the analysis of an organization based on the notion that power can be translated into control. This is accomplished by applying sanctions based on evaluation of performance, and it is effective to the extent that participants see the evaluation as rational and appropriate. This complex, empirically derived theory focuses on the supervisor as the evaluator of the individual worker, but it holds important implications for any evaluation effort.

Patton (1978), after reviewing the literature on the treatment environment, lists some of the paired opposite terms he found helpful in describing different programs. These terms are relevant to any organizational environment, including those of prevention programs. The terms are:

Formal	Informal
Centralized	Decentralized
Authoritarian	Participatory (Democratic)
Divisive	Cohesive
Standardized	Individualized
Hierarchical	Egalitarian
Controlled	Expressive

CONTINUED

1 OF 3

Partitioned	Integrated
Independent Parts	Interdependent Parts
Routinized	Individualized
Isolated	Community-Oriented
Low-Communication Interactions	High-Communication Interactions

Patton's list is so comprehensive and concise that it almost demands development as a test instrument for organizational analysis and process evaluation research.

QUANTITATIVE ASSESSMENT

Certain elements of prevention programs can be evaluated quantitatively. For example, the participant/staff ratio and frequency and duration of program contacts can be measured in a straightforward manner. (However, one still should follow the previous recommendation regarding fulltime equivalent participants and fulltime equivalent staff.)

Quantitative assessments, when routinized for the purpose of transmitting information to managers in order to facilitate decision making, are frequently referred to as Management Information Systems (MIS). An organization truly cannot function purposefully without some form of MIS. Very often, the term MIS is assumed to refer only to some formal, complex, often computerized process that is always conducted on a conscious level. The fact is in all organizations information is exchanged in a variety of ways, informal as well as formal.

What we refer to as an MIS is merely a formalization of the exchange process carried out to increase validity and reliability. The key components of the system are data capturing devices, processing and feedback channels, and procedures for analyzing the data by comparing them to plans and standards. Implementation requires the application of decisions based on this information, the result being greater managerial control of the organization (Bocchino 1972).

The major questions to be asked in the development of an MIS are essentially those of the journalist--who, what, why, when, and where. Properly adressed, these questions help shape development of a formal system to provide management with timely and accurate information. And this of course allows management to optimize the use of resources to most effectively achieve the organization's objectives. The rule of parsimony applies both to cost and to the amount of data fed back to managers: expenses should be minimal and only that information needed in the decision making process should be incorporated into an MIS.

Cooper (1973), writing about community mental health centers, describes five general areas of data essential to a minimum information system. His analysis is also applicable to drug abuse prevention programs. Cooper's five areas, about which information should be provided, are as follows:

- Target population
- Recipients of service
- Services provided
- Staff activities
- Finance.

Cooper stresses the importance of including all five elements since adequate analysis requires examination not only of each separate element but of their interrelationships. Cooper provides extensive lists of data to be considered within each area.

An efficient MIS can provide decision makers with much of the information discussed under process evaluation in the Guidelines. For example, the MIS will periodically report on the quantity of services provided and compare results with program objectives. The evaluation consultant, then, may play a major role in assisting an agency to develop a workable MIS.

A self-assessment system focusing on both outcome and process is found in Volume II of the Manual of an Evaluation System for Treatment Programs for Drug Abusing Youth (CONSAD 1976). The proposed system is intended to provide:

- A basic structure within which a formal self-assessment process may be conducted
- Standardized, comprehensive information supplied on a continuing basis
- Types of information needed to reliably document program activities and outcomes
- A basis for early identification of problem areas and a method for using information to improve service delivery
- A method which directly involves program staff and does not depend on outside sources for implementation
- Advice to help minimize the effort in instituting a good self-assessment.

Another highly useful discussion of self-assessment for the evaluation consultant is Program Development: A Manual for Organizational Self-Study (Blanton and Alley 1975). This thorough exposition is adressed to those involved in program development--the manager, planner, staff, user, and so forth.

Process evaluation is important for several reasons. It is a necessary tool in the formation of the organization. It allows objectives to be related to specific program components, and it identifies the key elements of an effective organization for its replication elsewhere.

REFERENCES

Bales, R. F. Interaction process analysis: A method for the study of small groups. Cambridge, Massachusetts: Addison-Wesley, 1950.

Bernstein, I. N. Validity issues in evaluative research: An overview. In I. N. Bernstein, (Ed.), Validity issues in evaluative research. Beverly Hills, California: Sage Publications, 1976.

Blanton, J., and Alley, S. Program development: A manual for organizational self-study. (NIMH, Contract No. HSM-42-72-143). Washington, D.C.: National Institute on Mental Health, 1975.

Bocchino, W. A. Management information systems: Tools and techniques. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

Broom, L., and Selznick, P. Sociology: A text with adapted readings. (5th ed.). New York: Harper and Row, 1973.

Chapple, E. D. The interaction chronograph: Its evolution and present application. Personnel, 1949, 25, 295-307.

CONSAD Research Corp. Manual of an evaluation system for treatment programs for drug abusing youth, Vol. II. (NIDA Contract No. 271-75-4066). Washington, D.C.: National Institute on Drug Abuse, 1976.

Cooper, E. M. Guidelines for a minimum statistical and accounting system for community mental health centers. DHEW Pub. No. (ADM) 74-14. Rockville, Maryland: National Institute on Mental Health, 1973.

Delbecq, A. L. and Van de Ven, A. H. A group process model for problem identification and program planning. Journal of Applied Behavioral Science, 1971, 7 (4) 466-492.

Dornbusch, S. M. and Scott, W. R. Evaluation and the exercise of authority. San Francisco: Jossey-Bass, 1975.

Edwards, W., Guttentag, M., and Snapper, K. A decision theoretic appoach to evaluation research. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. 1 Beverly Hills, California: Sage Publications, 1975.

Foster, G. Medical Anthropology: Some constrasts with medical sociology. Medical Anthropology Newsletter, 1974, 6, 1-6.

French, J. F. Analysis of a small work group by multidimensional scaling of perceived similarities. (working paper). Trenton, New Jersey: Division of Narcotic and Drug Abuse Control, 1977.

Garfinkel, H. Studies in ethnomethodology. Englewood Cliffs, New Jersey: Prentice Hall, 1967.

Homans, G. C. The human group. New York: Harcourt and Brace, 1950.

Hu, T., Booms, B., and Kaltreider, L. A benefit cost analysis of alternative library delivery systems. Westport, Connecticut: Greenwood Press, 1975.

Hu, T., Lee, M. L., Stromsdorfer, E., and Kaufman, J. A cost-effectiveness study of vocational education. University Park, Pennsylvania: Institute for Research on Human Resources, The Pennsylvania State University, 1968.

Hu, T., Swisher, J., and McDonnell, N. Cost-effectiveness of drug prevention programs. University Park, Pennsylvania: Center for Research on Human Resources, Institute for Policy Research and Evaluation, The Pennsylvania State University, 1978.

Hu, T. Econometrics: An introductory analysis. Baltimore, Maryland: University Park Press, 1972.

Kmenta, J. Elements of econometrics. New York: MacMillan, 1971.

Landy, D. (Ed.). Culture, disease and healing: Studies in medical anthropology. New York: MacMillan, 1977.

Logan, M. H., and Hunt, E. (Eds.). Health and the human condition: Perspectives on medical anthropology. Cambridge, Massachusetts: Duxbury Press, 1978.

Moreno, J. L. Who shall survive? Washington, D.C.: Nervous and Mental Disease Monograph, No. 58, 1934.

National Institute on Drug Abuse, A needs assessment workbook for prevention planning. Washington, D.C.: Prevention Branch, 1979.

Patton, M. Q. Utilization-focused evaluation. Beverly Hills, California: Sage Publications, 1978.

Richardson, F. L. The elusive nature of cooperation and leadership: Discovering a primitive process that regultes human behavior. In Eddy and Partridge, (Eds.), Applied anthropology in America. New York: Columbia University Press, 1978.

Stake, R. E. Evaluating educational programmers. Paris: Organization for Economic Cooperation and Development, 1976.

Twain, D. Developing and implementing a research strategy. In E. L. Struening, and M. Guttentag, (Eds.), Handbook of evaluation research, Vol. 1. Beverly Hills, California: Sage Publications, 1975.

Warheit, G. J., Bell, R. A., and Schwab, J. J. Needs assessment approaches: Concepts and methods. National Institute on Drug Abuse, Washington, D.C. 1977.

Weiss, R. S., and Rein, M. The evaluation of broad-aim programs: A cautionary case and a moral. Annals of the American Academy of Political and Social Science, 1969, 389.

CHAPTER 7: OUTCOME STUDIES IN EVALUATION RESEARCH

INTRODUCTION

In simple terms, evaluation research seeks the truth. It attempts to answer the question: Is a program effective according to stated goals and according to reliable and valid outcome criteria? The responsibility of the evaluator is to provide an accurate answer to this question, realizing that the answer may affect the fate of the program. Should the program be continued, initiated in other settings, changed in ways suggested by the results of the evaluation, or discontinued altogether? If evaluators are to provide the information base for such crucial decisions, a deep concern about the accuracy of their results and conclusions is appropriate.

The purpose of this chapter is to aid the evaluator in the difficult process of selecting an evaluation research design which yields information of sufficient accuracy to make valid decisions. Attempts to accomplish this purpose include the formulation of a series of questions intended to alert the reader to the variety of issues, problems, and challenges which the evaluator must consider in designing an evaluation study. Issues of internal and external validity in evaluation research are identified and discussed to a limited extent, and references to a more extensive discussion of these are provided. Descriptions of selected true and quasi-experiments follow. A sampler of advantages and disadvantages are presented for each design, along with references for more extensive reading and study.

SELECTING AN EVALUATION RESEARCH DESIGN

The research design for an evaluation study is conceived as the logical structure or plan which dictates the conduct of the study and specifies such conditions as how and when samples are drawn, when measures are taken, and how data are analyzed. The selection of the design is a function of two fundamental conditions: (1) the design must yield results of sufficient accuracy to meet the purposes of the study, and (2) the requirements of the design, particularly its demands on the program context, must be approved by key individuals or groups such as top administrators, the subjects participating in the study, the governing board and, in some cases, the community at large.

The goal of the evaluator is to implement the evaluation research design that will yield the most accurate and useful set of results as a basis for inferring the magnitude of effects and making decisions about the level of program effectiveness. In order to accomplish this goal, it is advisable for the evaluator to develop a hierarchy of design preferences and a strategy for informing and convincing key administrators of their possible implementation. At this point the evaluator should be able to describe the design and its demands on the program context in lucid, nontechnical language. The evaluator should be prepared to discuss possible disruptions that would be caused by the study and how the results of the evaluation would benefit the program in terms of understanding its influence on subjects and providing a basis for program improvement.

In the process of discussing possible research designs, and eventually reaching a solution within the constraints imposed by the program context, the evaluator must be keenly aware of whether or not the design appropriate for the program context will answer the key

questions asked of the evaluation study. The following list of questions is provided for the investigator as guidelines for evaluating the various aspects of a planned study:

ISSUES REGARDING THE EVALUATION RESEARCH CONTEXT

1. Do you as the evaluator understand both the formal and informal aspects of the organizational structure within which the evaluation study will be conducted? (Twain 1975; Gurel 1975; Weiss 1975.)

2. Are you aware of and well informed about the social and political forces which influence the above organizational structure and, possibly, the welfare and conduct of your study?

3. Have you identified the key actors in the organization, their position in the power structure, their vested interests, and their perception and evaluation of the demands and consequences of the evaluation research to be done?

4. Are you aware of the potential role and value conflicts between you, the applied scientist, who is emphasizing objectivity and demonstrated program effects, and the program director and administrator who are already convinced of the program's efficiency, effectiveness, and value? What can you do about this state of affairs? (Gorry and Goodrich 1978; Abt 1978.)

5. Has the turnover of key staff members been sufficiently low to provide an organizational context stable enough to facilitate the conceptualization, implementation, and completion of evaluation studies? (Levine and Levine 1977.)

6. Is there a record keeping system sufficiently comprehensive and accurate to provide a basis for various sample selection procedures? Does the record system yield attrition rates with sufficient accuracy to estimate sampling mortality in the proposed evaluation study? (Weinstein 1975.)

7. To what extent will the research context approve and facilitate a set of procedures for the selection and assignment of subjects which meet the demands of the research design, selected to provide information of sufficient accuracy for evaluating the program under consideration? (Boruch, McSweeney, and Soderstrom 1978.)

ISSUES OF THEORY APPLICATION

1. Should the investigator consider the application of selected theories to facilitate a more comprehensive understanding of the evaluation problem? (Hawkins 1978.)

2. How might the consideration of relevant theories help to conceptualize the evaluation problem in greater depth? How might it help to formulate specific directional hypotheses as a basis for the selection or development of operational definitions of key constructs?

3. Is there theory of sufficient relevance and utility to predict differential reaction of subjects to program components? What implications will predictions of differential reaction have for selecting the design of the evaluation study, including the determination of sample size and subject allocation?

4. How might the application of relevant theories help to understand the transactional relationship between the stimuli of program components and the different reactions and changes in clients of the program. How might such considerations enhance the appropriate selection of outcome or change measures?

ISSUES OF MEASUREMENT

1. How many outcome, or dependent, variables are necessary to measure the important components of hypothesized program effect or influence?

2. Are characteristics of properties of the program (the independent variables) reasonably constant over time? Can such characteristics be quantified or categorized to ensure or document the degree of program consistency?

3. To what extent (time) and degree (intensity) are subjects exposed to program content? Can estimates of these parameters be quantified?

4. Will the measures selected to estimate program effects have sufficient range to allow all subjects to change (to improve or get worse) once they enter the program?

5. Are outcome variables sufficiently reliable to measure change in subjects as they experience program content over time?

6. Will outcome variables interact with program content so that observed changes in subjects may be a function of program content by outcome variable interactions rather than the influence of program content alone?

7. Is the content of the selected measures appropriate for the population being studied? That is, will the items or questions comprising measures interact with the ethnic, cultural, social class, and education level of subjects to yield different levels of reliability and validity as a function of group membership?

ISSUES OF SAMPLING

1. Is the population sampled homogeneous and well defined or is it highly varied and not comprehensively described?

2. Is there a basis for estimating sample attrition, possibly from other studies previously completed? Will attrition rates be a function of membership in the treatment or control group and thus threaten both the internal and external validity of the study?

3. Are samples highly selected in some manner which may affect the study outcome? For example, were subjects high or low on performance criteria or were they self selected in some unknown manner?

ISSUES OF STATISTICAL VALIDITY

1. What criteria should be used to determine the level of statistical significance?

2. Will the data generated by the study meet the assumptions of standard data analytic procedures? Under what conditions can these assumptions be violated without invalidating the statistical test being used?

3. How should the sample size be determined? How might the determination of sample size be influenced by such factors as the hypothesis of differential reactions to program components, the assumption of interaction effects, and other considerations? (Cohen 1977.)

ISSUES OF GENERALIZABILITY AND REPLICATION

1. To what extent will you be able to generalize the results and conclusions of your evaluation study to similar programs and similar subject populations located in different treatment settings and different ecological contexts?

2. Can the populations from which subjects were allocated to treatment and control groups be clearly identified and meaningfully described? Can the population studied be identified in another treatment setting to facilitate replication of your study?

3. Can the program content and the evaluation methodology be described with sufficient precision to allow replication of your study in another treatment context?

THE VALIDITY OF EXPERIMENTS

Among the major goals of this chapter is the identification and development of standards for guiding the selection of evaluation research designs. The investigator must select, successfully implement, and bring to fruition a design which yields accurate enough information to make valid inferences and general conclusions. However the road from conceptualization through implementation and completion is indeed rocky, with many potential ambushes along the way. This section will be concerned with threats to the internal and external validity of true and quasi-experimental designs. Definitions will be presented along with references to articles and books which discuss these issues in greater depth and detail.

INTERNAL VALIDITY

The following are definitions of internal validity as found in the literature:

- Internal validity is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? (Campbell and Stanley 1966, p. 5).
- Internal validity is concerned with the question of whether the treatment as manipulated caused any change in the effect as measured (Cook, Cook, and Mark 1977, p. 111).
- The internal validity of a hypothesis is the number of its alternative hypotheses disproved (Wiggins 1968, p. 390).
- Internal validity refers to the degree to which a design allows one to rule out alternative explanations for the way in which a particular independent variable is causally related to the dependent variable of interest (Bernstein, Bohrnstedt, and Borgatta 1976, p. 108).
- Internal validity refers to the validity of any conclusions we draw about whether a demonstrated statistical relationship implies cause (Cook and Campbell 1976, p. 223).

Thus a research design is internally valid if, from a demonstrated statistical relationship between an independent and a dependent variable, we can correctly infer that change in the dependent or outcome variable was caused by a corresponding change in the independent or treatment variable and not by some unknown extraneous variables.

The probability of this conclusion being correct increases as evidence for the influences of extraneous variables on the dependent or outcome variable is ruled out. More precisely, as the number of extraneous variables ruled out increases the probability of a causal relationship between the treatment and outcome variables increases.

The following extraneous variables, if not controlled in the experimental design, might influence the dependent variable in some unknown manner:

History. This term refers to specific events or conditions, in addition to the treatment or experimental influence, which occur between first and subsequent measurements of the dependent or outcome variable, and which might influence the magnitude of these measurements.

For example, the outcome of an educational program designed to influence attitudes toward marijuana smoking could be profoundly affected by the publication of a study indicating that certain patterns of marijuana smoking were or were not hazardous to the health of the smoker. The release of such information following the pretesting of subjects would be considered an historical event which might influence the outcome of the study. The occurrence of such events are threats to internal validity unless the investigator anticipates and plans for their appropriate control within the logical structure of the research design.

Maturation. This refers to changes in subjects between pretest and subsequent testing which influence the observed outcome, but are not a part of the program or treatment of research interest. Subjects may grow older, wiser, stronger, acquire new knowledge, or enter into a biological change. While these factors may not be a part of the program or treatment of interest to the investigator, they may very well influence the outcome and must therefore be considered in developing the research design. For example, growing older and acquiring biological maturity may in themselves influence high school students beyond the effects of the program being evaluated. The presence of these potential influences should be considered in conceptualizing the evaluation study and in selecting the appropriate design.

Testing. This has to do with how the experience of taking a test affects the subject's subsequent performance on the same test. In pretest/posttest designs the concern is with the influence of the pretest experience on the posttest score. This potential influence is even more likely in time series or repeated measurements designs where multiple testing is required.

In completing an evaluation test or scale, it is reasonable to assume that learning may take place, attitudes may change, or emotions may be aroused in a manner that will influence a subsequent performance when the same set of stimuli are again encountered. The investigator must be aware of this potential influence and the necessity to estimate or control its effect under particular research conditions.

Instrumentation. This term refers to autonomous changes in the measuring instruments or procedures used to estimate parameters over repeated observations. Instrumentation is differentiated from testing in that in the former an effect is due to a change in the instrument or procedure of measurement itself while in testing the effect is due to change in the subjects as a function of their testing experiences.

If, for example, measures are derived from classroom observations or ratings, the observers may change in some unknown way as the study progresses and repeated observations are made. Observers may become more skilled, bored, or inferential as they rate samples of group behavior.

If measures are based on interview data, the quality of these observations may change as the interviewer gains experience and confidence, becomes more familiar with the interview schedule, or gains insight into the subject over repeated interviews.

In summary, effects due to instrumentation change must be anticipated in repeated observations of subjects and should play a role in research design selection.

Statistical Regression. This is an effect that can stem from subjects being classified into or selected from extreme groups on the basis of pretest scores, correlates of pretest scores, or some other basis. When subjects are assigned to groups on the basis of high or low pretest scores, the high groups will tend to score lower and the lower groups higher at the posttest. This will be especially true if the pre and posttests are unreliable, or include a considerable amount of measurement error. Under these conditions changes from pre to post are very likely due to statistical regression to the mean. Attributing such changes to treatment influences would almost certainly be incorrect. Investigators must be aware of these possibilities as they create experimental and control groups through various selection and classification procedures.

Selection. As used here, selection refers to an experimental effect that might be due to the lack of equivalence between the treatment and control groups at the inception of the study. Under these conditions differences obtained at the posttest might be due to unknown differences between the treatment and control groups rather than due to treatment influ-

ences. Therefore it is crucial to establish, and if possible empirically demonstrate the equivalence of treatment and control groups at the beginning of the evaluation study. Otherwise erroneous inferences regarding the treatment influence might threaten the internal validity of the evaluation.

Mortality. This term refers to an effect that might be due to different kinds of subjects dropping out of treatment and control groups as the study progresses from inception to completion. Differences obtained at the posttest might be due to differential mortality between the treatment and control groups rather than to treatment effects. In this situation the necessary condition of equivalent comparison groups would be violated and the internal validity of the study would be threatened. The monitoring of dropouts and the understanding of their selective nature is a crucial and frequently neglected issue in evaluation research. (Maintenance of samples of sufficient size to meet statistical assumptions is discussed in chapter 9.)

Interactions with Selection. The possibility exists that nonequivalent treatment and control groups will interact with other threats to internal validity and produce spurious treatment effects. For example, the influence of maturation on the treatment group might be different from that on the control group if the treatment and control groups are not equivalent. The lack of equivalence would be crucial if the treatment and control groups were comprised of subjects which differed in terms of their rates of change along dimensions relevant to the dependent variables of the study.

Ambiguity About the Direction of Causal Influence. Cook and Campbell (1976) point out that this threat to internal validity is not salient in most experiments since the order of temporal precedence is usually clear. They argue further, however, that the direction of causal influence is a threat to internal validity in simple correlation studies, especially when an equally plausible argument can be developed for the conclusion that A causes B, or B causes A.

Making causal inferences from correlations is a well known problem in the social and medical sciences, with no obvious solutions. The design of controlled experiments with well defined treatment stimuli provides a much firmer basis for making causal inferences. The use of randomized experiments provides an appropriate safeguard against most of the foregoing threats to internal validity.

Although randomized experiments provide safeguards against many threats to internal validity, Cook and Campbell (1976, pp. 228-230), describe five threats to both randomized and quasi-experiments. The reader is referred to the above reference for elaboration of these threats to internal validity.

EXTERNAL VALIDITY

The following are definitions of external validity presented in the literature:

- External validity asks the question of generalizability: To what population, settings, treatment variables and measurement variables can this effect be generalized? (Campbell and Stanley 1966, p. 5).

- External validity refers to the validity with which a causal relationship can be generalized across persons, settings, and times (Cook and Campbell 1976, p. 223).

- External validity deals with the generalizability of a causal relationship to, and across, populations of persons, settings, or times. Generalizability is limited to the classes of persons, settings, or time represented in an evaluation when it ends. These may or may not be reasonable representations of the populations of initial interest (Cook, Cook, and Mark 1977, p. 108).

- External validity refers to the degree of generalizability from one's study to some larger, hypothetical population of interest. (Bernstein, Bohrnstedt, and Borgatta 1976, p. 108).

Both internal and external validity are concerned with the demonstrated, or statistically significant effects of a program. Internal validity is concerned with whether or not the demonstrated effect is indeed due to the content of the program being studied, rather than to some extraneous influence. External validity is concerned with the extent to which one can generalize this (assumed to be) valid causal inference to other populations, settings, and times. External validity is a function of the extent to which a demonstrated program effect can be replicated with different populations, in different program settings, and at different points in time. Empirical evidence for external validity, then, is based on replication. The establishment of external validity is an inductive process; as program effects are verified through replication in different populations and settings and at different times, confidence in the external validity of the program content and influence increases.

Without evidence of replication, external validity is a matter of conjecture. If the target populations and the programs are comprehensively and meaningfully described and outcome variables measure sufficiently similar constructs, it may be possible to estimate the external validity of certain treatments and programs by comparing the results of studies found in the literature.

As a final note on the fragile nature of inferences drawn from one study, we quote from the wisdom of Cook and Campbell:

To infer a causal relationship at one moment in time, using one research setting, and with one sample of respondents, would give us little confidence that a demonstrated causal relationship is robust. A concern with the generalizability of findings across times, settings, and persons will be called a concern for external validity ... (1976, p. 226).

With this understanding of external validity, we turn to a selection of threats to accurate generalization, followed by references for more extensive reading. It is understood that all threats to internal validity are necessarily also threats to external validity.

Threats to Accurate Generalization

Selection. In the best of all possible evaluation worlds, a sample of subjects would be drawn from a defined universe followed by random allocation to experimental and control groups. The latter condition sometimes is obtained, but the former, crucial to external validity, almost never happens.

Thus, selection of the subjects to be allocated to experimental and control groups is plagued by a number of factors which make it difficult to identify and describe the population to which the investigator would like to generalize. Most clients of social and health programs are self selected in ways which are usually unknown to the evaluator and which vary across programs. Under these conditions the investigator must limit conclusions about program effectiveness to subjects using the program; subjects needing treatment, but not seeking it, may comprise quite a different population and may, in fact, be less likely to improve under the same treatment conditions. Generalizing to other treatment settings could only be done on the assumptions that selection criteria were similar over settings or unrelated to treatment outcome.

Other selection factors are selection by excellence, where the bias is likely to be toward positive results, and selection by expedience, where a variety of biases may be introduced. The former is well represented by providing psychotherapy to the young, attractive, bright, and articulate; the latter by the ever present college sophomores in psychology courses.

In quasi-experimental designs, where control groups are sometimes generated through matching procedures, the resulting sample represents a highly idiosyncratic population which is unlikely to be found in other settings.

For more detailed discussion of the effect of selection on external validity, see Campbell and Stanley 1966; Bernstein, Bohrnstedt, and Borgatta 1976; Cook and Campbell 1976; and Cook and Campbell 1979.

Measurement. Definitions of both internal and external validity assume that a significant experimental effect has been demonstrated. However, characteristics of measures selected or developed to assess the effectiveness of programs may prevent demonstration of the necessary significant effect. Measurement error in the form of low reliability and validity may seriously underestimate, and in some cases completely obscure program effects. The use of outcome variables which are insensitive to program content may also lead to incorrect conclusions and in some cases jeopardize continuation of the program.

Sensitization due to completing a pretest or posttest may in itself influence program effects and thus preclude generalization to populations that have not experienced these testing procedures. The content of measures may interact with individual characteristics to produce results which are a function of group composition rather than program content. If characteristics of individuals which interact with measures are unknown, generalizations to other populations which differ on these characteristics would probably be inaccurate, and thus external validity would be threatened.

Excellent discussions of measurement issues as threats to external validity are found in Campbell 1957; Cook and Campbell 1976; Bernstein, Bohrnstedt, and Borgatta 1976; and Campbell and Stanley 1966.

Space precludes an extensive presentation of the many threats to external validity. If the results of evaluation studies are to be taken seriously as a basis for policy determination, then the generalization of results to other populations, settings, and times is crucial. While replication of results is the most convincing evidence for external validity, the cost will almost certainly preclude extensive use of this procedure. Therefore, careful consideration of issues in external validity should be an important part of planning an evaluation study. The references cited above provide an ample basis for understanding external validity.

Statistical Issues. Principles relevant to research design are presented in chapter 9 of the Guidelines. This chapter also provides valuable guidance to the investigator planning the analysis of data generated within the logical structure of various research designs. (For an excellent discussion of statistical conclusion validity, see Cook and Campbell 1976.)

Construct Validity. For a discussion of construct validity as applied to evaluation research studies the reader is again referred to the excellent discussion in Cook and Campbell (1976, pp. 238-245).

We now turn to the description of selected true and quasi-experimental designs. Advantages and disadvantages are considered, followed by references to examples from the literature.

TRUE EXPERIMENTS

In this section a series of true experimental designs will be described, along with selected advantages and disadvantages of each. In particular, threats to internal and external validity will be considered. Since space limitations preclude lengthy discussion of these issues, the reader will be referred to literature providing a more comprehensive and indepth presentation.

True experiments always have one thing in common: subjects, or more generally, experimental units, are randomly assigned to treatments (that is, treatment and control groups). Experimental units can be individual people, classes, census tracts, and so forth, depending on the research design and its purposes. As Cook, Cook, and Mark (1977) point out, random assignment is the defining characteristic of a true experiment and it is important because it assures that the various treatment groups do not differ from each other at the start of an experiment. Random assignment creates the necessary conditions for comparability. That is, conditions are optimal for the comparison of treatment and control groups following the implementation of a given treatment.

Randomization to achieve comparability is differentiated from randomization to achieve representativeness. In the latter situation a random sample is drawn from a defined universe or population of experimental units before randomly assigning the units to treatments or treatment groups. This procedure allows generalization from the sample to the population within known limits due to sampling error and therefore increases external validity. While this is desirable, it is not a necessary condition for a true experiment, which only requires random assignment of experimental units to treatments. Random sampling for representativeness is often impractical in many evaluation research contexts, even though it may facilitate the interpretation and generalizability of results. Its drawback is that all units in the defined universe must be enumerated and assigned a known probability of selection--a major undertaking when that universe is large or geographically dispersed.

NOTATION FOR DESCRIBING DESIGNS

The notation used by Campbell and Stanley (1963, 1966) and repeated by Fitz-Gibbon and Morris (1978) will be used to diagram each design. The symbols are defined as follows:

- "R" means random assignment, that the group was randomly assigned.
- "O" refers to an observation or measurement of some kind, such as an attitude scale, a test of knowledge, a factor score, etc.
- "X" indicates the experimental program being evaluated.
- "C" designates a control group, the group or groups not receiving X. More than one control group is indicated by C_1 and C_2, etc. It is understood that C_1 and C_2 are randomly assigned and in that sense equivalent, unless otherwise stated.
- Dotted lines ("....") drawn horizontally and separating treatment groups, indicates that the groups were not randomly assigned and are therefore not equivalent.
- "E" designates the experimental or treatment group experiencing X.

DESIGN 1: THE PRETEST/POSTTEST CONTROL GROUP DESIGN

	TIME →			
	PRE			POST
Experimental Group	R	O_1	X	O_2
Control Group	R	O_3	C	O_4

Purpose

To assess the impact of treatment X on the experimental group when compared with an equivalent group C which did not receive treatment X.

Ideal Conditions

Subjects, or more generally, experimental units, are randomly drawn from a defined universe or population and then randomly assigned to experimental or control groups.

Subjects are tested under "blind" conditions, where the test administrator does not know if the subjects are in the experimental or control group. Measures are reliable, valid, and sensitive to the influence of treatment X. Sample sizes should be of sufficient magnitude to reject the null hypothesis when the treatment effect relative to the variance is close to the expected size. The treatment effect is constant over time.

Advantages

(1) Allows empirical validation of experimental and control groups as equivalent.

(2) Eliminates many threats to internal validity, including the effects of history, maturation, statistical regression, selection, and possibly testing. See Campbell and Stanley (1963) and Cook and Campbell (1976) for excellent discussions of internal validity.

(3) Minimizes the probability of threats to external validity. For an excellent discussion of external validity, see Bernstein, Bohrnstedt, and Borgatta (1976).

Disadvantages

(1) Difficulty in convincing key people in the evaluation context to allow random assignment to experimental and control groups.

(2) Pretesting might affect experimental results, although recent work by Lana (1969) indicates that such affects can be minimal.

(3) Pre or posttesting may interact with the treatment to produce effect which would not be detected in the population if it had not been tested. See Bracht and Glass (1968) and Bernstein, Bohrnstedt, and Borgatta (1976) for further discussion of these issues.

(4) Differential mortality (dropouts) may affect the equivalence of experimental and control groups.

Data Analysis

The data of this design may be analyzed using several strategies, including:

(1) Analysis of variance or t-test applied to post scores. See Dixon and Massey 1969, chapter 8, pp. 109-126, or a standard statistical text.

(2) Analysis of covariance, with post scores adjusted by pre scores. See Dixon and Massey 1969, chapter 12, pp. 222-236.

(3) Analysis of covariance computed in a multiple regression format. See Cohen and Cohen 1975, chapter 9, pp. 343-402.

(4) For a discussion of statistical analysis of data from nonequivalent group designs, which shed light on some problems in equivalent group design read chapter 4 by Charles S. Reichardt in Cook and Campbell (1979).

Program Example

An example of an early evaluation study which used a pretest/posttest control group design is one done by Swisher, Warner, and Herr (1972). They randomly assigned the entire ninth and eleventh grade classes (216 subjects) of a school to one of four treatment groups, including a no treatment control group. They then asked all subjects to take three pretests--a drug knowledge test, drug attitude scales, and a health habits scale which assessed drug use.

Next, the first treatment group was exposed to a "relationship counseling" drug abuse prevention program; the second treatment group was given a "reinforcement counseling with nondrug abusing models" prevention program; the third was given a "reinforcement counseling with ex-drug abusing models" prevention program; and the fourth (the control group) was given no special prevention program. (The control group received only the standard health class unit on drug abuse.) Each treatment took six weeks and at the end of the six weeks all subjects were asked to take posttests using the same scales. Separate 4 x 2 analyses of variance were run for each of the scales and for the ninth and eleventh graders, including the pre- and posttests for the four treatment groups. The results indicated that all four of the approaches increased the ninth and eleventh grade students' knowledge about drugs, and none of the four changed attitudes or reported drug use of either the ninth or eleventh graders.

This design had the advantage of exposing supposedly equivalent groups to four different prevention approaches so that the efficacy of the approaches could be compared. This design also allowed the evaluators to examine the pretest scores to see whether or not the groups were indeed equivalent before the treatments. It turned out that the pretests did not look equivalent for the four groups (although no statistical tests were performed to test this impression); perhaps an increase in the number of subjects would have solved this problem.

One of the disadvantages of the design, however, was that it did not allow the researchers to assess the effect that the pretesting or posttesting had on the subjects. For instance, the students' increase in knowledge could have been due to an interaction between the pretest and each of the four approaches. The design did not allow the researchers to evaluate this possibility.

DESIGN 2: THE PRETEST/POSTTEST CONTROL GOUP DESIGN WITH AN ADDITIONAL CONTROL GROUP, POSTTEST ONLY

	TIME →			
	PRE			POST
Experimental Group	R	O_1	X	O_2
Control Group One	R	O_3	C_1	O_4
Control Group Two	R		C_2	O_5

Purpose

Same as Design 1, but includes, in addition, an estimate of the influence of pretesting on the posttest.

Ideal Conditions

May be appropriate when: (1) The time between pre- and posttesting is relatively short, (2) the construct measured is a "hot" issue, (3) the subjects are sophisticated, and (4) any time effects of pretesting are expected. However. this design does not distinguish between true effects of X and interactions between X and O_1 (for example, sensitization artifacts). X-O_1 interactions are best assessed with a Solomon Four-Group Design. (See Design 3.)

Advantages

Same as Design 1, with additional strength of knowing the influence of pretesting on posttest results for control groups.

Disadvantages

Requires a second control group which may exceed the capacity of the subject pool to get adequate N's and the tolerance of key personnel in the evaluation context.

Data Analysis

Postscores could be compared by applying one way analysis of variance, followed by a "gap" test to identify the number of populations. One could also test for differences between the two control groups at the posttest. If the null hypothesis was accepted, and no differences due to testing were inferred, then the methods for Design 1 could be applied.

Program Example

This design could have been used in the Swisher, Warner, and Herr (1972) study if they had had a much larger subject pool to start with. Specifically, the subjects from each grade would be randomly assigned to five treatment groups instead of to just four. The fifth group would receive no special prevention, just as the fourth group (control group) did not in the Swisher et al. (1972) study. However, unlike the fourth treatment group, the fifth treatment group would not take the three pretests. The posttest results for the fourth and fifth groups could then be compared to see if the pretest affected the posttest.

The reason that the subject pool must be much larger than Swisher's (1972) in order to use this design is that researchers must be able to assume that the fifth group's pretests would have been equivalent to those of the fourth group's if the tests had been given. It is essential to be able to assume this because the design does not test that assumption.

DESIGN 3: THE SOLOMON FOUR-GROUP DESIGN

		TIME →		
		PRE		POST
Experimental Group One	R	O_1	X_1	O_2
Control Group One	R	O_3	C_1	O_4
Experimental Group Two	R		X_2	O_5
Control Group Two	R		C_2	O_6

Purpose

To extend Designs 2 and 3 to explicitly consider the influence of pretesting on both the experimental and control groups.

Ideal Conditions

This more demanding design may be selected when the influence of pretesting is thought to be strong and might interact with X.

Advantages

(1) As Campbell and Stanley (1963) point out, the effect of X is replicated in four comparisons. And if $O_2 > O_1$, $O_2 > O_4$, $O_5 > O_6$, and $O_5 > O_3$, the strength of the inference regarding the influence of X is increased. (2) The interaction of pretesting and X, along with the main effects of X and pretesting, are determinable.

Disadvantages

This design requires more resources in terms of the number of experimental units and it may be harder to win the approval of administrators. Also, because of the greater number of groups and subjects, the likelihood of differential dropout effects increases.

Data Analysis

Since there is no single procedure for analyzing the six observations simultaneously (See Campbell and Stanley 1966, pp. 25), the data of this design are usually analyzed as a two by two analysis of variance, with interaction. Postscores are the dependent variable. (See Winer 1971, chapter 5, pp. 309-428; Dixon and Massey 1969, chapter 10, pp. 150-192).

Program Example

An example of this design would be the further extension of Swisher, Warner, and Herr's (1972) study to include a sixth, seventh, and eighth randomly assigned group. These groups would not take the pretest but would be exposed to one of the three different prevention programs and the posttests. The interactions between the effects of the pretests and the effects of the different prevention programs could then be examined.

DESIGN 4: THE POSTTEST-ONLY CONTROL GROUP DESIGN

		TIME →	
			POST
Experimental Group	R	X	O_1
Control Group	R	C	O_2

Purpose

To assess the impact of X by comparing equivalent groups with and without X.

Ideal Conditions

When it is difficult to pretest or it is felt that pretesting would seriously influence posttest results or the reaction of subjects to X. See Campbell and Stanley (1963) for a more detailed description of this design.

Advantages

The demands of pretesting are eliminated, along with the potential influence of pretesting on the outcome.

Disadvantages

(1) Since it is not possible to test X and C groups for equivalence, the random assignments procedure is crucial.

(2) It is not possible to obtain direct measures of individual changes due to X.

(3) Biases due to attrition are difficult to estimate due to lack of pretest information on subjects.

Data Analysis

This design requires a test of the differences between post score means. (See Dixon and Massey 1969, chapter 8, pp. 109-126). If the size of the experimental effect is required, a point-biserial correlation coefficient could be computed. (See Cohen and Cohen 1975, pp. 35-37).

Program Example

An example of this evaluation design was presented by Bry (1977). In 1973, eighty seventh graders were identified as being at risk for drug abuse because they were having related adjustment problems. These high risk seventh graders were randomly assigned to an experimental or a control group. Some unobtrusive premeasures were taken at that time, but none could be taken reflecting the primary variables of interest, that is, drug use, school dropout, criminality, and employment. The main reason no premeasures were taken was to prevent the control group from being affected by being identified and "measured". Also, by and large, the subjects were too young to be exhibiting many of the behaviors of interest. Since a prevention program was being tested, it was assumed that the behaviors the program was designed to prevent had not yet occurred.

The students in the experimental group were given a long-term drug abuse prevention program called the Early Secondary Intervention Program (ESIP). Three and one-half years later, interviewers who were not connected with the prevention program and who did not know which young people had been in the program were sent out to find them. Sixty-three of the young people were found, half from the experimental group and half from the control group.

Chi square and t-tests were used to assess the differences between the two groups. The levels of behaviors of interest that were reported by the control group were assumed to be what would be expected for high risk young people who had experienced no preventive intervention. Thus, it was assumed that any significant difference between the reports of the experimental group and the control group were due to the prevention program.

The results indicated that ESIP reduced the reports of serious criminal behavior (property destruction and grand theft), nonmarijuana drug use (pills, hallucinogens, and heroin), and school dropout and retention (having to repeat a grade). The results also indicated that ESIP increased the amount of teenage employment.

There was no difference between the groups' reported alcohol or marijuana use. As with all of the study's results, the researcher must question whether or not the two treatment groups were equivalent before the prevention program began and whether or not the subject attrition was differential between the two groups. This research design does not allow the researcher to examine these two questions.

DESIGN 5: FACTORIALLY ORGANIZED, PRE/POST CONTROLLED DESIGN

			TIME →		
			PRE		POST
Experimental Group	M	R	O_1	X	O_2
Experimental Group	F	R	O_3	X	O_4
Control Group	M	R	O_5	C	O_6
Control Group	F	R	O_7	C	O_8

Purpose

To test hypotheses regarding variation in response to treatment among categories or groups of subjects. Both the determination of appropriate categories and their order of response to treatment X may derive from theory, a review of empirical literature, or the experience of the investigators.

Another purpose is to test for the possible interaction of treatment/control conditions with categories of subjects, such as sex, as indicated in the diagram above.

Ideal Conditions

When there is sufficient reason or evidence to indicate that certain categories of subjects (males for example) will respond more favorably to treatment X than other categories of subjects (females).

Advantages

The pre/post controlled design, with only two comparison groups, may prove in a number of instances to be simplistic in its logic. That is, it does not allow for the identification of "good" and "poor" responses to treatment X in the experimental group which, when added together, will yield a mean score similar to that of the control group and thus support the null hypothesis and the inference that treatment X is ineffective. In fact treatment X may be quite effective for certain categories of subjects. The challenge for evaluators, as they design their studies, is to create a design which allows the identification of subject groups which vary under the influence of treatment X. The advantage of this design is that it allows the possibility of such differences to emerge.

Disadvantages

More subjects are required if the cell sizes are to be sufficiently large to adequately test for interaction affects.

Other than the need for additional subjects, this design would generally not place a greater burden on the research context. It would require the evaluators to think about their studies in terms of more differentiated hypotheses regarding outcome and of the categories of subjects which would be most likely to yield variability in response to treatment X. This procedure may lead to more carefully conceived studies.

Data Analysis

The data of this design could most simply be analyzed in a 2 X 2 analysis of variance with postscores as the dependent variable. Sex and experimental versus control would serve as independent variables. The preferred analysis would adjust postscores for prescore effects in a factorially organized analysis of covariance. (See Winer 1971, pp. 781-796). An assumption of this analysis is homogeniety of regression of postscores on prescores over the four groups. Assumption failure may cause differential adjustment of cell means, in which case the results are not valid. (See Cohen and Cohen, 1975, pp. 375-376. See also Winer 1971, pp. 594-599 and 772-775).

Program Example

Bien and Bry (in press) used this design when they were faced with scheduling non-equivalencies in implementing an experimental school based, drug abuse prevention program for high risk students. One-half of the subjects could be taken out of their morning classes to attend the prevention program, and the others could participate only in the afternoons. Experience suggested that the young people might be more receptive to the program in the morning than in the afternoon. Thus, half of the subjects that could be available only in the morning were randomly assigned to the experimental group and the other half were assigned to the control group, with the same procedure being followed for those subjects who were available only during afternoons.

The effects of the program were assessed by examining whether or not it had reduced the correlates of drug abuse, that is, low student grades, poor attendance, and frequent discipline referrals. Thus, the premeasures and postmeasures of each subject were compared and an improvement score (or gain score) for each subject was calculated for each variable. Then, two-way analyses of variance were performed on each variable with treatment group and time of program (morning or afternoon) as the main effects.

The results indicated that the prevention program had reduced deterioration in the grades of the students who could be assigned to the program in the morning, but not of the students assigned to the program in the afternoon. These results have been replicated (Bry and George, 1979) so it appears that the factorially organized design allowed the investigators to learn about an important variable which leads to differential treatment effects.

DESIGN 6: FACTORIALLY ORGANIZED, REPEATED-MEASUREMENTS CONTROLLED DESIGN

		TIME →								
			T_1		T_2		T_3			T_k
Experimental Group	M	R	O_1	X	O_2	X	O_3	. . .	X	O_k
Experimental Group	F	R	O_1	X	O_2	X	O_3	. . .	X	O_k
Control Group	M	R	O_1		O_2		O_3	. . .		O_k
Control Group	F	R	O_1		O_2		O_3	. . .		O_k

Purpose

To assess the influence of treatment X when applied sequentially over k treatment sessions. Specifically, the following null hypotheses may be tested:

- There are no experimental/control effects
- There are no group (for example, sex) differences
- There are no group/experimental condition interactions
- There are no changes in subjects over treatments or trials (all subjects pooled)
- There are no trial by group interactions.

Curve fitting may also be applied to determine the best fitting functions to the means of the four groups over k trials.

Ideal Conditions

This design is used in the study of treatment, particularly drug or educational effects, where observations of improvement or change are made at periodic intervals. It easily lends itself to testing the hypothesis of differential reaction or differential effects. If relevant theory is strong enough, specific hypotheses focused on selected group difference (beyond experimental/control differences) can be tested. It may at times be useful to examine group differences and group-by-trial interactions in a series of post hoc analyses to generate hypotheses for subsequent studies.

Advantages

Carefully planned studies, guided by relevant theory and specific hypotheses, may help to identify group specific treatment effects otherwise obscured by considering only samples from general rather than stratified populations. Thus, internal validity may be enhanced through subject classification.

The observation of subjects over trials provides a powerful basis for assessing treatment effects as time and exposure increase. From examination of change over time it is possible to estimate the shape of the change function and to estimate maximum and minimum periods of experimental influence. Group differences with regard to change can, as indicated above, be studied either as hypothesized differences or as post hoc hunches.

Disadvantages

The maintenance of subjects over a series of k observations is always a problem and the evaluator must anticipate and estimate the dropout rate during the planning stage of the study. Obviously, to make inferences about the influence of the full range of treatment, the experimenter must retain a sufficient number of subjects to reach valid conclusions. Classification of subjects into theoretically meaningful categories further reduces the degrees of freedom available and the power of the statistical tests. Therefore the evaluator should create a research structure in which the subject pool is large enough to accomplish the goals of the study.

Since dropouts from studies are always a threat to the internal and external validity of an evaluation study, the experimenter may wish to study this troublesome phenomenon in order to understand its selective nature and possible influence on the study. It would seem that this area of study, crucial to most evaluation studies, has been systematically nelected.

Data Analysis

This design provides for a number of tests of significance including:

(1) Differences between groups
(2) Interactions of groups
(3) Differences within subjects
(4) Interactions of groups with subjects
(5) Following the above tests, trends in repeated measures may be fit to linear, quadratic, or higher order functions.

For a thorough presentation with illustrations, see Winer 1971, pp. 514-59. To analyze repeated measures in a multiple regression format see Cohen and Cohen 1975, pp. 403-425.

Program Example

Feldi (in preparation) used this research design to examine the effects of a drug abuse prevention program which had been offered over a three year period of time in two different school systems. Subjects were randomly assigned to either an experimental or control group at the beginning of the three years. Specifically, of interest was whether there was an optimal program duration and location. The two school systems were quite different so there was reason to believe that the program might have differential effects.

Feldi looked at the effects of the prevention program upon the correlates of drug abuse, such as low grades, poor attendance, poor promptness, and discipline referrals. She calculated gain scores for each subject on each correlate at the end of the first, second, and third years of the program. A two-way repeated measures analysis of variance for unequal N's was performed, with treatment and location as the main effects.

The results indicated that the program reduced the amount of tardiness among students in both school systems. This effect was more pronounced in one school system than the other, and was the greatest at the end of the third year of the program. There was no evidence that other variables were affected by the program, but there were some very interesting findings about what happens to the school records of the students in the two different school systems over time. Thus, the experimental design enabled Feldi to examine questions beyond the mere effects of the prevention program.

There is almost always subject attrition in a repeated measures design, but the presence of the pretest in this design enables the researcher to test whether the subjects lost from the experimental group had premeasures similar to those lost from the control group. Feldi assumed that there had not been selective attrition since all of it was caused by families moving out of the school system; however, she could have validated this assumption.

QUASI-EXPERIMENTAL DESIGNS

In most cases the results of true experiments provide the firmest basis for making causal inferences or for estimating the effectiveness of a particular treatment or education program. Other things being equal, the key issue in comparing true and quasi-experimental designs and finding the former superior, is the degree of equivalence between the experimental and control groups in the two types of designs. Because we know so little about the many factors which might influence program outcomes, subjects are randomly assigned to experimental and control groups in an effort to randomize, and therefore equalize, the many possible influences. Randomization is the preferred method of obtaining equivalence. However, it is not always possible to have a control, or even a comparison group. Even in the latter circumstance it may be better to have some data as a basis for conclusions rather than to rely completely on opinion, conjecture, or political expediency.

The purpose of quasi-experimental designs is to approximate, as closely as possible, those conditions which enhance the validity of the true experimental design. A major factor

is the selection of a comparison group so that its equivalence with the experimental group is maximized. Therefore, in planning a quasi-experiment involving a comparison group the investigator must go to great lengths in selecting a sample from the most equivalent population. If, then, differences between the experimental and comparison groups are established, there is a firm basis for inferring treatment effects rather than differences due to selection. In other words, once differences are established it is incumbent upon the investigator to rule out possible explanations of experimental control differences other than treatment effects. The investigator is looking for the most plausible set of inferences and conclusions.

This section does not include a full description of a well known design, the "One Group Pretest-Posttest Design," because of its inherent weaknesses. Campbell and Stanley (1966) describe it as a "bad example" because it is so vulnerable to extraneous variables that can jeopardize internal validity. History may compete with X to influence pre/posttest differences. When one considers the many other factors in the lives of students, participants, and other groups that can influence these differences, the need for a control group becomes apparent.

Another threat to the internal validity of this design is maturation, or some change in the individual which, in the absence of a control group, may erroneously be attributed to X. These and other disadvantages of this design are discussed at length in Campbell and Stanley (1966, pp. 7-12) and Cook and Campbell (1976, pp. 247-248). However, Campbell (1979) has presented a cogent argument for the potential benefits of the "One-Shot Case Study" which runs counter to his previous stand. This discussion is equally appropriate to the one "One Group Prettest-Postest Design."

With this as background, we turn to a selection of quasi-experimental designs.

DESIGN 7: THE TIME-SERIES EXPERIMENT

	TIME →								
	T_1	T_2	T_3	T_4		T_5	T_6	T_7	T_8
Experimental Group	O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8

Purpose

To identify discontinuity in a sequence of observations resulting from the introduction of an experimental change, X. As described by Campbell and Stanley: "The essence of the time-series design is the presence of a periodic measurement process on some group or individual and the introduction of an experimental change into this time series of measurements, the results of which are indicated by a discontinuity in the measurements recorded in the time series" (1966, p. 37).

Ideal Conditions

This design is particularly appropriate when observations are unobtrusive and routine, and respondents are not reacting to repeated testings. As Cook and Campbell point out, "The most important feature of time-series designs is that there be a sufficient number of pretest data points covering a sufficiently extended time period so that all plausible patterns of variation can be ascertained" (1976, p. 275).

Thus, sufficiently reliable and valid measurements made over an extended period of time, followed by an event or experimental influence and further periodic observation are crucial to this design. Such conditions frequently obtain in schools, clinics, and other institutions as well as in defined populations observed over time on such factors as mortality and recidivism rates.

Advantages

(1) If accurate records are available, the data needed to describe performance (or whatever) prior to an unanticipated change or planned experimental effect can be collected retrospectively. At times, as in the case of mortality and census data, the necessary information is in the public domain.

(2) Multiple observations allow a more comprehensive evaluation of the experimental effect, including maximum and minimum points of influence, the amount of lag following implementation and, more generally, the shape of the response line of best fit.

Disadvantages

(1) The effects of history cannot be determined without a control group or comparable institution.

(2) The nature of the experimental effect may be limited to those populations experiencing repeated testing. This disadvantage would, of course, be restricted to subjects responding to reactive measurements and would not apply, generally, to the routine data collected by an institution.

(3) The effects of testing may at times be a factor, although such effects should become apparent before X occurs.

(4) The interaction of X with testing and with selection may interfere with a clear inference regarding the influence of X.

(5) This design is vulnerable to many of the threats characteristics of designs without control groups, however selected. Campbell and Stanley (1966) point out, "Where a better controlled design is not possible, we will use it." Confidence in the results are enhanced if the findings can be replicated in several settings.

Data Analysis

For further references, see Glass, Willson and Gottman (1975); Campbell and Stanley (1969, p. 42); and Cook and Campbell (1979).

Program Example

An interesting example of a time series design appears in the National Institute of Drug Abuse Research Monograph No. 17, Research on Smoking Behavior. A researcher used the design to examine the effect of television antismoking ads on per capita intake of cigarettes. A dramatic discontinuity in per capita intake was found during the three years when the antismoking ads were shown. This drop was well beyond the normal variation in per capita intake that occurred both during the six years before and after the three year period of ads. The discontinuity is so clear that statistics are not even needed to evaluate it.

The research monograph report is not the primary source, but one can assume that the basic requirements for the time series design have been met. Specifically, the measurement of per capita intake of cigarettes is probably an unobtrusive measure that is collected routinely in a periodic manner whether or not an anti-smoking campaign is being waged.

The main disadvantage of this design is the lack of a control group. It is not possible to know whether or not there was another cause of the drop in intake besides the ads. There is no way to knowing, for instance, whether or not the purchase of all "luxury" items decreased during that period due to a recession or inflation.

DESIGN 8: THE NONEQUIVALENT CONTROL GROUP, PRETEST/POSTTEST DESIGN

	TIME →		
	PRE		POST
Experimental Group	O_1	X	O_2
Control Group	O_3	C	O_4

Purpose

To assess the impact of treatment X on the experimental group when compared with a control group.

It is crucial to realize that the E and C groups are not formed by random assignment from some defined population. Nonrandom assignment is indicated by the dotted line between the two groups, following the notation used by Campbell and Stanley (1969) and subsequently by Fitz-Gibbon and Morris (1978).

Ideal Conditions

This design may generate useful information when it is not possible to form E and C groups by random assignment. It is very important to select a control group which is very similar to the E group, particularly on characteristics known or thought to be related to the outcome or dependent variable. In other words, every effort should be made to maximize the equivalence of groups E and C, since the internal and external validity of the evaluation depends to a large extent on the assumption of equivalence.

Advantages

(1) A comparison of E and C groups provides information on their equivalence, but only on the pretest variable, however.

(2) This design has distinct advantages over the one-group, pretest/posttest design, in that it provides at least some basis for comparing pre/post change.

(3) By releasing the demands for randomization it may be possible to carry out this design in research contexts which would not accept randomization of subjects. Assuming that somewhat qualified information is better than none, successful completion of this design may yield valid enough results to facilitate decision making.

(4) Evaluation research should not be judged solely in terms of results. The procedures involved in the completion of this type of study will frequently identify important characteristics of treatment or educational processes that will serve as a basis for program improvement.

(5) If the E and C groups are similar on characteristics theoretically linked to the outcome variable and on pretest scores, it may be assumed that the design controls for the main effects of history, testing, and maturation. See Campbell and Stanley (1969, pp. 47-50) and Cook and Campbell (1976) for a more detailed discussion of these issues.

Disadvantages

(1) E and C groups may not be equivalent on salient characteristics related to the dependent variable even though they are similar on pretest scores. Thus, selection may affect internal validity.

(2) Interaction effects of the E and C groups with maturation or testing are more likely to occur when the groups are not formed by random assignment. Such intervention effects, according to Campbell and Stanley (1969) could be mistaken for the influence of X and therefore pose a threat to internal validity.

(3) Selection of either group with extreme (high or low) scores on the pretest variable will lead to regression to the mean in that group and result in another possible threat to internal validity (Campbell and Erlebacker 1975).

Data Analysis

Analysis of covariance is the preferred analysis for this design, although there are many pitfalls in make causal inferences (as thoroughly reviewed by Charles S. Reichardt in Cook and Campbell (1979) chapter 4, pp. 147-205). Chapter 9 of Cohen and Cohen (1975), which presents analysis of covariance in a multiple regression format, could be appropriately applied to data of this design.

Program Example

Visco and Finotti (1974) used this experimental design when they were asked to evaluate a New York City School's drug abuse prevention program for high risk students which had been in operation for two years. Thus, they could not ask for random assignment of subjects to experimental and control groups.

Instead, school personnel looked at each program student's premeasures (absences, grades, drug related referrals, other antisocial behavior) and went through the school records to find another similar student who had not been in the program. The new student became the nonequivalent matched pair of the program student. Each program participant was thus paired with a "control" subject.

Visco and Finotti used t-tests for matched pairs to learn whether or not the program students' grades, absences, drug related referrals, and other antisocial behavior were different at the end of the first and second program years than those of their matched pairs. They concluded that the prevention program had a positive effect because most of these comparisons were statistically significant, with the program groups showing the most positive postmeasures.

These results cannot be accepted without qualification, though, because of the weaknesses in this experimental design. Close examination of the pretest data shows that the experimental and control groups were probably not equivalent before the program began, despite their painstaking matching procedure (Bry 1978). The control students had more absences, lower grades, and fewer drug and antisocial referrals than did the prevention program students. More control group subjects than experimental subjects dropped out of school before the study was over; thus, it looks like the experimental and control groups were from two different populations. Consequently, these population differences could have accounted for the differences between the groups on the posttests.

DESIGN 9: THE TIME-SERIES DESIGN WITH A NONEQUIVALENT CONTROL GROUP

	TIME →					→			
	T_1	T_2	T_3	T_4		T_5	T_6	T_7	T_8
Experimental Group	O	O	O	O	X	O	O	O	O
Control Group	O	O	O	O	C	O	O	O	O

Purpose

To determine the effect of program X on the E group when observed at regular intervals prior to and following the implementation of X, and when compared with a nonequivalent control group, C.

Ideal Conditions

When the effects of history and maturation are likely to influence group response to X and the effects of making a number of observations at regular intervals are minimal. This latter condition is most likely to occur when the observations are unobtrusive and routine.

This design may be used in the study of institutions where archival data and control institutions are available. Planned propective studies should yield the best results, although routinely collected archival data of high quality coupled with policy, legislative, or institutional change may be studied to advantage with this design.

Advantages

(1) As Campbell and Stanley (1969) point out, this design gains in certainty as the number of sequential observations increases, and from comparison with periodic control group observations.

(2) The effects of history and maturation are more accurately assessed with the greater number of observations made prior to the implementation of X.

(3) Changes in the influence of X can be estimated from the observation made at regular intervals following X.

(4) Comment: Campbell and Stanley (1963) evaluate this design as "an excellent quasi-experimental design, perhaps the best of the more feasible designs."

Disadvantages

(1) Under certain experimental conditions the collection of repeated measurements may introduce testing effects of considerable magnitude.

(2) Unless the C group is carefully selected, an interaction of the E and C groups with testing may occur due to differences between those groups. This is most likely to occur under reactive (testing) conditions.

(3) Under reactive conditions a testing-X interaction is possible.

(4) This design is subject to validity threats characteristic of nonrandomly assigned subjects. Careful and thoughtful selection of control groups or comparable institutions is emphasized.

Data Analysis

See Glass, Willson, and Gottman (1975), Campbell and Stanley (1969), and Cook and Campbell (1979).

Program Example

This design could be used to examine some alternative explanations for the discontinuity of per capita cigarette intake in the previously cited time series study. One alternative explanation for the discontinuity is that some economic factor, such as world recession, may have accounted for the results. A comparison of the United States data (where the anti-smoking ads were sanctioned and funded for just three years) with per capita cigarette intake data from a similar country where similar policy shifts did not occur would shed light on the recession hypothesis.

Another alternative explanation is that normal variation within a ten year period may account for the observed discontinuity. This hypothesis could be examined by comparing the 1963-1976 data, where the discontinuity occurred, with another thirteen year period of data, such as 1953-1966. The amount of information that can be gained by using this design depends on the investigator's ability to choose a nonequivalent control group which is similar to the experimental group, on dimensions which are most related to the hypothesized effects of the prevention program.

QUALITATIVE STRATEGIES IN EVALUATION RESEARCH

Now that we have dealt with designs that emphasize quantitative evaluation, it is as important to examine designs which use qualitative methods. Most often, such methodologies are used in the formulation of specific research questions which can then be further examined through quantitative methods. However, it is important to emphasize that the findings which have come out of qualitative studies in various fields have played a major role in policy development.

Researchers usually start with an explicitly defined dependent or criterion variable and devote most of their effort to examining competing hypotheses and ordering explanatory factors so as to most effectively account for variation in that variable. As indicated by Bernstein (1976), however, the sequence is often reversed for program evaluation: one begins with a fairly well defined independent variable--the program to be evaluated--and proceeds with the task of comprehensively describing its range of delivery, tracing its impact, and measuring its outcome in terms of often only vaguely defined sets of goals and objectives.

This reversal suggests a place for qualitative methods in evaluative study, including various forms of observation, case history, and interviewing techniques. Such strategies are likely to prove useful for purposes of sensitizing evaluators to complexities of program success and failure, and for more traditional purposes of developing and implementing research efforts (Twain 1975). They offer specific and oftentimes dramatic illustrations of the range of program impact, and concrete insights into why and to what extent efforts exceed or fall short of meeting stated purposes. Information generated by such methods may serve to orient and redirect ongoing delivery efforts. Further, where attention is given to issues of reliability and validity, qualitative data can play a crucial role in developing effective instruments and perceptive analytical designs. There is, one might argue, no substitute for working from a thorough understanding of "target" populations and programs designed to serve them. Qualitative strategies offer the best means for developing such an understanding.

This section provides a brief outline of qualitative methods likely to prove useful in evaluation studies, drawing from the relevant literature where such strategies have been successfully employed. Specific examples of applications in the general area of drug abuse and prevention programs are relatively rare, however, reflecting what amounts to a general mistrust of information generated by so-called "soft" methods, particularly where findings may be used for making decisions to either extend or terminate efforts. Given such purposes, evaluators tend to rely on information that easily lends itself to quantification and statistical summary, avoiding problems regarding the "impressionistic" or "subjective" nature of data generated by qualitative techniques. Unless one is willing to argue that numbers speak for themselves, however, the utility of qualitative data remains an open issue for debate. Its use as an adjunct to information generated by standardized procedures and

interpretation of statistical results seems to offer a reasonable and easily defensible position (Gurel 1975).

OBSERVATIONAL METHODS

Observational methods fall into two general categories, depending upon the degree of researcher involvement in events or situations being studied. At one extreme are what might be referred to as "pure" observational strategies, aimed toward developing an "unobtrusive" description of subject attitudes and behaviors.

At the other extreme are "participant" observation approaches, involving direct researcher involvement in processes studied, aiming to move beyond simple description toward an indepth understanding of motivational factors and situational variations affecting the behavior of subjects.

Following Webb, et al. (1966), "pure" observation seeks to eliminate "reactivity" associated with measurement procedures typically used in quantitative approaches. The strategy begins from a realization that data collection itself has the potential of altering phenomena being studied and proceeds in terms of information unintentionally generated as a byproduct of typical behavior. Among the types of measures developed from such information are:

- Physical traces--including "erosion" and "accretion" indicators. The former might involve selective depletion of literature made available to enrollees as reflecting specific topics of interest. The latter might involve an ongoing examination of seating patterns as revealing information about degree of participation, and so on.

- Content measures--involving examination of written records, printed materials pertaining to program goals and objectives, though produced without explicit knowledge of the study.

- Concealed measures--including general impressions of expressive movements, verbal cues and phrases, clothing styles, and so forth.

Examples of the use and development of such measures can be culled from a variety of studies including Lindesmith's (1938) classic investigation of drug culture slang, Clausen's (1957) description of drug use and adolescent personality types and, more recently, Agar's (1975) analysis of linguistic materials collected from heroin addicts. An overview of the current status of pure observational measures and strategies has been compiled by Sechrest (1979).

It is further possible to adapt "pure" observational strategies to a quasi-experimental framework by introducing objects or confederates into specific events or situations being studied in such a way as to focus or direct attention to topics of immediate interest or concern. Ethnomethodologists have adapted such strategies to the study of group norms and mores by, for example, documenting reactions to purposefully executed violations of social expectations (Garfinkel 1967).

The collections of such information may be enhanced by the use of tape recorders, hidden cameras, one way mirrors, and the like, offering a permanent or verifiable record of observations obtained and analyzed. Where incidence and prevalence are of interest, checklists and tabulation sheets might be employed.

The strength of "pure" observational strategies, then, lies in their promise of generating "firsthand" information, free from distortions arising from the use of questionnaires or direct researcher intervention. Their weakness is largely a function of the kinds of situations amenable to observation, typically limited to behavior in public places (Goffman 1963). Other problems arise from the high time investment required for data collection and possibilities of systematic oversight. Multiple observers may help overcome boredom and fatigue factors affecting the reliability of observation information, and also offer a means of avoiding

preconceptions involving single researchers. Such an approach generally enhances the internal validity of "pure" observational data. The information, however, remains almost purely descriptive, providing little insight into motivational factors underlying behaviors observed.

Participant observation offers the best alternative to overcoming the limitations of "pure" observational data. Like pure observation, the method is deliberately unstructured in design at first, becoming progressively more structured as the study proceeds. It assumes that the researcher will adopt some degree of commitment to the perspective of those being studied. By sharing in their daily experiences, different roles are available depending upon the degree of commitment (Gold 1969).

The "complete participant" role refers to a situation where the researcher is wholly concealed in his purpose, acting as a fullfledged member of the group being studied. The "participant-as-observer" role refers to a situation where the researcher makes his presence and intentions known to group members and attempts to form a series of relationships with subjects who, basically, serve as informants. The former approach is exemplified by Becker's (1964a) report on marijuana use and, more recently, by Agar's (1973) study of the heroin subculture. The latter approach is exemplified by Ray's (1964) study of heroin addiction and, more recently, Hughes', et al. (1971) investigation of buyer and dealer role-playing in illicit drug traffic. A general assessment of the strengths and limitations of both approaches as applied to drug studies is provided by Plant and Reeves (1976). A review of participant observation techniques in alcohol research is provided by Clark (1979), and Carvan's (1966) ethnography of a bar provides a classic example.

Both roles are subject to change as a study proceeds, identifying maturation as a key concern in assessing the completeness and validity of findings. Subjects, for example, may prove willing to provide more accurate information as their relationship with the researcher develops. Similarly, the observer may become more aware of subtle behavioral patterns and cues as s/he becomes more familiar with the subjects.

Data based on the "complete participant" approach may also be biased in terms of limitations imposed on settings in which it is gathered, and categories of participants involved. Data collected by the "participant-as-observer" approach, in turn, may suffer from reactive effects, including social desirability factors and incomplete or misleading reporting. On the positive side, both methods offer a comprehensive understanding of behavior of interest as it occurs in natural settings, including insight into patterns and processes of group interaction. Such information is typically ignored in quantitative approaches to program evaluation, not because of irrelevance, but rather, because it is so difficult to collect and parsimoniously summarize.

CASE HISTORY METHODS

Case history methods are closely linked to the participant-as-observer approach, both, in effect, relying on the use of "informants" to provide relevant data covering a broad range of topics of interest. Such similarities are evident in studies conducted by Preble and Casey (1969) and Hughes, et al. (1971), each employing information provided by subjects regarding patterns of "drug-copping." The case history method, in pure form however, emphasizes the personal experiences and definitions held by a single person, group, or organization. The hallmark of the method lies in its ability to capture the development (or unfolding) of events over time, offering data regarding the social history of occurrences as seen by those involved. As applied to an evaluation of drug prevention, such information could serve as an invaluable source of insight into the range of program impact and the study of differential effects.

Case history materials include records and/or documents either previously in the possession of informants or, more typically, specially prepared at the request of the researcher. A relevant example of a combination of both data sources can be found in Street and Loth (1953). Information can be further categorized as follows:

- Primary--information based on the direct involvement of informants including personal letters, diaries, eyewitness reports, and so forth.
- Secondary--information indirectly referring to the experiences of informants and other individuals or groups, and data sources such as actuarial and administrative records, program listings, attendance sheets, and a wide range of hearsay reports.

Ideally, primary information should be cross-validated in terms of available secondary sources for purposes of establishing its temporal and factual assertions. Information derived from both sources may be summarized and employed in the following forms:

- Complete--involving a total summary of the informant's experiences.
- Topical--emphasizing selected aspects or phases of the informant's experiences.
- Edited--focusing on comments, impressions, opinions drawn from complete or topical reports.

Like information derived from participant observation, case history data are subject to maturation and reactivity effects, each in turn providing possible threats to internal validity. Also, where secondary sources are used the researcher must be cautious of bias and incompleteness introduced by the simple fact that such records and reports were prepared for specific purposes other than those of the given study. Generally, the greater the confidential nature of case history data, the higher their validity. More generally, however, the best role is not to place too much emphasis on any single document and always look for corroborative evidence when summarizing either primary or secondary reports.

INTERVIEW METHODS

Interview methods include administration of psychometric instruments to individuals and groups, surveys, telephone and personal interviews, completion of forms, orally administered instruments, and computer interviews. The latter is a relatively new technology which allows user paced branching interviews with online data collection and reduction. Applications of this technology to a variety of interview situations and topics may be found in Slack, et al. (1976, 1966), Maultsby and Slack (1971), Gustafson, et al. (1977), de Domal (1973), Bleich (1971), Fisher et al. (1977), and Klitzner, et al. (1979). Although such applications are relatively new, they hold promise for reducing the biases sometimes found in face to face interviews, where demand characteristics and social desirability can reduce the reliability and validity of findings.

Interviews are a commonly used strategy in evaluation research, providing a major proportion of information regarding the characteristics of program participants, their knowledge, attitudes, and behaviors, both before the program and afterward. Interviews may be classified by their degree of structure (or standardization), ranging from unstructured, to focused, to structured (Merton and Kendall 1946). Each type is most appropriate for gathering certain kinds of information and each is marked by its own peculiar advantages and disadvantages.

Interviewing allows for flexibility in data collection. Data can be collected from large numbers of individuals in a single category (students in a school, a sample of community members, or program participants) at one or more points in time (before, during, or after a program is in effect). Compared to direct observation, interviewing tends to be less costly and to yield information that is more amenable to quantitative analysis. Furthermore, the highly controlled nature of the collection process makes the data well suited for experimental methods of evaluation.

The primary weakness of the interview lies with the difficulty in establishing its external validity. The context of the interview might cause subjects to systematically and knowingly bias their responses. Problems that can arise include deliberate falsification to avoid

embarrassing or self-incriminating behavior, an inability to accurately recall past behavior, inaccurate placement of the time period for which information is requested, and difficulty in making accurate counts of frequent behaviors, such as smoking or some forms of vandalism.

The cost and feasibility of data collection by interview techniques are influenced by the availability of existing instruments. Well constructed and tested instruments have predetermined reliability and validity parameters, often for settings that closely fit the context of the evaluation. Particular attention must be paid to the characteristics of the audience for whom such instruments are intended. Reliability and validity are functions of both the instrument and the subjects tested. In addition to such obvious issues as age, sex, comprehension level, and so on, a major concern in the field of prevention is that indicators, measures, and the context of data collection should be responsive to racial, ethnic, and cultural considerations.

When existing instruments are not appropriate, the evaluator has the option of developing one, although this is recommended only as a last resort, given the expense, time, and complexity of the task. Chapter 8 examines the use of interviews in surveys, while chapter 9 discusses issues in instrument development. The following discussion distinguishes types of interviews ranging from qualitative (unstructured) to quantitative (structured), and the advantages of each approach.

Unstructured. Approaching the level of everyday conversation, this type of interview emphasizes individual perspectives and experiences in much the same way as does the case history approach. There is no attempt to parallel any formal schedule of questions over subjects. It is best suited for exploratory study and might be used as a first step in developing more specific interviewing schedules.

Focused. This method employs a standardized set of questions without any standardized schedule. The researcher works from a list of topics or issues involving information required from each subject, but the phrasing and ordering of questions is redefined to best accomodate each particular respondent. Like the unstructured approach, focused interviewing allows for spontaneity of response and enables the researcher to probe a variety of topical areas. The method is well suited for purposes of investigating issues of selective use and differential impact, and seems particularly appropriate for developing and reformulating survey instruments.

Structured. These interviews are based on a standardized schedule where the wording and ordering of questions is identical for each respondent. Differences in response, thus, can be attributed to differences between respondents rather than to differences in applying the item schedule. Results lend themselves to quantification and statistical summary, similar to results from written surveys. The approach is best suited for gathering background information on "target" groups and for purposes of testing hypotheses.

Substantively relevant examples of interview approaches to the study of drug use and abuse are provided by Becker's (1964b) investigation of marijuana subculture, Ray's (1964) examination of abstinence and relapse among heroin addicts, and Winick's (1964) descriptions of physician narcotic addicts. The first employs a combination of unstructured and focused approaches while the latter two use standardized interview techniques.

Like information derived from participant observation and case history approaches, interview data are subject to effects of maturation and reactivity. Rapport is the common term used to describe the degree of validity in responses, necessitating some sort of demonstration that a combination of threats to internal validity have been removed, including the following:

(1) The degree to which respondent "resistance" and "desirability" factors have been overcome. Tendencies to withhold or report misleading information are likely higher in evaluation studies where respondents know that findings may be used to terminate program efforts. Also, there is often a "right," or "most desirable" answer to questions adressed to subjects in evaluative studies, that is, indicating that the program is meeting or exceeding its stated goals and objectives.

(2) The degree to which issues of interview bias have been controlled, allowing for truthful and spontaneous subject response. Particularly in evaluations of direct service programs, including drug treatment and prevention, interviewing takes place in the same social setting in which the respondent receives service. Where the interviewer is a member of the program staff or is viewed as having a vested interest in evaluative outcomes, issues of truthful response are particularly salient.

(3) The degree to which interviewers have managed to penetrate subcultural meanings and language differences. Respondents are likely to be different from researchers in a wide variety of characteristics related to social distance. To the extent that such differences affect their willingness or ability to articulate their views, interview results are less valid. Specifically, threats to internal validity include patterns of acquiescence-defference and complexity of language used as related to class, age, sex, and ethnic factors. Efforts should be made to obtain a match between interviewers and respondents on relevant social characteristics.

ANECDOTAL DATA

The introduction to part III recommends that comprehensive evaluations can be developed by combining methods from both the hypothetico-deductive and the holistic-inductive paradigms. A major distinction often made between these paradigms is the dialectical form of the debate surrounding quantitative and qualitative methods (Reichardt and Cook 1979), and the claim that the former is somehow better because it is objective, while the latter is subjective.

Even when cost or other considerations limit the extent to which formal qualitative methods can be used in an evaluation, information obtained from observations of and comments by program personnel, clients, and others can often shed light on quantitative findings. This information is most frequently used in the form of anecdotes. Ethnographic studies often use anecdotal data to illustrate findings arising from the field setting. For example, Agar (1973) used quotes from two "jailhouse" poems to point out the disparity in self-image among addicts. The first, entitled "King Heroin," uses Standard American English, with no addict argot, for example:

I've captured men's wills, destroyed their minds.
Caused men to commit brutal crimes.
Now I can made a mere schoolboy forget his books.
Make a world famous beauty neglect her looks;
Made a good husband forsake his wife.

The addict is presented in this poem as a social failure, caught by his overwhelming need for heroin.

In the poem "Honky Tonk Bud," a very different and more positive picture of the addict emerges. The addict argot (circa 1960) is used so extensively that, to those not familiar with it, the poem becomes incomprehensible, for example:

He was choked up tight with a white-on-white, had a cocoa front
that was down.
Sported a hand-painted tie that hung down to his fly, and he had
on a gold dust crown.

This anecdotal material enriches Agar's discussion of the two self-images of the addict based on different social category identifications, and his further discussion, based on symbolic interactionism, of differing self-images in response to different significant others.

Anecdotal data need not be used only to highlight ethnographic research. They can be used in conjunction with quantitative analysis as well to:

- develop questions
- reinforce and support conclusions
- enhance and illuminate findings
- clarify contradictions
- add richness or body to findings

The use of anecdotal data cannot be justified unless it is anchored by careful analysis. Properly used, anecdotal data can become a solution to the problem reflected in the old saying, "Too many research studies use statistics in much the same way as a drunk uses a lamp post--for support rather than illumination." In fact, within the field of statistics, there is a growing recognition of this problem, leading to an increasing use of exploratory data analysis, the major proponent of this move being John Tukey (1977), who could reasonably be called the "ethnographer of data analysis."

Let us look at two examples in the field of prevention. It is generally recognized that information regarding the harmful effects of various substances, in the absence of other prevention activities, is not sufficient to change the attitudes or behavior of current users. In this light, consider the following anecdote, as told by an ethnographer:

> I was waiting to see the principal of a grammer school in a typical New York ghetto neighborhood, watching the clerks and secretaries going about their work, when a ten year old boy came in and started to look in the ashtrays for usable cigarette butts. The workers paid him no mind, but a passing teacher, seeing what he was doing, stopped to give him a kindly lecture on the harm of cigarette smoking. The boy seemed to be attentive, so the teacher expounded at length, finally ending with "Now, do you have any questions?"
>
> "Yeah," said the boy, "You got a match?" (Preble 1980)

This anecdote provides a concrete example of the inadvisability of providing information only, and expecting it to modify behavior.

Another example involves the presentation of information regarding the outcomes of a particular prevention program to a legislative body which had the responsibility to determine the allocation of funds to various prevention activities. Quantitative findings were presented which showed that program participants improved somewhat on measures of self-esteem and school grades, while dropouts, disciplinary actions, and reported drug use declined.

As part of the presentation of findings to the legislative body, several program participants testified, relating their own personal experiences in the program. It was the perception of those prevention and evaluation personnel involved in the presentation that the testimony from the involved youths, which supported and enhanced the quantitative findings, was the major factor in the legislative determination to not only support, but indeed to expand the program (Kaufman 1980).

As these examples show, anecdotes can be considered data in the same way as statistics, in that they become a foundation for conveying information to the audience. It can be argued that anecdotes are essentially rhetorical devices, used to persuade the reader to a particular belief. However, we often forget that statistical findings are used in much the same way in evaluation reports--to support an argument or position with the intent of persuading the reader to accept and adopt the stance of the writer regarding the issue at hand.

Tukey (1977) argues that the data analyst approaches data in essentially three ways: (1) as a detective, to explore the field and find within it relevant information, (2) as a lawyer, who develops arguments based on findings, and (3) as a judge, who makes decisions based on the information at hand. Anecdotal information has a place in all three categories. At the very least, the properly chosen anecdote can be viewed as a reexpression of quantitative findings.

METHODOLOGICAL ISSUES

It is, perhaps, misleading to treat observation, case history, and interviewing as separate and distinct methods of information gathering. Qualitative strategies are almost always used in combination, specific techniques either emphasized or deemphasized depending upon the stage of inquiry, differing groups, contexts, and/or topics being analyzed (Foster 1974). Each of the drug studies cited, in fact, relies on a combination of these methods, with consistent understanding of relevant attitudes and behavior. This holistic orientation, in turn, is the hallmark of ethnographic study (Weppner 1977) and is generally seen as the single most important contribution that the application of qualitative methods to evaluation research can make (Landy 1977; Logan and Hunt 1978).

Combining methods is a basic way of dealing with problems of internal invalidity in qualitative data, offering a defense against criticisms regarding the "subjective" or "impressionistic" nature of such information. Generalizations based on a combination of data sources, tapped by different methods, can be presented in terms of attaining a higher degree of validity than findings based on reliance on any single method. Defenders of qualitative strategies, in fact, may argue that their information has more internal validity checks than information generated by survey or experimental techniques in that it is developed through ongoing inductive testing against social-psychological "realities" (Glaser 1965). The overall idea is that one plays off each data source and method against all others in order to maximize the scope and sensitivity of information and hypotheses developed in field efforts.

Without careful attention to threats to external validity common to all qualitative methods, however, combinations of strategies will simply result in a multiplication of error. If, as suggested by ethnographers (Basham and Degroot 1977), qualitative studies of drug-related behavior may eventually provide a solid comparative basis for a priori examination of possible program alternatives, issues of external validity assume an immediate and crucial importance. Shared sources of invalidity arise from the very strength of qualitative methods--their indepth focus on fewer cases in hopes of attaining extremely detailed information. Let us look briefly at several of these sources of error.

Sampling Considerations

For purposes of establishing external validity the researcher must demonstrate that cases studied and reported are representative of either the general "target" population or some specific subgroup to which s/he wishes to generalize. This includes some sort of demonstration that temporal and situational factors have been adequately considered and treated.

Theoretical sampling provides, perhaps, the best criterion for establishing external validity. In formulating generalizations the researcher may rely on "crucial" cases, time, and situational contexts, whereby supportive and contradictory examples are chosen for intensive analysis in an explicitly structured effort to define limitations of information gathered. Diversity is the key consideration, however, and studies may be critically reviewed in terms of the range of efforts made to minimize effects associated with group selection, temporal, and situational circumstances upon which generalizations are based.

Where generalization was not of immediate interest in the study reported, that is, in cases where research was limited to goals of developing ongoing monitoring systems or fine tuning survey instruments rather than formulating general descriptions, "oddball" sampling may be legitimately employed. This emphasizes groups or contexts of particular interest and should be so reported. While results may complement findings based on theoretical sampling procedures, reviewers must remain aware of substantive limitations involved in findings and, in turn, of possible differences in attempting to weigh suggested program alternatives derived from such studies.

Subject Mortality

Since qualitative strategies are based on small samples to begin with, any loss of subjects is likely to prove particularly troublesome. Such losses are especially likely to occur in prevention programs where subjects who have received either the most or the least help may selectively "drop out." There is, of course, no simple way of correcting for "selection-residualization" effects, but qualitative strategies may offer insight into the kinds of subjects who stay or leave, as well as identifying such processes when they first begin. While always remaining a major threat to external validity and substantive generalizations based on such studies, insights provided by qualitative strategies should prove important in assessing the seriousness of mortality effects.

Comparing Studies

The high time investment required for generating qualitative information is a problem common to observation, case history, and interviewing techniques and is perhaps a major reason why such methods are so infrequently employed in evaluation studies. This in turn suggests that studies using such methods may be quite different from those where no such efforts were employed, either in their overall purpose, their structure and organization, or some combination of all three aspects. Generalizations provided and comparative implications drawn across studies are invalid where such differences are left unexamined. Again, however, there is no simple solution for taking such differences into account unless, of course, study summaries explicitly outline issues of purpose, structure, and organization. Where time is an issue in eliminating the use of qualitative methods in a study perhaps the only alternative is to specify theoretical samples in advance and outline quasi-experimental interventions that may increase the overall time-to-information gained ratio. A priori specifications, of course, require that problems for study be fairly well defined in advance of initiating evaluative efforts, hopefully a relatively common case where such research is conducted under stringent time limitations.

Where explicit attention is given to issues of internal and external validity, then, qualitative methods can provide a unique and useful set of information that both supplements and compensates for data derived from the application of quantitative procedures. Rather than restricting study to a deductive process of theory formulation and rejection, so-called "soft" strategies provide an ongoing source of information which, in turn, enables progressive formulation and refinement of measures, concepts, models, and theories comprehensively documenting the range of program impact and processes through which efforts either meet or fall short of attaining stated goals and objectives. In a general sense the inclusion of such strategies offers our best guarantee of forcing a clear and careful consideration of all the complexities involved in program organization, delivery, and effect.

REFERENCES

*Asterisks indicate books or articles on presentation and discussion of true and quasi-experimental design.

Abt, C. C. The public good, the private good, and the evaluation of social programs: How inept government requirements increase costs and reduce effectiveness. Evaluation Quarterly, 1978, 2(4), 620-630.

Agar, M. Ethnography and the addict. In A. Weiss (Ed.), Readings in anthropology 75/76. Guilford, Connecticut: Dushkin Publishing Group, 1974.

_____ Ripping and running: A formal ethnography of urban heroin addicts. New York: Seminar Press, 1973.

_____ Selecting a dealer. American Ethnologist, 1975, 2, 47-60.

*Baltes, P. B., Reese, H. W., and Nesselroade, J. R. Life-span developmental psychology: Introduction to research methods. Monterey, California: Brooks/Cole Publishing Company, 1977.

Basham, R., and Degroot, D. Current approaches to the anthropology of urban and complex societies. American Anthropologist, 1977, 79, 414-420.

Becker, H. S. Becoming a marijuana user. In H. S. Becker, Outsiders: Studies in the sociology of deviance. Glencoe, Illinois: Free Press, 1964a.

_____ Marijuana use and social control. In H. S. Becker, Outsiders: Studies in the sociology of deviance. Glencoe, Illinois: Free Press, 1964b.

Bernstein, I. N. Validity issues in evaluative research: An overview. In I. N. Bernstein (Ed.), Validity issues in evaluative research. Beverly Hills, California: Sage Publications, 1976.

Bernstein, I. N., Bohrnstedt, G. W., and Borgatta, E. F. External validity and evaluation research: A codification of problems. In I. N. Bernstein (Ed.), Validity issues in evaluative research. (Sage Contemporary Social Science Issues No. 23). Beverly Hills, California: Sage Publications, Inc., 1976.

Bien, N. Z., and Bry, B. H. An experimentally-designed comparison of four intensities of school-based prevention programs for adolescents with adjustment problems. Journal of Community Psychology, in press.

*Blalock, H. M. Jr., and Blalock, A. B. Methodology in social research. New York: McGraw Hill Book Co., 1968.

Bleich, H. L. The computer as consultant. New England Journal of Medicine, 1971, 248, 141-147.

*Boruch, R. F., and Rindskopf, D. On randomized experiments, approximation to experiments and data analysis. In L. Rutman (Ed.), Evaluation research methods: A basic guide. Beverly Hills, California: Sage Publications, Inc., 1977.

Boruch, R. F., McSweeny, A. J., and Soderstrom, E. J. Randomized field experiments for program planning, development, and evaluation: An illustrative bibliography. Evaluation Quarterly, 1978, 2(4), 655-695.

Bracht, G. H., and Glass, G. V. The external valididty of experiments. American Education Research Journal, 1968, 5, 437-474.

Bry, B. H. First interviews to determine the long-term effects of the early secondary intervention program (Final Research Report). Freehold, N. J.: Monmouth County Board of Drug Abuse Services, January 1977.

_____ Research design in drug abuse prevention: Review and recommendations. International Journal of Addictions, 1978, 13, 1157-1168.

Bry, B. H., and George, F. E. Evaluating and improving prevention programs: A strategy from drug abuse. Evaluation and Program Planning, 1979, 2, 127-136.

Campbell, D. T. "Degrees of freedom" and the case study. In T. D. Cook and C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills, California: Sage Publications, 1979.

_____ Factors relevant to the validity of experiments in social settings. Psychology Bulletin, 1957, 54, 297-312.

_____ Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D. T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensating education look harmful. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

*Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Company, 1969.

Cavan, S. Liquor license: An ethnography of a bar. Chicago: Albin Press, 1966.

Clark, W. The contemporary tavern. Social Research Group Paper #B102. Berkely: University of California Berkeley, 1979.

Clausen, J. A. Social patterns, personality and adolescent drug use. In A. Leighton, J. Clausen, and R. Wilson (Eds.), Explorations in social psychiatry. New York: Basic Books, 1957.

Cohen, J., and Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. New York: John Wiley and Sons, 1975.

*Cook, T. D., and Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

*Cook, T. D., Cook, F. L., and Mark, M. M. Randomized and quasi-experimental designs in evaluation research: An introduction. In L. Rutman (Ed.), Evaluation research methods: A basic guide. Beverly Hills, California: Sage Publications, 1977.

*Cook, T. D., and Campbell, D. T. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally, 1979.

De Dombal, F. T. Surgical diagnosis assisted by computer. Proceedings of the Royal Society of London, B., 1973, 184, 433-440.

Dixon, W. J. and Massey, F. J., Jr. Introduction to statistical analysis (3rd ed.) New York: McGraw-Hill, 1969.

*Edwards, A. L. Experimental design in psychological research. (4th ed.), New York: Holt, Rinehart and Winston, 1972.

Feldi, K. Behavioral treatment of alienated adolescents. (Doctoral dissertation, Columbia University, in preparation).

Fisher, L. A., Johnson, S., and Porter, D. Collection of a clean voided sample: A comparison among spoken, written, and computer based instruction. American Journal of Public Health, 1977, 67, 640-644.

Fitz-Gibbon, C. T., and Morris, L. L. How to design a program evaluation. No. 3 in the Program Evaluation Kit of the Center for the Study of Evaluation. Beverly Hills, California: Sage Publications, 1978.

Foster, G. Medical anthropology: Some contrasts with medical sociology. Medical Anthropology Newsletter, 1974, 6, 1-6.

Garfinkel, H. Studies in ethnomethodology. Englewood Cliffs, New Jersey: Prentice-Hall, 1967.

Glaser, B. G. The constant comparative method of qualitative analysis. Social Problems, 1965, 12, 436-445.

*Glass, G. V., Willson, V. L., and Gottman, J. M. Design and analysis of time series experiments. Boulder, Colorado: Colorado Associated Universities Press, 1975.

Goffman, E. F. Behavior in public places: Notes on the social organization of gatherings. New York: Free Press of Glencoe, 1963.

Gold, R. L. Roles in sociological field observations. In G. J. McAll and J. L. Simmons (Eds.), Issues in participant observation. Reading Massachusetts: Addison-Wesley, 1969.

Gorry, G. A., and Goodrich, T. J. On the role of values in program evaluation. Evaluation Quarterly, 1978, 2(4), 561-572.

Gurel, L. The human side of evaluating human services programs: Problems and prospects. In M. Guttentag and E. L. Struening (Eds.), Handbook of evaluation research, Vol. II. Beverly Hills, California: Sage Publications, 1975.

Gustafson, D. H., Greist, J. H., and Strauss, F. F. A probabilistic system for detecting suicide attempts. Computers and Biomedical Research, 1977, 10, 83-89.

*Haggard, E. A. Intraclass correlation and the analysis of variance. New York: Dryden Press, 1958.

Hawkins, D. F. Applied research and social theory. Evaluation Quarterly, 1978, 2(1).

Hughes, P. H., Crawford, G. A., Barker, N. W., Schumann, S., and Jaffe, J. H. The social structure of a heroin copping community. American Journal of Psychiatry, 1971, 128, 551-558.

Kaufman, N. J. Personal Communication, May 1980.

Klitzner, M.D., Danziger, S.D., Bosworth, K., and Gustafson, D. Bridging the adolescent health gap. Working paper, University of Wisconsin, 1979.

Lana, R. E. Pretest sensitization. In R. Rosenthal and R. L. Rosnow (Eds.), Artifact in behavioral research. New York: Academic Press, 1969.

Landy, D. (Ed.). Culture, disease and healing: Studies in medical anthropology. New York: MacMillian, 1977.

Levine, A., and Levine, M. The social context of evaluative research: A case study. Evaluation Quarterly, 1977, 1(4), 515-542.

Lindesmith, A. The argot of the underwold drug addict. Journal of Criminal Law and Criminology, 1938, 29, 279-286.

Logan, M. H., and Hunt, E. (Eds.). Health and the human condition: Perspectives on medical anthropology. Cambridge, Massachusetts: Duxbury Press, 1978.

Maultsby, M. C., and Slack, W. V. A computer based psychiatric history system. Archives of General Psychiatry, 1971, 25, 270-275.

McNemar, Q. Psychological Statistics (5th ed.). New York: Wiley, 1975.

Merton, R. K., and Kendall, P. L. The focussed interview. American Journal of Sociology, 1946, 51, 541-557.

National Institute on Drug Abuse. Research on smoking behavior (Research Monograph No. 17, GPO No. 017-024-00694-7). Washington, D.C.: U.S. Government Printing Office, 1977.

*Nunnally, J. C. The study of change in evaluation research: Principles concerning measurement, experimental design, and analysis. In E. Struening, and M. Guttentag (Eds.), Handbook of evaluation research, Volume I. Beverly Hills, California: Sage Publications, 1975.

Plant, M. A., and Reeves, C. E. Participant observation as a method of collecting information about drugtaking: conclusions from two English studies. British Journal of Addiction, 1976, 71, 155-159.

Preble, E. Personal Communication, May 1980.

Preble, E., and Casey, J. J. Taking care of business: The heroin addict's life on the street. International Journal of the Addictions, 1969, 4, 1-24.

Ray, M. B. The cycle of abstinence and relapse among heroin addicts. In H. S. Becker (Ed.), The other side: Perspectives on deviance. New York: New Rochelle, New York: Arlington, Free Press of Glencoe, 1964.

Sechrest, L. (Ed.) Unobstrusive measurement today. San Francisco: Josey-Bass, 1979.

Slack, W. V., Hicks, G. P., and Reed, C. G. A computer based medical history. New England Journal of Medicine, 1966, 274, 194-198.

Slack, W. V., Porter, D., and Witschi, J. Dietary interviewing by computer. Journal of the American Diet Association, 1976, 69, 514-517.

Street, L. (pseudonym), and Loth, D. I was a drug addict. New York: New Rochelle, Arlington House, 1973.

Swisher, J. D., and Crawford, J. L., Jr. An evaluation of a short-term drug education program. The School Counselor, 1971, 18, 265-272.

Swisher, J. D., Warner, R. W. Jr., and Herr, E. L. Experimental comparison of four approaches to drug abuse prevention among ninth and eleventh grades. Journal of Counseling Psychology, 1972, 19, 328-332.

Twain, D. Developing and implementing a research strategy. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

Visco, E. P., and Finotti, J. F. Spark program analysis (Report No. HF-347). Pamona, California: Geomet, September 1974.

Webb, E. J., Campbell, D., Schwartz, R., and Sechrest, L. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Weinstein, A. S. Evaluation through medical records and related information systems. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

Weiss, C. H. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

_____ Evaluation research in the political context. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

_____ Interviewing in evaluation research. In E. L. Struening and M. Guttentag (Eds.), Handbook of Evaluation Research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

*Welkowitz, J., Ewen, R. B., and Cohen, J. Introductory statistics for the behavioral sciences. New York: Academic Press, 1976.

Weppner, R. S. Street ethnography. In R. S. Weppner (Ed.), Annual review of drug and alcohol abuse, 1977, 1.

Wiggins, J. A. Hypothesis, validity, and experimental laboratory methods. In H. M. Blalock, Jr. and A. B. Blalock (Eds.), Methodology in social research. New York: McGraw Hill, 1968.

*Winer, B. J. Statistical principles in experimental design. (2nd ed.), New York: McGraw Hill, 1971.

Winick, C. Physician narcotic addicts. In H. S. Becker (Ed.), The other side: Perspectives on deviance. Glencoe, Illinois: Free Press, 1964.

CHAPTER 8: METHODS FOR THE STUDY OF IMPACT

Evaluation research is usually concerned with the effects of a program on a particular population such as high school students, clients of a prevention program, or members of a trade union. However, it seems reasonable to assume that certain programs will have, in addition, considerable influence on associates of the target population. For example, parents of the high school students, family members of clients, and spouses of union members may be affected by changes in members of the respective target populations. In this chapter, the generalized effect, or the effect on and beyond the program population will be referred to as impact. The purpose of this chapter is to present several types of epidemiological and social science methodologies for estimating the impact of a program on a community. It is understood that the evaluation research designs previously presented in chapter 7 and the statistical principles formulated in chapter 9 can be applied to estimate the impact of programs on subjects beyond program participants. The essential feature of impact evaluation is the establishment of a causal chain between the prevention programing and the improved quality of health within the community.

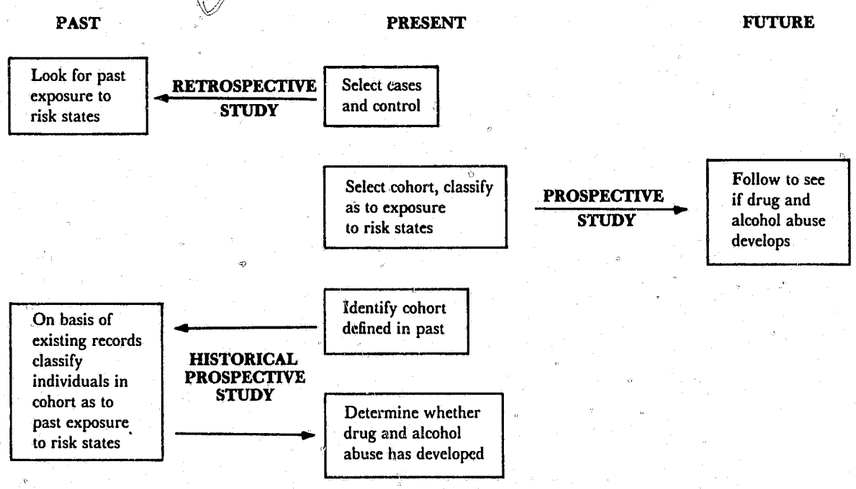
As with outcome evaluation, the strongest evidence for impact is generated from experimental studies in which treatment and control groups are established, preferably through random procedures. However, study constraints may limit the application of experimental or quasi-experimental designs. To augment these paradigms, several observational methodologies from the science of epidemiology are appropriate for impact evaluation. These include the retrospective study, the prospective study, and the historical prospective study. Figure 7 summarizes the three study designs.

Under the retrospective study design, individuals presently considered as drug or alcohol abusers are compared with groups of individuals without the problem to determine their differences in past exposure or early indication of risk states associated with alcohol or drug abuse. Prospective studies identify groups of individuals who do not evidence drug or alcohol abuse as a problem to determine if there are differences in the rate of future occurrences of the problem. Finally the historical prospective study assesses the emergence of drug/alcohol problems within a group of individuals previously identified as at risk to drug or alcohol abuse.

In addition to these epidemiologic methodologies, this chapter will present three other techniques relevant to impact evaluation: two stage panel studies, social area analysis and cost benefit/cost effectiveness analysis.

The two stage panel study described later in this chapter provides a number of possibilities for studying program impact. The sampling of households with parents and high school age children as subjects would provide baseline values. With students randomly assigned to two kinds of drug or alcohol education programs, the parents may serve as a potential impact population. The first-stage interview can also be used to assess the drug or alcohol education needs of both parents and children by estimating current levels of knowledge, attitude, and use. Information derived from the needs assessment may be used to advantage in determining the content of the two drug or alcohol education programs. Stage two interviews would assess change in knowledge and attitude in both parents and children. Measures of relationship between parents and children can be used to predict acquisition of knowledge and change in attitude by parents.

Figure 7. Epidemiologic Models for Impact Evaluation



Social area analysis, discussed in the section following panel studies, exploits the use of information in the public domain, such as census data, vital statistics, and social service data, to describe the socioenvironmental context of populations. Characteristics of the social context of populations vary greatly over, for example, census tracts of large cities, and provide a powerful basis for predicting variation in the use of services. Patterns of change in dense populations create stressful conditions and, under certain circumstances, rapid increases in rates of drug use. The influence of the environment in shaping attitudes and behavior is gaining increased attention as we learn more about the limits of predicting important behaviors from a knowledge of individual characteristics. Social characteristics serve as powerful mediating variables and influence the impact of programs beyond the target population. Variation in small areas also has implications for program design in that populations of different cultural, ethnic, and social class backgrounds have different needs.

Finally, cost benefit and cost effectiveness analysis have relevance for impact evaluation. The procedures are based on the theory that one cannot make a judgement about either costs or benefits of any program without relating them to each other.

Cost benefit/cost effectiveness analyses can provide valuable information to prevention program administrators and policy makers in a variety of ways, such as:

(1) To account for use of public funds. Local, State, and Federal governments are primary sources of funds for prevention activities. Government officials and the public require information on how funds are expended for prevention.

(2) To compare the cost of alternate prevention services. Cost analysis can be used to compare alternate methods for providing prevention services. With cost data on alternate services, the analysis seeks to identify the least costly program alternative that can accomplish the desired objective.

(3) To identify the efficiency of the operation. Cost analysis can shed light on the questions of the efficient use of resources and the optimum size of an operation, given community needs.

(4) To allocate or reallocate resources. The information of costs and benefits of a program can help prevention program administrators to modify or improve the process of the program and reallocate resources among alternative programs.

We turn now to a description of these methods and their applications.

EPIDEMIOLOGIC STUDIES

RETROSPECTIVE (CASE CONTROL) STUDY

When it is desired to explore causal factors in the occurrence of a social problem, a retrospective or case control study often forms the initial approach. Case control studies are relatively inexpensive and quick to carry out, and they permit one to explore multiple potential antecedents. In addition, they are specifically suitable for investigating etiologic factors in complex or unusual problem areas.

Several points about retrospective studies should be mentioned.

First, cases should consist of a random sample of newly identified drug and alcohol abusers during a specified time period. In this way, even though the study design is retrospective, the sample represents incident rather than prevalent cases and avoids bias toward overrepresentation of cases of long duration that results from the use of prevalence data. Second, the control group should be selected to closely resemble the cases, except, of course, for the presence of drug or alcohol abuse. To achieve an appropriate control group it is necessary to duplicate relevant selection factors that influence the composition of the case group. (For example, select controls from the same hospital or neighborhood as the cases.) And third, in virtually every study there are factors or confounding variables that might distort the relationship between drug and alcohol abuse and their risk states. A

confounding variable may be defined as a parameter of the system that causes change in the dependent variable and that varies systematically with the hypothesized causal variable under study. Confounding may be handled through matching or statistical analysis. Age, sex, race, ethnic group, and sociometric status are frequently used as matching variables to create closely balanced case and control groups. In some circumstances, confounding variables are not neutralized by matching but are allowed to vary. The resulting confounding is then eliminated through statistical analysis, as in age adjustment or covariance analysis. Whichever the approach, it is necessary to consider the existence of confounding factors and determine their presence or absence in the members of the study group.

PROSPECTIVE (COHORT) STUDY

In contrast to the retrospective approach, prospective or cohort studies start with a group of individuals (a cohort) who: (1) are free of drug or alcohol abuse and (2) vary in exposure to a given set of risk states believed to be associated with this problem. The cohort is followed over time to determine if there are differences in the rate at which a drug or alcohol problem develops in relation to the intensity, duration, or magnitude of the risk states.

Cohort studies provide a much stronger study design than the retrospective study. They are, of course, much more time consuming and expensive to carry out because the number of people who must be enrolled is inevitably much greater, given that the problem is expected to occur in only a small proportion of the sample. They do however, have several important advantages. It is generally agreed that cohort studies are less subject to bias than are retrospective studies. In part, this is because individuals in the cohort are classified as to exposure before the drug or alcohol problem develops, whereas in case control studies the level of risk is ascertained after the development of drug or alcohol abuse. Further, in most prospective studies there is no need to select controls because the study group includes both at risk and not at risk individuals. Even when the whole group has some degree of risk, it may be possible to stratify the cohort by degree of risk and compare rates of drug and alcohol abuse for the subgroups. Sometimes this may not be possible and it will then be necessary to designate a comparison group or to select available population statistics for comparison.

Additional advantages of cohort studies are (1) that they provide validation of the criteria used for determining risk, (2) that they permit calculation of the incidence rates of drug and alcohol abuse among at risk persons, and (3) that they permit the detection of multiple outcomes related to patterns of risk states. For example, prospective studies to test the association between cigarettes and lung cancer uncovered additional effects of cigarette smoking, such as increased risk of death from other types of cancer and increased death rates from heart disease.

Of course, cohort studies pose problems in addition to the time and expense involved in carrying them out. During the period of time spanned by the study, a certain number of subjects is usually lost to followup and, even more serious, losses may be differentially related to risk status. Further, improved methods of identification of drug/alcohol abuse and diagnostic methods may complicate the interpretation of results. Despite these limitations, prospective studies can yield important impact information.

HISTORICAL PROSPECTIVE STUDY

Prospective studies are not restricted to data collection over some forward span of time. On occasion it is possible to do a cohort study that takes advantage of the fact that relevant information was recorded for a defined group at some time in the past. This variation of the cohort study has been referred to as the historical prospective study, or nonconcurrent cohort study, to indicate that the observations on the study group were made prior to the initiation of the study. An example of a nonconcurrent cohort study is one (Court-Brown and Doll 1965) in which the investigators traced the deaths that had occurred between 1934 and 1954 among some 13,000 patients treated with radiation for rheumatoid arthritis of the spine. By inspection of current medical records, the researchers found that this group had

a substantially higher death rate from aplastic anemia and leukemia than that of general population. In this way, a historical prospective study can utilize both past and current data sources to assess causal relationships between risk status and the appearance of drug or alcohol abuse.

PANEL STUDIES

The design strategy and the analytical techniques of panel studies were initially developed in the late 1930's by Lazarsfeld (1948) in order to study change in the attitudes and behavior of individuals. In a panel study the same individuals are interviewed at different points in time. This type of longitudinal approach differs from more often conducted trend studies in which different samples of individuals are studied cross-sectionally at various points in time. A trend study makes it possible to ascertain what proportion of individuals are engaged in certain types of behavior, for example, the use of various mood altering substances at different points in time. Because in a panel study the same individuals are reinterviewed over time, it is possible to specify both the number and the unique characteristics of individuals that have changed. Thus, in impact evaluation of prevention programs, a panel study makes it possible to additionally determine what types of individuals are most likely to change in their use of such substances and under what situations these changes are most likely to occur. This methodology provides a specific framework and process to provide data concerning the following three characteristics:

- Direction and mode of changes in individuals. For example, how important is peer group influence in the use of marijuana; or, what impact does marijuana decriminalization legislation have on the use of marijuana and other drugs.

- Determination of relationships between earlier life styles and later behaviors. Here, for example, one may be interested in the relationship between child abuse or lack of significant others in childhood to experimentation with and use of opiates or, what socioeconomic and personality variables might explain the extent of marijuana use after decriminalization legislation.

- Assessment of differential changes in groups which have been exposed to different intervention strategies or policy decisions. For example, the effectiveness of individual, group, and community level approaches in the prevention of opiate users may be compared; or, the impact of marijuana decriminalization on use of marijuana among youth in one State can be compared to similar youths in a nondecriminalized State.

Though systematic longitudinal panel studies are infrequent in drug use investigations, the few that do exist point out the significance and potential of this strategy. A prime example is the nationwide study by Johnston (1973, 1974) of some 2,200 tenth-grade male high school students who were initially studied in 1966 and subsequently reinterviewed a number of times. This sample was randomly stratified to represent all young men beginning high school in the continental U.S. in the fall of 1966. By 1970 a third reinterview had been conducted of 73 percent of the original sample. This panel type study, due to its prospective dimension, makes it possible to determine directly rates of drug use and to explicate risk or causal factors associated with the development and changes in drug use. Indeed, a number of findings from this panel study question cross-sectional and other types of data collected in studies that simplistically attribute low grades, crime, juvenile delinquency, and social isolation to drug use in young persons.

A basic requirement in organizing and implementing a panel study is maintaining contact with and relocating respondents for subsequent reinterviewing. Methodological problems related to this requirement are potentially considerable, and it can involve a major effort and expense to stay in touch with an adequate proportion of the study group. Subject attrition can be related to a number of social factors such as socioeconomic status, age, and mobility. For example, persons who move frequently due to work opportunities are more difficult to maintain in the followup phases of a panel study. Also, in drug use studies, patterns of

drug use in themselves may be related to subsequent study participation or nonparticipation. For example, drug users may avoid or refuse interviews because of involvement in illegal activities or the perceived stigma associated with drug use.

These difficulties in maintaining and relocating followup respondents for a panel study are not necessarily insurmountable. In a recent report (Clarridge, Sheehy, and Hauser 1978) on a 17-year followup of over 10,000 Wisconsin high school graduates, tracing procedures located 97 percent of this group and 88 percent were ultimately interviewed by telephone. In addition, the followup group was found fully representative on key variables with the original panel group. The authors attributed their followup success to extensive telephone use, certain characteristics of the tracing operations, amounts of information available for tracing (such as a prior postcard survey), and having had enough resources to do it. The last factor is of key significance as noted by the authors:

> A major factor in our success--the allocation of sufficient resources for the tracing operation--reflects the belief of the project directors in the importance of locating virtually every member of the sample. Too frequently, a good deal of time and money are spent analyzing data from incomplete and biased samples (op. cit., p. 202).

Similar success has been reported in a Baltimore panel study of some 5,000 children who had been studied a number of times over twelve years, with 88 percent of the original panel reinterviewed and fully representative on key variables (Hardy 1971; Richardson, Hardy, and Dallas 1975). This followup success was attributed to the continuous maintenance of a tracing operation that contacted the panel group every six months, by utilizing a number of direct and indirect strategies including telephone calls, return postage-guaranteed postcards, and even household visits to obtain relocation information from relatives, friends, and neighbors.

Another difficulty confronting panel studies is related to testing effects; that is, repeated surveys may sensitize an individual and lead to responses indicative of reactions to the interview situation and not of change in attitude or behavior. For example, in studying patterns of drug use over time, repeated measurement may lead individuals to become increasingly aware of and influenced by acceptable social norms of drug behavior. This may lead to changes of interview responses in the socially acceptable direction without parallel changes in actual drug attitudes or behavior. This is more likely when the time interval between panel interviews is short, say, several weeks, as compared to a year. Attempts to partially control for such testing effects include (1) the use, whenever feasible, of questions and tests that minimize normative connotations, and (2) the use of multiple items measuring the same phenomenon in order to cross-check for an individual's consistency in responses. Estimates of testing effects can be obtained by adding to a panel study matched control groups which are measured only once. For example, in a two-wave panel study, one control group would be used for the first wave and a second, but independent, control group would be used for the followup wave. Since the two independent control groups would not be sensitized to measurements over time, comparison of these groups allows one to estimate the presence and magnitude of testing effects for the panel's experimental and control groups.

Methodologists in recent years have presented a number of new and complex types of study designs and data analysis models. Though beyond the scope of this chapter, some of these strategies are noted and references provided for the reader wishing further information. Labouvie (1976) advocates two extended cohort-type[1] longitudinal research designs beyond the parameters of more simple panel type experimental and control group designs. One is a "cohort-sequential" design in which a set of cohorts is studied at different age levels, thus providing a longitudinal series for each of several generations. The other is a "cross-sequential" design in which a fixed set of cohorts is studied at different times. An asset of these cohort panel studies is their sensitivity in describing age related types of change.

Various data analysis strategies for panel studies have been cogently described by Duncan (1969) and Goodman (1973), and the interested reader is urged to see these sources for specific details. One analytic schema that has been used to ascertain causal priorities in panel study data utilizes cross-lagged correlations based upon the strategy that repeated measurement of the same two variables can provide information about the direction of any

causal relationships between them (Kenny 1973, Pelz and Andrews 1964). Figure 8 illustrates the framework for this analytic strategy wherein A and B represent two variables that have been measured at two points in time, 1 and 2, in a panel study. The analytical question is whether A is a stronger cause of B than B is of A. To test this relationship the two cross-lagged correlations of A1-B2 and B1-A2 are compared within the framework of correlations (synchronous) A1-B1 and A2-B2 and correlations (test-retest) A1-A2 and B1-B2. If A is the stronger cause, correlation A1-B2 would be greater than B1-A2.

Another recent analytical venture has been the work of Heise (1970). Heise presents a model for obtaining causal inferences from panel type data through the use of path analysis. Path analysis is a technique whereby an established or hypothesized model of behavior is translated into a series of equations describing the effect over time of one group of variables on another group of variables. Though the viability of this technique depends upon a theoretically sound model, path analysis within a given model framework does provide a formalized structure for analysis of the longitudinal data of panel studies.

In general, panel studies have several distinct advantages that make them valuable as an impact methodology. Though panel studies usually require more effort, time, and expenditure than other study designs, they have the critical advantage of providing the opportunity to study change more precisely on the individual, contextual, and societal levels, since the same individuals are studied over time. In these ways panel studies can contribute to a more indepth and comprehensive understanding regarding drug or alcohol use, in terms of the direction and mode of individual change, relationsips between antecedent social and psychological/risk states and later behaviors, and the differential change in groups exposed to different intervention strategies or policy decisions.

SURVEY METHODS

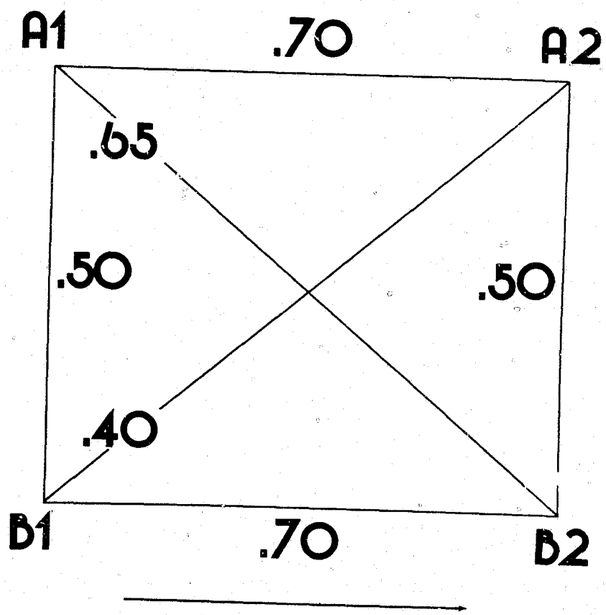
The use of the survey as an impact evaluation technique provides the opportunity for an examination of attitudes, beliefs, and practices of individuals or organizations which may not otherwise be derived through experimentally controlled settings. In evaluation research, the decision to use the survey is based on the assumption that this approach will provide information on the extent to which programs have accomplished their objectives, and second, on the availability of resources, such as financial, personnel, and time. Once the decision has been made, however, the evaluator must adress a number of issues before carrying out the survey. Each of these issues is discussed below.

CHOOSING THE POPULATION

Impact evaluation seeks to measure the secondary effect of the program on the community--defined as a geographical or population area. The first step in identifying the appropriate population to be surveyed is directly associated with programatic aims. For example, if a program is designed to serve senior high school students in a small town in the midwest, it would not be appropriate to survey all households in that state to measure the impact of the program. However, the town in which the program is located or the local school system may be appropriate populations to survey. Thus, the evaluator should be cognizant of programatic aims in determining which is the most meaningful community group to study. Second, the evaluator should consider the size of the target community in order to determine whether or not the entire community or some subsample(s) of it should be surveyed. As mentioned in an earlier chapter, program impacts may affect the three components of the societal social system: societal structure (individual, family, neighborhood, and community); institutions (school, hospital, police, courts, and other governmental institutions); and economy (physical or monetary resources). Within this framework, different sampling designs may be used to select appropriate respresentatives of each of these components. (A discussion of sampling techniques is included in chapter 9.)

For example, surveys of drug and alcohol abuse agencies and practitioners may provide an important source of information relevant to evaluation needs. Such a survey would, at the very least, offer both an understanding of the number and types of requests for drug

Figure 8
Hypothetical Correlations in a Cross-Lag and Synchronous Common Factors Analysis



or alcohol prevention services and the capacity of existing resources to respond to these requests.

SURVEYING APPROACHES

There are three major approaches to surveying. These are the mail questionnaire, the telephone interview, and the inperson or face to face interview. While either of the three, or in some cases, all of them may be used in any one study design, each approach has associated with it a number of advantages and disadvantages. Once the evaluator has determined what the objectives of the evaluation will be, the size of the community, and the meaningful group to be surveyed, s/he can decide which survey approach will be most feasible in terms of cost and the quality of obtained information. (For a general discussion, see Babbie 1973, Sudman 1976, and Warwick and Lininger 1975).

Mail Questionnaires

The mail questionnaire is most often used for very large samples where respondents are distributed over a large geographic area. The U.S. Bureau of the Census, for example, utilizes mail questionnaires for its decennial census, since the cost of surveying such a large, geographically dispersed population (that is, all households in the U.S.) using any other technique would be prohibitive. Thus, the mail questionnaire has a major cost advantage over other survey approaches in that it reduces the size of the research staff needed to carry it out.

There are several major disadvantages, however, in using mail questionnaires. First, the overall response rates may be lower than desired and as a result, may not provide a large enough sample size. Second, results may be biased due to differential response rates among various subgroups in the population. In particular, it could be expected in attempting to obtain direct information on drug or alcohol use patterns, that current users might not have as high a response rate as nonusers. In fact, heavy users might be underrepresented in the survey because they tend to be more transient, hence more difficult to reach by mail or, for that matter, all other survey techniques. Third, in any study design which uses pre and post surveys--and one would expect to use this type of design in doing impact evaluation--the results of mail questionnaires are more difficult to analyze due to biases that may arise from having different response rates for various subgroups of the population at the pre and posttesting times.

Telephone Interviews

The telephone interview is also often used with large, geographically dispersed samples. In addition, this survey approach can serve as a screening tool for identifying appropriate persons to be reinterviewed at a later date. Unlike mail questionnaires, interviews conducted by telephone lend themselves to reinterviewing since all that is required is for the same number to be called. Another advantage of telephone interviewing is that an accurate record of the progress of the research can be provided daily. Specifically, since each respondent contact can be classified as being a "completed interview," a "refusal," or a "call back," the evaluation staff can more easily estimate numbers of additional interviews needed to reach a desired sample size. Because telephone interviews are often associated with commercial marketing techniques, (for example, promotional campaigns aimed at the consumer) the rates of refusals may be somewhat higher than anticipated. Like mail questionnaires, followup calls can be made to reduce the level of nonresponse. Another disadvantage of telephone interviewing is that it often requires that certain physical and organizational structures be available to the surveyors. For example, telephone interviewers usually require a self contained room and flexible hours to maximize contacts with designated segments of the population.

Face-to-Face Interviews

Inperson or face to face interviews are the most expensive and often most valuable survey techniques (Weiss 1975). They can be used to solicit individual or organizational responses, the latter involving a representative of the organization. A major advantage of this survey technique is that a higher response rate is regularly obtained and, like the telephone interview, a daily tally of progress can be made. It also provides the opportunity to establish rapport with the respondent which may lead to more accurate responses and pave the way for a followup interview. The costs of carrying out the inperson interview can be a major disadvantage. This is especially true in cases where it is necessary to travel long distances (for example, across the State) to interview respondents. Another disadvantage is the tendency of some interviewers to "lead" respondents or interpret questions for respondents, contributing to biases in the results.

In summary, the above survey approaches are influenced heavily by the availability of funds and time. Decisions on which approach maximizes the amount of information within the resources available should be made early in the evaluation process. Also, time needed for activities such as interviewer training, testing, and supervision should also be considered.

QUESTIONNAIRE DESIGN

Once the evaluator has determined the survey approach to be used for the study, the construction of items for the questionnaire can begin. This process requires a great deal of forethought and preparation to ensure that instruments capture responses that will be meaningful to the study. In addition to these points it should be noted that length and type of interviews (unstructured, focused, or structured) are constrained by the specific survey approach to be used. For example, it may be preferable to keep mail or telephone questionnaires in a structured format to decrease length of time respondents need to complete the form and to encourage high response rates.

In addition to the above points, several others should be adhered to in developing survey instruments. They include:

(1) Knowledge of the characteristics of the population to be interviewed helps to reduce any cultural differences that may exist between the interviewer and respondents.

(2) Placing of interview items in an order which facilitates the flow or logical sequencing of questions. The objective is to maintain respondent interest and have low rates of refusals.

(3) Avoidance of leading responses and ordering items in a way that the answer to one might trigger and bias a response to the next question.

(4) Deciding whether the answer to any question could be more readily obtained from another source, such as census materials or other evaluation studies.

(5) Making sure that instructions for completing the questionnaire or interview are clear and standardized for all respondents and/or interviewers.

(6) Considering the match of demographic characteristics between interviewers and respondents when choosing interviewers, in order to maximize unbiased responses.

(7) Making sure that all questionnaires or interviews end with a thank you or a note of appreciation.

PILOT TESTS

It is strongly suggested that the questionnaire and data collection procedures be subjected to a pilot test. No evaluator wishes to spend precious research dollars and time on a study design that fails to meet its objectives. A list of points which highlight the importance of pretests follows (adapted from Babbie 1973):

(1) Either the entire survey instruments or portions thereof may be pilot tested.

(2) Instruments should be pilot tested in the manner intended for the final study; self-administered questionnaires and interview schedules should be pretested in the appropriate manners.

(3) The selection of the subjects for instrument pretests can be profitably kept flexible and varied. Subjects should be reasonably appropriate respondents for the questions under consideration. In more rigorous pilot testing of the research instrument, little concern should be given to strict representatives; rather, the attempt should be made to achieve the broadest range of respondent types--including those who may represent a small minority of the population.

(4) Data collection procedures should be pretested to ensure that steps needed to properly organize the work of carrying out the survey are covered.

(5) The pilot test is the most appropriate time to estimate the degree to which nonresponses are systematic by population characteristics, thus producing biased results.

USES OF SOCIAL AREA RESEARCH AS AN IMPACT METHODOLOGY

Social area research establishes the community framework within which prevention programs operate and against which they can be assessed with regard to their ability to change levels of drug and alcohol abuse.

IDENTIFYING TARGET GROUPS

The beginning of a social area approach requires the gathering of two basic pieces of data; the first is very simple, the other extremely complex. First, what are the geographic boundaries or catchment area responsibilities of the program under study? That is to say, to what population is this service being targeted? And second, what group of individuals are in need of the services that are provided (Kramer 1976). In epidemiological terms, one wants to find those individuals "at risk."

This target finding procedure will vary according to the specific nature and intent of the program, but in all cases, the data must be geographically based. For example, if school drug education is the focus, available census data on the age distributions of children as well as the incomes of families with children in the area are of great utility. Additional data can be obtained on school enrollment as well as on truancy--a possible predictor of antisocial acting out behavior. Estimates by local police on the presence and activities of gangs may be helpful in understanding the peer pressures faced by adolescents.

Arrest data, as well as a police estimate of drug related criminality will be valuable to further assess where the problem is greatest. Census data should be utilized for estimates of the size and location of the most probable age, sex, and ethnic groups to engage in drug related behavior. Numbers of persons entering treatment facilities may also indicate needs. Measures of unemployment and income characteristics are available in the census and are further indicators of social class and lifestyle (Goldsmith 1972; NIMH, 1974, 1975a, 1975b).

By judiciously observing the geographical distributions of these various measures of the population we can make good initial estimates of the location of potential users of a service program.

ARCHIVAL DATA

Archival data refer to data that have been collected by a program, agency, or other organization in the normal course of its activities, and not solely to support the needs of an evaluation. Archival data can include participant program activity and financial records, social indicators, school records, criminal justice files, health records, periodic program or agency reports, census data, geographic data, telephone directories, vital statistics, and the records of prior research or evaluation studies. These data may frequently yield information pertaining to a population over a broad geographical area and covering an extended period of time. They tend to be easily quantifiable, inexpensive to get, and, if in a computer readable form, analyzable at relatively low cost.

Access to archival data is often restricted by governmental regulations. For example, Section 5/3(a) of the General Education Provision Act, Public Law 93-380--the Buckley Amendment--restricts release of school based records. Written consent of the parent is required for release of information to anyone other than school officials or authorized government representatives. The parental consent must specify the records to be released, the reasons for such release, and to whom the records will be released. In addition, copies of the records must be provided to the parents and the student if desired. Data released for evaluation purposes may not include personal identity information such as names or social security numbers.

Even in the case of archival records from prior evaluations, release of information is governed by Federal regulations. Of primary importance are requirements for informed consent, confidentiality of a subject's identity, and control of risk to the participant. These regulations are applicable to the primary collection of data as well as to secondary and subsequent uses of data.

Archival data are not usually collected with specific evaluation objectives in mind. Rather, they most often represent responses to needs for global aggregate data from which policy makers hope to obtain gross indications of existing patterns. Frequently, they are collected as a minor adjunct to the work of nonresearch personnel, who often see such collection as having little or no importance. For example, emergency room nurses are often expected to complete forms on drug abuse incidents, while at the same time responding to their major role of dealing with immediate medical crises. It is no wonder that the reliability and validity of the resulting data suffer. In general, such problems are common to all archival data not collected under extremely well controlled conditions.

The following sections describe ways in which archival data can be used to assess needs, monitor trends, and examine the impact of community based prevention programs. These often ignored sources provide useful information which can be used in evaluation studies.

Catchment Areas

The concept of the catchment area has rekindled interest in the use of small areas data for planning and evaluation. The catchment area has emerged as a cornerstone for Federal funding in mental (P.L. 88-164, 1963) and physical (P.L. 94-63, 1975) health. Catchments are defined as geographical areas where some facility is located or is planned to be responsible for providing or making accessible services to all residents (U.S. DHEW, 1970). As generally defined, such areas include between 75,000 and 200,000 residents and are described in terms of standard census units (tracts, minor civil divisions, counties). State agencies have been required, at least for mental health, to file a definition of such areas along with a plan for long range program development. Thus, catchment areas can be used to provide social and economic descriptions of communities which are part of an evaluation study.

Census Data

The major source of information on a catchment basis has been data provided by the U.S. Bureau of the Census. Apart from fitting required definitions of catchment areas, such information is both readily available and of uniform enough quality to permit small area comparisons (Ferris and Lee 1972; Muhlin and Milcarek 1974; Siegel 1974).

The government has encouraged the use of census data for purposes of needs assessment, developing for example, the Mental Health Demographic Profile System (MHDPS). (This system has been outlined in a series of publications. See Goldsmith and Unger 1970, 1972, 1974; Rosen 1974; Rosen, Lawrence, Goldsmith, Windle, and Shambaugh 1975). Computer tapes for standard catchment area profiles have been provided to all cooperating State mental health authorities across the United States. Data available for each catchment area include measures of age, sex, socioeconomic status, ethnicity, and family and housing composition, along with comparable data for the nation, the county, and state in which the catchment area is located. In addition to aggregating within catchment area boundaries, the MHDPS can also summarize this information by census tract, by county, and (in some states) by civil divisions. Finally, the system at the National Institute of Mental Health (U.S. NIMH 1976; Bachrach 1974; Tabue 1976) is capable of linking census data with other public information on catchment areas (that is, service utilization statistics, public health statistics, vital statistics, arrest records, and so on). Thus, where standardized sources of information on incidence and prevalence of drug or alcohol consumption patterns are available (usually through state health departments), census variables can provide a general expression (or indication) of small area differences on a number of variables which, in turn, can be linked to incidence and prevalence.

Geocoding

Information systems relying solely on census data are extremely limited in terms of evaluating the impact of specific program efforts. For these purposes, additional data may be collected through geocoding procedures (Costa 1972; Spuck 1974; Smith 1979). Geocoding (geographic coding) employs the use of numerical or symbolic codes to tabulate records according to location (for example, street addresses coded to census tract of residence). Such records could include any existing information made available by agencies and practitioners involved in prevention programs in a given community. Manuals linking adresses to catchment areas defined by the census bureau (including zip code, block group, and school district as well as tract, minor civil division, and county) are available (U.S. Bureau of the Census 1970). If information is available in machine readable form, as might be obtained from government agencies or university research departments, the Census Bureau has developed several computerized routines which will interrelate information from separate sources (Glimpse 1978). By using census definitions, a wide range of descriptive information is available for use as denominator data in formulating problem specific rates (Denne 1978).

Where such applications are not feasible, the familiar "pin map" may be used to define incidence levels within catchment boundaries and to study such problems as (1) the location of services, (2) the indentification of target populations; and (3) the distribution of resources within catchment areas. By plotting known characteristics of individual cases and/or agencies by location on a catchment map, it is possible to define the distribution of problems and resources across the area. Grid counts can then be generated by combining known characteristics to obtain the total number of similar cases and/or agencies in a grid cell. Totals might then be placed on separate maps in the form of contour plots to geographically analyze available data (Earickson 1970).

In summary, geocoding enables the evaluator to use a variety of community and prevention program data that are already available. It uses the individual as the unit of measurement but also affords possibilities for small area analysis. Geocoding does not require an inordinate amount of manpower or money, and clearly lends itself to collaborative efforts among agencies providing services within a given catchment area.

EXISTING PROGRAMS

Unless one is dealing with a new planned community, the alcohol and drug problems to be described will have existed for some time in the past, with a history of various attempts to deal with those problems. One must therefore locate and describe existing programs in order to understand what possible impact a new program will have. Knowing the nature of the population is as important as learning what services are offered.

One should not underestimate the complexity of this step. A myriad of prevention services, sometimes not well publicized or clearly classified as prevention oriented, will be available from various civic, church, and educational groups as well as the more traditional Government services. It is the job of the careful social area analyst to discover them and understand their scope. The evaluator can then combine this new knowledge of program availability with measures of need to assess the equitability of service distribution as it currently exists.

SERVICE UTILIZATION

As we are discovering, the provision of services does not guarantee that the services will be utilized. The literature is replete with reasons why persons elect to use or not use available services. Utilization depends upon the accessibility and cost to clients as well as public attitudes toward or knowledge of the service. One must carefully determine what parts of programs are over or under utilized and from what catchment areas the clients come, and compare utilization to the initial needs assessment profiles. This will provide information on how unmet needs are being adressed and may point out the need for better program publicity.

Because of ethnicity, age, or style, populations will often use certain types of services more than others. Although the precise dynamics are not fully understood, one can observe and modify programs to suit the target populations (Warheit, Bell, and Schwab 1977; NIMH 1974, 1975a, 1975b).

CRITERIA OF EFFECT

The task of measuring the community impact of a program is very complex (Hargraves 1977, Lukoff 1974). The question of what qualifies as a success must be made explicit. Is it increased program participation? User satisfaction? Praise by a local school or police department for substantially reducing alcohol or drug problems? Decreases in police, hospital, or school records noting alcohol or drug involvement? As with any social condition, what constitutes a remedy is problematic. All of the above indicate some form of program accomplishment, but success is difficult to define. If, for example, one is to assess the impact of a drug education program, measures of attitudes and knowledge about drugs must be obtained from residents in the community both before and after the introduction of the program.

If these "pre" measures were never obtained, it might be feasible to compare the program to a similar community in which no program was in progress (Struening 1975). A community selected for a demographic and social profile which is as close as possible to the target community can serve as a nonequivalent control against which to measure change. The assumption again is that all differences are due to programatic intervention, but in order to make that assumption, the community matching procedure must be done using the most extensive and up to date information available (NIMH 1975a).

The evaluator may want to measure a reduction in alcohol or drug problems in the population. This may be reflected in a decrease in arrests for alcohol or drug related offenses or a reduction in the number of complaints from schools, after taking into account changed policies or procedures. Impact might also be determined from reductions in drug and alcohol related deaths (Lettieri and Bachenheimer 1974), or by studying perceptions of

those "key informants" who are aware of the "drug problem" in the community. (See chapter 7, Qualitative Strategies.)

Many methods for measuring impact are essentially the same as those enumerated in chapter 7 for measuring outcome. For example, archival data on relevant measures such as alcohol purchases can be analyzed using multiple time series analysis by choosing one community in which an intervention has been introduced, and using one or more other similar communities for comparison.

Social area analysis procedures entail matching communities "under treatment" to others without treatment. Comparisons between the two will indicate if they differ significantly, reflecting the impact of a prevention program (Lukoff 1974). One must remain somewhat objective in these comparisons, again realizing that local histories as well as maturation are always threats to the external validity of a measure. The complexity of these procedures cannot be over stressed. Careful planning and measurements are essential for demonstrating the efficacy of prevention programs.

COST BENEFIT/COST EFFECTIVE ANALYSIS

Cost benefit and cost effectiveness analysis are economic techniques devised to evaluate a public or social program by providing information on the optimal allocation of limited resources among competing needs. They are tools of analysis which assess alternative courses of action and help decision makers maximize the net benefit to society. The essence of such analyses is their ability to evaluate the total value of benefits or effectiveness against the total costs.

Although cost benefit analysis and cost effectiveness analysis have the same objective, they differ in scope and degree. Cost benefit analysis compares the monetary value of benefits of a program with the monetary value of its cost. The benefit to cost comparison enables the evaluator to use monetary value as a common measurement in assessing the relative attractiveness of program alternatives. The earliest application of cost benefit analysis was in the area of water resource projects (Eckstein 1958; Krutilla and Eckstein 1958).

The evaluation of costs and benefits of water resources projects were easily measured in monetary terms. For instance, the costs of engineering design, construction of dams, and maintenance and operating costs were calculated from engineering data and previous cost estimates of similar projects. The benefits of projects such as hydroelectric power and water usage for industrial and agricultural use were estimated by using market prices of these outputs. The value of flood control was also assessed by deriving estimates of the avoidance of property destruction and loss of life. Based on these monetary estimates of costs and benefits, one can compare the net benefits or a benefit to cost ratio of a water resource project. This information has been used to allocate limited investment resources among several alternative water projects.

The crucial assumption for conducting cost benefit analyses is that costs and benefits can be valued in monetary terms. Monetary terms can be obtained either from market prices or equivalent valuation. However, many outcomes of social programs have no market valuation. For instance, what is the market price for prevention programs that help to improve participants' self-esteem and self-concept? How do we obtain a price for improvement of family relationships as a result of prevention programs? While some outcomes of prevention programs may be evaluated in monetary terms, such as the amount of cost savings in treatment due to the reduction of drug or alcohol abuse, some outcomes are difficult to express in market values because their markets do not exist.

In such situations, effectiveness indicators are used for evaluation purposes. These indicators are measured in psychological or physical terms. When program effectiveness indicators are linked to program costs, the approach is considered a cost effectiveness analysis.

Cost effectiveness analysis was originally developed in the evaluation of weapons systems (Hitch and McKean 1960). The purpose of using cost effectiveness analysis for national defense systems is essentially to achieve a particular defense objective with minimum costs. In recent years, cost effectiveness analysis has been applied to many social programs, such as vocational education, health, library services, and drug treatment.

Cost effectiveness analysis in prevention evaluation provides information about the effects of resources in relation to the value of resources used for prevention programs. The effects may be measured in nonmonetary or monetary terms. Thus, cost effectiveness analysis is used in a much broader way than cost benefit analysis. The costs of prevention programs, such as space, equipment, supplies, and personnel can be calculated from accounting records. The effectiveness of the program, such as changes in consumption patterns and property crime can be measured in nonmonetary terms, while possible increases in employment can be measured in wages or earnings.

Cost benefit and cost effectiveness analyses normally include three steps. The first, and most important step, is the identification of costs and benefits (effectiveness) of a given program with the specification of costs being relatively easier than the identification of benefits. The objectives of a program may be used as a guide to identify some of the potential benefits or effectiveness of the prevention activity. Costs can be calculated from the value of resources used for the activities of a prevention program. This procedure may appear to be obvious, but in practice it raises a number of fundamental issues of methodology and economic theory. For instance, should the donation of church space for a local prevention program be considered as a cost or not? If considered as a cost, to whom is it a cost? Or, should the reduction of government welfare payments due to a prevention program be considered as a benefit and if so, to whom? The real issue is how far one should go to enumerate and evaluate program benefits and program costs to a third party.

Costs and benefits (effectiveness) can be classified in two broad categories: private and social. Private costs and benefits (effectiveness) refer to the value of resources incurred to and benefits gained by individual program participants, while social costs and benefits (effectiveness) refer to costs incurred to and benefits gained by not only program participants but also by the public. Therefore, social costs and benefits (effectiveness) include both public and private aspects of the society as a whole.

For instance, private costs include participants' expenses, such as fees, supplies, transportation, and the potential earnings that an individual gives up to participate in the program. Social costs of a program include not only the private costs but also operating costs, capital expenses, and the donations from other social organizations. Private benefits (effectiveness) of prevention programs may include the possible reduction of participants' drug and alcohol problems, and improvement of participants' scholastic achievement and self-concept. The possible social benefits (effectiveness) of a prevention program are the reduction of treatment costs, an improvement in work or school productivity (due to improvement of health condition, work skills, or self-concept), or the reduction of crime.

The second step in cost benefit (effectiveness) analysis is the quantification of these costs and benefits (effectiveness). The costs and benefits are usually reflected as prices, fluctuating with the market forces of supply and demand. Therefore, quantifications are often expressed in monetary terms. For instance, the participants expenses and the foregone earnings of participants can be measured in dollars. The value of resources donated by social organizations can also be measured in dollars.

In the benefit area, the reduction of treatment costs or the improvement of productivity can be measured in dollars. In certain circumstances, however, market forces may fail to reflect all costs and benefits. For example, if the participants' reduction of drug and alcohol abuse is a primary benefit of a prevention program, the most explicit and useful indicator for that program's policy maker may not be the monetary value of the reduction of abuse, but may be the quantitative measures of the reduction of use in terms of change of type, frequency, dose, and consumption level. Similarly, participants' improvement in scholastic achievement can be measured in terms of grade point average, test scores, and the completion of a scholastic program and various psychological scales measure personality charcteristics such as self-concept. Community involvement can be measured in terms of

one's participation in political activities and civil organizations, an absence of criminal records, and so forth. These quantifications of benefits in the absence of market valuation are possible effectiveness measurements of a prevention program. Thus, it is possible to develop indicators of these benefits. The question is whether the indicator is a good approximation or a poor one.

Once the indicators are quantified, the final step is the comparison of benefits or effectiveness and costs of a program. In benefit and cost comparisons, there are three alternative investment criteria used to evaluate a program: the present value of net benefits, the benefit to cost ratio, and the internal rate of return. The basic purpose of benefit and cost comparisons is to choose the most desirable program from among a set of alternatives, by selecting the one with the maximum net present value of benefits. The net present value of benefits is defined as the total discounted benefits at the present value minus the total discounted costs at the present value. The reason for discounting costs is to convert all past or future monetary values into current prices so that there is a yardstick for comparison. One chooses, of course, the program with the largest net present value.

Mathematically, the difference between present value of benefits and costs can be expressed as:

$$\sum_{t=1}^{n} \frac{B_t}{(1+i)^t} - \sum_{t=1}^{n} \frac{C_t}{(1+i)^t} \qquad (1)$$

where B_t denotes benefits in period t, C_t denotes costs in period t, is the sum of periods from period 1 to period n, i is the rate of interest used for discounting, and t is the time period.

The benefit to cost ratio, a commonly used investment criterion, is the ratio between the present values of benefit (the numerator) ad the present value of costs (the denominator). A program should have a benefit to cost ratio greater than or equal to one in order to be worthwhile. The higher the ratio, the greater the payoff.

Mathematically, benefit to cost ratios can be expressed as:

$$\sum_{t=1}^{n} \frac{B_t}{(1+i)^t} \Big/ \sum_{t=1}^{n} \frac{C_t}{(1+i)^t} \qquad (2)$$

If the ratio is equal to one, it implies that the present value of benefit is equal to the present value of cost of a program. If the ratio is greater than one, it implies that the present value of benefit is larger.

The internal rate of return is that which makes the discounted value of costs equal to the discounted value of benefits. The estimated rate of return is used to compare against the interest rate, representing the rate of social or private investment. If the rate of return for the program is higher than the interest rate for social or private investment, then investment in the program is worthwhile. If all the alternative programs have rates of return higher than the interest rate, one should choose the program with the highest rate of return.

Mathematically, internal rate of return can be expressed as:

$$\sum_{t=1}^{n} \frac{C_t}{(1+r)^t} = \sum_{t=1}^{n} \frac{B_t}{(1+r)^t} \qquad (3)$$

where r is the internal rate of return and is the unknown in the equation. It should be noted that r is different from i, as shown in expressions (1) and (2). The interest rates or discount rate, i, is an assumed value which is used for deriving present value. On the other hand, r is a solution which makes the present value of benefits equal to the present value of cost. The derived value of r is compared to the assumed value of i to make investment decisions.

In the real world, the results of each of these criteria may not be consistent with each other. Therefore, the choice of the criterion is crucial and depends upon the specific circumstances of a program. Moreover, in order to apply these criteria, a cost benefit comparison makes assumptions as to the size of the discount rate and time periods to be discounted.

For cost effectiveness, two criteria are relevant: average costs per effectiveness and marginal costs per effectiveness. In cost effectiveness comparisons, costs are measured in monetary terms while effectiveness may be measured in nonmonetary terms (that part of the analysis which is expressed in monetary terms can apply benefit to cost comparison criteria). Therefore, the central issue is whether to measure the average costs per unit or to measure the marginal costs (additional costs) for additional units of effectiveness. For example, assume that the result (effectiveness) of program A is a five percent reduction of marijuana users, and the total cost of the program is $50,000. Cost effectiveness is estimated by dividing the reduction in percent of marijuana users (5 percent) by the cost of the program ($50,000). This ratio shows that for every $10,000 in expenditures program A reduces marijuana use by one percent. Now, assume that program B reduces marijuana users in a comparable school over an equal time period by 3 percent, with a program expenditure of $60,000. The average cost per unit of effectiveness is $20,000 in program B. Based on a simple comparison, program A is more cost effective than program B.

Mathematically, the average cost per effectiveness can be expressed as:

$$\sum_{t=1}^{n} \frac{C_t}{(1+i)^t} \Big/ \sum_{t=1}^{n} E_t \qquad (4)$$

where E_t is the measure of effectiveness in period t (E_t is not measured in monetary terms).

The marginal cost per effectiveness unit is another criterion for cost effectiveness comparison. Unlike the average cost per effectiveness unit comparison, the marginal cost per effectiveness unit comparison is concerned about the expansion or reduction of a given program. For instance, if the additional cost is $2,000 for preventing an additional participant from abusing drugs in a program (C), while the additional cost of another prevention program (D) is $1,500 an expansion of program D will be more cost effective than an expansion of program C.

Mathematically, the marginal cost per effectiveness can be expressed as:

$$\sum_{t=1}^{n} \frac{MC_t}{(1+i)^t} \Big/ \sum_{t=1}^{n} ME_t \qquad (5)$$

where MC_t is the difference in cost between two periods (t and t-1) and ME_t is the difference in effectiveness between the same two periods.

It should be noted that both average costs and marginal costs are often affected by the size of the program. This factor should be taken into account when comparison of two programs are made. Without considering program size, the results of such comparisons may be misleading. For example, a media prevention program may be much less costly than an intervention prevention program, since a nationwide media campaign can reach hundreds of

thousands of students. But because of high fixed costs, a media campaign may not be cost effective for a small, local school with a few thousand students.

SUMMARY

It is apparent that cost benefit and cost effectiveness analysis are important tools for analyzing the impact of prevention program. They are especially useful when one plans to allocate limited amounts of resources among competing programs. However, the applicaton of cost benefit/effectiveness is not straightforward. Several precautions are necessary. First, some indicators are only partial or proxy measures for total costs and benefits. Certain benefits and costs are difficult to quantify or to express in monetary terms. Second, in estimating the benefits and costs of a program based on actual data, economists often make explicit assumptions in order to develop cost benefit/effectiveness comparisons (for example, the choice of discount rate, the choice of time period, and the monetary imputation). Third, the benefit to cost ratio may be misleading if it is not calculated for all components of the program on an individual basis. Policy makers tend to be interested in the additional costs of program expansion and its cost effectiveness. Thus, the ratio for the program as a whole may provide misleading guidance for such decision.

In view of these considerations, evaluators should make their assumptions explicit and provide alternative estimates given different assumptions. Similarly, policy makers should recognize that conclusions can be altered by the change of the assumptions of the analyses.

ENDNOTES

[1]A cohort is a group of people who enter a given social system at the same point in time. For example, an age cohort is a group of people who were born in the same period of historical time and are thus approximately the same chronological age.

REFERENCES

Babbie, E. R. Survey research methods. Belmont, California: Wadsworth, 1973.

Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand-McNally, 1968.

Clarridge, B. R., Sheehy, L. L., and Hauser, T. S. Tracing members of a panel: A 17-year follow-up. In K. F. Schuessler (Ed.), Sociological Methodology. San Francisco: Jossey-Bass Inc., 1978.

Cook, T. D., and Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand-McNally, 1976.

Costa, C. H. Application of geocoding and mapping. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Court Brown, W. M., and Doll, R. Mortality from cancer and other radiotherapy for ankylosing spondyhitis. British Medical Journal, 1965, 2, 1327-1332.

Denne, J. D. A standardized geographic portrayal for patient origin studies, Public Data Use, 1978, 6, 34-43.

Duncan, O. D. Some linear models for two-wave, two-variable panel analysis. Psychological Bulletin, 1969, 77, 177-182.

Earickson, R. The spatial behavior of hospital patients: A behavioral approach to spatial interaction in metropolitan Chicago. (Research Paper No. 124). Chicago, Illinois: University of Chicago, Department of Geography, 1970.

Eckstein, O. Water-Resource development. Cambridge: Harvard University Press, 1968.

Glimpse, W. G. The needs for availability of user software to process and analyze bureau machine-readable products. Public Data Use, 1978, 6, 3-12.

Goldsmith, H. F., and Unger, E. L. Differentiation of urban subareas: A re-examination of social area dimensions. (Laboratory Paper No. 35). Adelphi, Maryland: Mental Health Study Center, National Institute of Mental Health, 1970.

Goldsmith, H. F., and Unger, E. L. Social areas: Identification procedures using 1970 census data (Laboratory Paper No. 37). Washington, D.C.: Mental Health Study Center, National Institute of Mental Health, 1972.

Goldsmith, H. F. and Unger, E. L. Social area analysis: Procedures and illustrative applica tions based upon the mental health demographic profile system. In L. M. Siegel, C. Attkisson, and A. H. Cohn (Eds.), Mental health needs assessment strategies and techniques, Part II. San Francisco, California: University of California, The Program Evaluation Project, Langley Porter Neuropsychiatric Institute, 1974.

Goodman, L. Causal analysis of data from panel studies and other kinds of surveys. American Journal of Sociology, 1973, 78(5), 1135-1191.

Groves, W. E. Patterns of college student drug use and lifestyles. In E. Josephson and E. Carroll (Eds.), Drug Use: Epidemiological and sociological aspects. New York: John Wiley and Sons, 1974.

Haberman, P. W., Josephson, E., Zanes, A., and Elinson, J. High school drug behavior: A methological report on pilot studies. In S. Einstein and S. Allen, (Eds.), International conference on student drug surveys. Farmingdale, New York: Baywood Publishing Co., 1972.

Hardy, J. B. The Johns Hopkins collaborative perinatal project: Proceedings of a symposium on factors affecting the growth and development of children. Johns Hopkins Medical Journal, 1971, 128.

Hargraves, W. A., Attkisson, C. C., and Sorenson, J. E. (Eds.). Resource material for community mental health program evaluation. (DHEW (ADM) 77-328). Washington, D.C.: U.S. Government Printing Office, 1977.

Harris, C. W. (Ed.). Problems in measuring change. Madison, Wisconsin: The University of Wisconsin Press, 1967.

Heise, D. R. Causal inference from panel data. In E. F. Borgatta and G. W. Bohrnsted (Eds.), Sociological methodology 1970. San Francisco: Jossey-Bass, Inc., 1970.

Hitch, C. J. and McKean, R. N. The economics of defense in the nuclear age. Cambridge: Harvard University Press, 1960.

Johnston, L. D. Drugs and American youth: A report from the youth in transition project. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1973.

Johnston, L. D. Drug use during and after high school: Results of a national longitudinal study. American Journal of Public Health, 1974, 64 (Supplement), 29-37.

Josephson, E. Trends in adolescent marijuana use. In E. Josephson and E. Carroll (Eds.), Drug use: Epidemiological and sociological aspects. New York: John Wiley and Sons, 1974.

Kandel, D. Interpersonal influences on adolescent illegal drug use. In E. Josephson and E. Carroll (Eds.), Drug use: Epidemiological and sociological aspects. New York: John Wiley and Sons, 1974.

Kenny, D. H. Cross-lagged and synchronous common factors in panel data. In A. S. Goldberger and O. D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 1973.

Kramer, M. Issues in the development of statistical and epidemiological data for mental health services research. Psychological Medicine, 1976, 6, 185-215.

Krutilla, J. V. and Eckstein, O. Multiple purpose river development. Baltimore: Johns Hopkins Press, 1958.

Labouvie, E. W. Longitudinal designs. In P. M. Bentler, D. J. Lettieri, and G. A. Austin, (Eds.), Data analysis strategies and designs for substance abuse research. (Research Monograph NO. 13) Washington, D.C.: National Institute on Drug Abuse, December, 1976.

Lazarsfeld, P. F., Berelson, B., and Gaudet, H. The people's choice. (2nd ed.) New York: Columbia University Press, 1948.

Lettieri, D. J., and Bachenheimer, M. S. Methodological considerations for a model reporting system of drug deaths. In E. Josephson and E. Carroll (Eds.), Drug use: epidemiological and sociological approaches. New York: John Wiley and Sons, 1974.

Lukoff, I. F. Issues in the evaluation of heroin treatment. In E. Josephson and E. Carroll (Eds.), Drug use: Epidemiological and sociological approaches. New York: John Wiley and Sons, 1974.

National Institute of Mental Health. A model for estimating mental health needs using 1970 census socioeconomic data (DHEW Publication No. (ADM) 74-63). Washington, D.C.: U.S. Government Printing Office, 1974.

National Institute of Mental Health. A typological approach to doing social area analysis (DHEW Publication No. (ADM) 76-262). Washington D.C.: U.S. Government Printing Office, 1975a.

National Institute of Mental Health. Mental health demographic profile system description (DHEW Publication No. (ADM) 76-263). Washington D.C.: U.S. Government Printing Office, 1975b.

Nunnally, J. C. The study of change in evaluation research: Principles concerning measurement, experimental design, and analysis. In E. L. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

Pelz, D. C., and Andrews, F. M. Detecting causal priorities in panel study data. American Sociological Review, 1964, 29(6), 836-847.

Richardson, A. H., Hardy, J. B., and Dallas, J. Family planning and population control: attitudes, knowledge and behavior in the inner-city. Report to the Maryland State Department of Health, 1975.

Rosen, B. M. A model for estimating mental health needs using 1970 census socioeconomic data (DHEW (ADM) 75-167). Washington, D.C.: U.S. Government Printing Office, 1974.

Rosen, B. M., Lawrence, L., Goldsmith, H.F., Windle, C.D., and Shambaugh, J. P. Mental health demographic profile distribution (DHEW (ADM) 76-263). Washington, D.C.: U.S. Government Printing Offices, 1975.

Smith, N. L. Techniques for the analysis of geographic data in evaluation. Evaluation and program planning, 1979, 2, 119-126.

Spuck, D. W. Geocode analysis. In H. J. Walberg (Ed.), Evaluating educational performance. Berkeley, California: McCutchan, 1974.

Struening, E. L. Social area analysis as a method of evaluation. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

Sudman, S. Applied sampling. New York: Academic Press, 1976

United States Bureau of the Census. Guide to data for health systems planners. Washington, D.C.: U.S. Government Printing Office, 1970.

Warheit, G. J., Bell, R. A., and Schwab, J. J. Planning for change: Needs assessment approaches. Rockville, Maryland: Department of Health, Education and Welfare: Alcohol, Drug Abuse and Mental Health Administration, 1977.

Warwick, D. P., and Lininger, C. A. The sample survey: Theory and practice. New York: McGraw-Hill, 1975.

Weiss, C. H. Interviewing in evaluation research. In E. L. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

CHAPTER 9: EVALUATION RESEARCH DESIGN AND DATA ANALYSIS

INTRODUCTION

Design of prevention evaluations and analysis of the data collected necessarily demand a number of statistical considerations. These involve issues in social science methodology, statistical theory, and data analysis techniques. This chapter is intended to provide a brief guide to the current state of research design, data analysis and computing resources, and to some of the as yet unsettled issues that surround recent developments in these fields. Coverage and treatment are not meant to be exhaustive; indeed, it would require several times the number of pages presented here merely to list the topics encompassed by the terms "design" and "analysis." Nor is there any attempt to instruct readers in statistical theory or practical techniques; a general familiarity with basic statistical concepts and terminology is assumed. There is, however, ample citation of relevant papers and textbooks.

This chapter falls naturally into three parts. The first treats current topics in evaluation design, including power analysis, design issues surrounding the need for multiple comparisons, and optimum inference strategies based on those considerations. A survey of sampling strategies is included in this design section. The second part of the chapter is devoted to data analysis, including brief reviews of several burgeoning areas of development: log-linear models, exploratory data analysis, nonparametric tests, structural equation models, and path analysis. Also included is an outline of intervention data structures and more theoretical treatments of likelihood inference and the models underlying nonparametric tests. The third part focuses on computing resources available to perform statistical analyses, including a general overview of programs, criteria for selecting statistical programs, an evaluation of general purpose programs, and index of special purpose programs, and a brief discussion of other computing tools. Other important topics such as design selection, panel studies, social area research, and qualitative strategies are not included, being covered elsewhere in this volume.

ISSUES IN RESEARCH DESIGN

POWER ANALYSIS

In planning an evaluation, it is desirable to control two types of error. The first, Type I error, is the probability (usually designated as alpha) of mistakenly rejecting the null hypothesis when it is true. The second error, or Type II error, is the probability (beta) of failing to reject the null hypothesis when it is false. Any given statistical test of a null hypothesis can be viewed as a complex relationship among the following four parameters: (1) the power of the test, that is, the probability of rejecting the null hypothesis (defined as one minus beta); (2) the region of rejection of the null hypothesis as determined by the alpha-level and whether the test is one-tailed or two-tailed (as alpha increases, power increases); (3) the sample size n (as n increases, power increases); and (4) the magnitude of the effect in the population or the degree of departure from the null hypothesis (the larger the effect, the greater the power).

These four parameters are so related that when any three of them are fixed, the fourth is completely determined. After the significance criterion (alpha-level) and the sample size (n) have been decided for a given evaluation plan, the power of the test can theoretically be computed. But because the magnitude of the effect to be studied is not known, the power of a test can only be estimated. Cohen and Cohen (1975) offer several useful things to keep in mind when estimating the size of effects. These are: (1) effect sizes in studies that are closely related to the investigation in question will reflect the magnitude of effects that can be expected; (2) an investigator can sometimes posit a minimum effect that would have either practical or theoretical significance; (3) it is possible to use certain suggested conventional definitions of "small," "medium," and "large" effects (see Cohen 1969), either by choosing one of these values (for example, the conventional "medium" effect size is a population r of .30) or by estimating power for all three population effect sizes; in the latter case the evaluation plan would be revised according to the estimated relevance of the various effect sizes for the problem under investigation.

Because little or nothing can be done after the evaluation is completed, estimation of statistical power is of primary value as a preinvestigation procedure. If power is found to be insufficient to achieve a necessary outcome, the plan may be revised so as to increase its power, either by increasing the sample size or by increasing alpha. When power turns out to be insufficient and revision is impossible, the plan ought to be dropped entirely; this is the reason for power analysis of research plans prior to their execution. A more complete general discussion of the concepts and strategy in statistical power analysis has been presented by Cohen (1965, 1969); further discussion of power analysis in multiple regression and correlation analysis (that is, for B, r, R, partial coefficients, and partialled sets) is included in Cohen and Cohen (1975). The latter text also contains tables for power analysis of multiple regression and correlation.

MULTIPLE COMPARISONS

The fundamental demand of an effective strategy of statistical inference is the balancing of Type I and Type II errors in a manner consistent with the substantive issues of the evaluation. These issues include the level of certainty necessary for decision making, the developmental stage of the program, and the primacy of a given statistical test in the logic of the overall evaluation plan.

In general, this takes the form of attempting to maintain a reasonably low rate of Type I errors, while at the same time not allowing Type II errors to become unduly large or, equivalently, maintaining reasonable power against realistic alternatives to the null hypothesis. Especially when it is necessary to deal with multiple hypotheses, however, procedures for significance testing and power analysis become exceedingly complex. This complexity has at least four highly salient dimensions: (1) whether the Type I error rate is calculated per hypothesis, per group of related hypotheses, or for even larger units (like entire investigations); (2) whether the significance criterion (alpha-level) is held constant or varied over the multiple hypotheses; (3) whether the hypotheses are mutually independent (orthogonal) or dependent; and (4) whether the hypotheses are planned in advance or stated after the data have been examined (post hoc), the latter procedure being sometimes referred to as "data dredging" or "data snooping." Each of the possible combinations of these alternatives has one or more specific procedures for testing the multiple hypotheses, and each procedure has, in turn, its own set of implications for the statistical power of its tests. The following are examples of types of multiple hypotheses, each of which has it own set of procedures.

All Simple Comparisons among Pairs of Means

There exists a large collection of statistical methods for dealing with the problem of making all simple comparisons among pairs of means. These methods vary in their definition of the problem, particularly in their conceptualization of Type I error, and therefore also vary in both power and in results. One of the oldest and simplest procedures is Fisher's

"protected t-test" (Miller 1966; Games 1971; Carmer and Swanson 1973). Only if an ordinary overall F-test is significant at the alpha-level are the means compared, this being accomplished via an ordinary t-test; such tests are protected from large Type I error rates per group of related hypotheses by the requirement that the preliminary F-test must meet the alpha criterion. The Tukey test (Winer 1971) controls the error rate per group of related hypotheses at the alpha-level. The Newman-Keuls test and Duncan test both approach Type I error via "protection levels" which are functions of alpha-level, but the per-hypothesis Type I error risks for the former are constant and for the latter vary systematically (Winer 1971). Bonferroni or Dunn tests employ the principle of dividing an overall alpha-level into as many (usually equal) parts as there are hypotheses, and then setting the per-hypothesis significance criterion accordingly (Miller 1966). These tests of all pairs of means are the most commonly employed, but this is by no means an exhaustive list (see Miller 1966; Games 1971). Each of them approaches the control of Type I errors differently, and therefore carries different implications for the rate of Type II errors and hence for power.

Some Simple Comparisons among Pairs of Means

Sometimes only differences between some pairs of means may be of interest, as when a single control or reference group is to be compared with all other groups. In this special case, the Dunnett test, the alpha-level of which is controlled per group of related hypotheses, would apply (Winer 1971). For the more general case where not all pairwise hypotheses are to be tested, Fisher's protected t-test and Bonferroni tests (and others) can be used. Once again, tests differ in their strategies of Type I error control, and hence have different power characteristics.

Orthogonal Comparisons

Planned (a priori) orthogonal comparisons are generally considered the most elegant multiple comparison procedures and have reasonable power characteristics, but they can be employed only infrequently in social science and evaluation research because the questions put to data in these applications usually are not orthogonal, that is, independent of each other. With g groups, it is possible to test up to g - 1 null hypotheses on orthogonal comparisons (linear contrasts). These may be simple or complex, the latter meaning ones which involve more than two means. When the maximum possible number (g - 1) of orthogonal contrasts are each tested at an alpha-level, the Type I error rate per group of related hypotheses is larger. When orthogonal contrasts are used, however, it is common practice not to reduce the per-contrast rate alpha-level below its customary value in order to reduce the error rate per group of related hypotheses (Games 1971).

Nonorthogonal, Numerous, and Post Hoc Comparisons

The number of different contrasts of all kinds is infinite for more than two groups. Evaluation researchers may wish to make nonorthogonal comparisons, or to make comparisons which were not contemplated in advance of data collection, but rather suggested post hoc by the sample means found in the research. Whether such procedures are viewed favorably as "exploratory analysis" or unfavorably as "data dredging," they often play an important role in social and policy research. Unless Type I error is controlled in accordance with such post hoc procedures, however, the error rate per group of related hypotheses of spuriously significant t-tests on comparisons becomes unacceptably high. The Scheffe test (Miller 1966; Games 1971; Edwards 1972) is designed for such circumstances, and permits all possible comparisons to be made, whether they are orthogonal or nonorthogonal, planned or post hoc; they are subject only to a controlled Type I error rate per group of related hypotheses. Because the Scheffe test is so permissive, however, in many applications it results in very conservative tests, that is, in tests of relatively low power (Games 1971).

Miller (1966) is a comprehensive reference for the statistics of testing the multiple null hypotheses generated from a collection of groups defining a research factor. For social

science and evaluation researchers, Games (1971) provides a comprehensible article-length exposition. Most experimental design texts present some of the more common multiple comparison procedures for means with worked examples and the necessary tables (for example, Winer 1971; Edwards 1972).

OPTIMUM INFERENCE STRATEGIES

Prevention evaluators can afford neither to make spurious positive claims (Type I error) nor to fail to find important relationships (Type II error). All other things equal, these two types of errors are inversely related, so that some balance is needed. At the same time, the complexity encountered with only a single nominal scale makes it clear that any effort to treat this problem in comprehensive detail is outside the bounds of practicality, particularly in most prevention evaluation applications. An alternative is provided by Cohen and Cohen (1975), who present several general principles and simple methods for keeping both Type I and Type II errors acceptably low and in reasonable balance. The three major elements of this approach include parsimony in the number of variables employed, use of a hierarchical strategy, and adaptation of the Fisher protected t-test to multivariate regression and correlation analysis. Each of these three elements is treated here in brief detail.

Minimizing the Number of Independent Variables

Prevention evaluation demands that the number of variables be sufficient to cover the substantive issues involved, while costs in time, money, and increased complexity are kept to a minimum. In addition to the time and money costs of more variables, there are also important costs in terms of the validity of statistical inference which are often overlooked. The more variables--dependent or independent--in an investigation, the more hypotheses are tested (either directly or implicitly) and the greater the probability of spurious significance (that is, Type I error per investigation). It is rare in research proposals and reports of evaluation research, however, to find data analysis appraised from this perspective, and many studies are not reported in sufficient detail to make it possible for readers to do so--variables that fail to prove useful might never be mentioned in print. As the number of variables increases, the greater are the standard errors of estimates and the lower is the power of tests. For this reason, having more variables when fewer would do increases the risks both of finding things that are not so and failing to find things that are so. Nor does a large n solve the problems that accompany large numbers of variables--Type I error rate per investigation depends on the number of hypotheses but not on n. Even potentially high power conferred by a large n may be dissipated by a large number of variables.

Fortunately, it is often the case that a large number of variables is not necessary. It may be that only a few (or even one) of the variables are really central to the construct and the remainder peripheral and largely redundant; these latter variables are better excluded. Sometimes variables may all be roughly equally related to the construct and define a common factor in the factor analytic sense, in which case they should be combined into an index using factor scores or sums of the variables (see the discussion of weighted scoring in chapter 4). Composite variables cannot only represent the construct with greater reliability and validity, but can do so in the sense of a single variable (this is true of both dependent and independent variables). As Cohen and Cohen (1975, p. 161) conclude, "Insofar as variables and hence hypotheses are concerned, an important general principle in research inference is succinctly stated: 'less is more'--more statistical test validity, more power, and more clarity in the meaning of results." In general, the issue here is one of balancing statistical and program significance. The decision to combine variables which convey different (if overlapping) information to decision makers for the sake of controlling Type I error rate should be carefully weighed within the context of the overall goals of the evaluation effort.

Exploiting the Hierarchical Model

The hierarchical model of multivariate regression is one in which each of the independent variables is entered cumulatively according to some specified hierarchy which is dictated in advance by the purpose and logic of the evaluation. The hierarchical model calls for determination of R^2 and the partial coefficients of each variable at the point at which it is added to the equation. Because at each stage the R^2 increases, the ordered series of R^2 in hierarchical analysis is called the "cumulative" R^2 series (Cohen and Cohen 1975).

This hierarchical model can be an important element in an effective inference strategy. Particularly when used with Type I error at each level, the use of the hierarchical model prevents variables of lower priority (which are likely to account uniquely for little variance in the dependent variable) from reducing the power of the tests on variables of higher priority by using some of their variance; this increases the standard errors of the tests' partial coefficients and reduces the degrees of freedom of error. Less weight ought to be given to significant results for design factors of low priority, particularly if many independent variables are involved, because the Type I error rate per investigation is likely to be large. In this way, dilution of the statistical significance of the high priority factors is avoided. As Cohen and Cohen (1975, p. 162) conclude, "The principle is 'least is last'--when research factors can be ordered as to their centrality, those of least relevance are appraised last in the hierarchy, and their results taken as indicative rather than conclusive."

When an order of priority for the independent variables is not specified a priori, other methods can be used, including simultaneous entry, or stepwise entry, in which each variable is entered in sequence based on the extent to which it explains variance in the independent variable after the effects of preceding entries have been removed. Since different methods using sequential entries produce different proportions of explained variance for each predictor, there should be justification for the method chosen (Cohen and Cohen 1975).

Generalizing the Protected t-Test

The simplicity and practicality of Fisher's protected t-test (Miller 1966; Games 1971; Carmer and Swanson 1973) have already been discussed. What is surprising is how effective the protected t-test is in keeping Type I errors low while affording reasonably good power. In an extensive investigation of ten pairwise procedures for means compared empirically over a wide variety of conditions, Fisher's test was unexcelled in its general performance characteristics (Carmer and Swanson 1973).

The protected t-test covers only one set of information on research factors, namely, that defined by nominal scales (that is, a collection of groups). Cohen and Cohen (1975) generalize the protected t-test procedure, applying it to what they term the "functional sets" (large and sometimes diverse groups of independent variables that function as a single covariate in a particular study) that organize multivariate regression analysis. This procedure is effective for several reasons: because the number of sets is typically small, the Type I error rate per investigation does not mount up to anywhere near as large a value over the tests for sets as it would over the tests for the frequently large total number of independent variables. The tests of single independent variables are protected against inflated Type I error rates per set by the requirement that their set's F-value meet the alpha significance criterion. With these Type I errors under control, both the F- and t-tests are relatively powerful; hence both types of errors in inference are kept relatively low and in approximate balance.

POPULATION SAMPLING

Sampling involves various ways of selecting a portion of a total population--that portion observed in order to provide information about the total population. For scientific as well as for purely practical reasons, it is necessary to use sampling procedures that have measurable errors. Procedures should also be capable of characterization relative to bias and variability. The fundamental procedure satisfying these conditions is simple random sampling, a method in which each unit has an equal chance of being selected (usually per-

formed with the aid of random numbers). Systematic sampling, by contrast, is a variation on simple random sampling which proceeds from a random start to select elements at preset intervals.

Samples may be selected in stages by breaking down the population into subgroups. In a multistage random sample, random samples are selected at each stage. If a complete count of sampling units is taken at any stage other than the last stage, the procedure is known as a stratified sample. If the complete count comes at the final stage, it is a cluster sample. The probability of selecting any subgroup may be made proportionate to some function of the size of the subgroup, and the number of units selected from any subgroup may also be made proportional to some such function. Such proportionate sampling tends to reduce sampling errors.

Stratification and clustering can be combined to yield efficient samples, particularly when stratification and/or clustering is based on geographic properties, such as in area sampling. Area sampling reduces the complexity of preparing sampling lists and permits clustering of units in bunches. In double sampling, a first sample is employed to provide information which can be used to design an efficient second sample. Such sampling can also be used to reduce the number of observations required, on the average, for arriving at a conclusion. When double sampling is generalized, it yields sequential sampling, a method of drawing one item or set of items at a time and using the data obtained to decide whether to continue to sample.

All sampling methods based exclusively on random selection and complete counts are probability samples, which yield measurable errors. This is not true of judgment samples, however, which rely on the evaluator's judgment rather than on controlled methods of selection. The ultimate basis for selecting a sampling procedure should be minimization of the costs of getting the sample and the expected cost of errors which may result from the method.

Four books are generally recognized as the best guides to sampling theory. Cochran (1963) is one of the best mathematical introductions, and hence widely used as a textbook in university courses on sampling; it is mute on many practical problems of survey sampling, however. Kish (1967) is the standard reference for most professional survey samplers; it contains a wealth of detail on every aspect of survey sampling except, of course, for the most recent developments. Hansen, Hurwitz, and Madow (1953) still affords a good understanding of large scale government surveys; it is written from the perspective of the U.S. Bureau of the Census statisticians who invented or codified most of the basic theory and methods. Deming (1960) is the standard reference for business oriented sampling statisticians, and contains a skillful integration of theory and application.

Two other books might serve to supplement these basic texts. Babbie (1973, chapters 3 and 4) is a nonmathematical introduction to sampling concepts and types written at the undergraduate level; its materials are more valuable for appreciation of sampling in general than for theory or practical advice, however, and there is no mention of modern methods like telephone sampling. Sudman (1976) is intended as a self help guide for low budget sampling (market research, students, and so on); it is the best source on recent developments like random digit telephone sampling.

Controlled samples based on mathematical probability theory have three basic advantages: (1) they rule out the human biases that might be involved in the more casual selection of items to be observed; (2) they enhance the likelihood that a sample drawn from a population will be representative, that is, will have essentially the same distribution of characteristics as the population from which it is drawn; and (3) probability theory provides a set of computational methods for estimating the degree of error to be expected in a given sample. The basic principle of probability sampling is that every sample will be representative if all members of the population from which it is drawn have an equal chance of being selected (samples with this property are often labeled EPSEM, for "equal probability of selection method"). Even when the EPSEM properties are unobtainable, however, as long as every member of the population has a known, nonzero probability of being selected, probability theory provides computational methods for estimating the expected error of the sample.

SAMPLING DESIGNS

Because the main body of statistics used in sampling theory assumes a simple random sample, many novices conclude that this is the best (or the most accurate) possible method. This is not true. Stratification, for example, has the effect of improving the representativeness of a sample by reducing the degree of sampling error. Moreover, with all but the simplest sampling frame, simple random sampling is not, for practical purposes, possible. The wide variety of sampling designs defined in the previous section are applicable to prevention evaluation; these are described below in more detail, along with their practical advantages and disadvantages.

Simple Random Sampling

In simple random sampling, each population member is assigned a unique number; the sample is then selected via use of random numbers. Simple random sampling has three advantages: it requires only minimum knowledge of the population a priori, it avoids classification errors, and it facilitates analysis of data and computation of errors. Disadvantages of simple random sampling include that it does not make use of knowledge of the population which might be possessed by the evaluator, and it yields larger errors (for the same sample size) than does stratified sampling.

Systematic Sampling

Systematic sampling exploits the natural ordering of a population. A random starting point is selected between the number one and the nearest integer to the sampling ratio (N/n). Items are then selected at the interval nearest (at the whole number) to the sampling ratio. If the population is ordered with respect to some pertinent property (for example, source of referral to the program), then systematic sampling gives stratification effect, and thus reduces variability compared to a simple random sample; this is the major advantage of systematic sampling. In addition, it facilitates both the drawing and checking of the sample. If the sampling interval is related to a periodic ordering of the population, however, increased variability may be introduced; this is the major disadvantage of systematic sampling. When there is such stratification effect, estimates of error are likely to be high.

Multistage Random Sampling

Multistage random sampling involves stages, all of which are a form of random sampling. A major advantage is that sampling lists, identification, and numbering are required only for units belonging to subgroups actually selected. If sampling units are geographically defined, multistage random sampling cuts down on field costs like travel expenses. On the negative side, errors are likely to be larger for multistage random sampling than for simple random or systematic sampling for the same sample size. Errors increase as the number of sampling units selected decreases.

In a major variant of multistage random sampling, sampling units are selected with probability proportionate to their size. This procedure has the advantage of reducing variability; its major disadvantage is that lack of a priori knowledge of the size of each sampling unit increases the variability.

Stratified Sampling

If a complete count of sampling units is taken at any stage other than the final one (which is a cluster sample), the sampling procedure is known as stratified sampling. There are three major variants: proportionate sampling, optimum allocation sampling, and disproportionate sampling. Each of these three variants is discussed here in turn.

In proportionate stratified sampling, selection from every sampling unit at other than the last stage is random with probability proportionate to size. This assures representativeness with respect to the property which forms the basis of classifying units, and therefore yields less variability than simple random sampling or multistage random sampling. Proportionate stratified sampling also decreases the chance of failure to include members of the population because of the classification process; characteristics of each stratum can be estimated, and hence easy comparisons made. On the negative side, proportionate stratified sampling requires accurate information on the proportion of the population in each stratum; otherwise error will be increased. If stratified lists are not available, these may be costly to prepare. There is also the possibility of faulty classification and hence increased variability.

Optimum allocation sampling procedures are the same as those in proportionate sampling except that the sample is proportionate to the variability within strata as well as to their size. This assures that there will be less variability for the same sample size than in proportionate stratified sampling. The major disadvantage is that optimum allocation requires knowledge of variability of pertinent characteristics within each stratum.

Disproportionate stratified sampling proceeds as in the proportionate and optimum allocation variants, except that the size of the sample is not proportionate to the size of the sampling units but is determined rather by analytic considerations or convenience. This procedure is more efficient than proportionate stratified sampling for comparison of strata or where different errors are optimum for different strata. The major disadvantage is that disproportionate stratified sampling is less efficient than proportionate sampling for determining population characteristics, that is, it yields more variability for the same sample size.

Cluster Sampling

In cluster sampling, sampling units are selected via some form of random procedures; the ultimate units are groups, which are selected at random but counted exhaustively at the final stage. Cluster sampling has several advantages: if clusters are geographically defined, it yields the lowest field costs; it requires listing only units in selected clusters; characteristics of clusters as well as those of the population can be estimated; it can be used for subsequent samples because clusters rather than units are selected, and substitution of units may be possible. The disadvantages are larger errors for comparable sample sizes than with other probability samples, and the requirement that each member of the population be uniquely assigned to a cluster; inability to do so may result in duplication or omission of units.

Stratified Cluster Sampling

In stratified cluster sampling, clusters are selected at random from every sampling unit. This reduces the variability of ordinary cluster sampling, but combines the disadvantages of stratified sampling with those of cluster sampling. In addition, because cluster properties may change, the advantage of stratification may be reduced and the sample may not be usable for subsequent research.

Sequential Sampling

Sequential sampling is a procedure whereby two or more samples of any of the types discussed above are taken, with results from earlier samples used to design later ones (or to determine if they are necessary). Sequential sampling provides estimates of population characteristics which facilitate efficient planning of succeeding samples, and thereby reduces error in the final estimate. In the long run, sequential sampling also reduces the number of observations required. It has several disadvantages: it complicates the administration of field work, it requires more computation and analysis than does nonrepetitive sampling, and

it can be used only where a very small sample can approximate representativeness and where the number of observations can be increased conveniently at any stage of the research.

Judgment Sampling

In judgment sampling, a subgroup of the population is selected which, on the basis of available information, can be judged representative of the total population. Either a complete count or a subsample of this group might be taken. Judgment sampling has the advantage of reducing costs of preparing samples and field work because ultimate units can be selected closely bunched. The procedure has serious problems, however. Variability and bias of estimates cannot be measured or controlled, and considerable knowledge of the population and subgroup selected is required (or else strong assumptions made).

Quota Sampling

In quota sampling, the population is classified by pertinent properties that determine the desired proportion of sample from each class. Quotas are then fixed for each observer. Quota sampling reduces the costs of preparing samples and field work because ultimate units can be selected so as to be closely bunched (as in judgment sampling). Quota sampling has the additional advantage that it introduces some stratification effect, with an accompanying reduction in variability. The disadvantages of quota sampling are that it introduces observer bias in the classification of subjects, and thus, makes for nonrandom selection within classes.

Social Network Sampling

Because social scientists have traditionally borrowed their formal procedures of inference from classical statistics, quantitative research in social and policy fields has long been constrained by a peculiarly statistical concept of "population." In the words of the noted statistician, Frederick Stephan, "Statisticians think of populations essentially as sets of objects for which any interrelationships that exist can be ignored" (1969, p. 89). In many social science and evaluation research applications, however, the relationships that link individual units of analysis may have more substantive importance than any property of these units themselves. In such cases, standard statistical procedures apply only by assuming a population of relations (as, for example, when inferences about divorce are based on samples from a population of divorces, rather than from a population of people interrelated by marriage and divorce).

Such diversion of substantive concerns for the sake of statistical tractability has proved increasingly unsatisfactory in a growing number of applied fields--ecology and ethology, sociometry and small group analysis, and the studies of kinship and formal organization, interpersonal influence and power structures, diffusion of innovations, and the spread of rumors. Each of these areas has begun to develop its own quantitative methodologies based on applications of mathematical graph theory, and frequently grounded in iterative computer techniques. Unfortunately, these methodologies usually require complete enumeration or census procedures; there has been relatively little work on sampling and statistical inference for graphs (available work may be found in Goodman 1961; Bloemena 1964; and Frank 1971). For this reason, social network analysis has been largely confined to small groups.

Granovetter (1976) has proposed a method for sampling social networks to estimate average acquaintance volume (the mean number of links per individual) and network density. His "subgraph" approach encounters practical difficulties, however, particularly in large populations which are necessarily sparse. Density is an inverse function of population size, while average acquaintance volume has a practical upper limit in the number of relations that any one individual can maintain. This means that, in populations much larger than 100,000, the expected number of relations to be found can fall below one per interview for interviews that contain up to 500 yes-no relational questions (for details, see Beniger 1976, pp. 228-229; Morgan and Rytina 1977).

An alternative to Granovetter's subgraph approach to network sampling is afforded by the "subgroup" methodology of Beniger (1976). Under this approach, a population is first partitioned by criteria salient to respondents into a manageable number of mutually exclusive and exhaustive subgroups. These subgroups are then used, in place of individual names, for at least the first round of network interviewing. This allows estimation of densities by means of a measure called "estimated density spaces" (EDS), a practical approximation of simple network density.

INSTRUMENTATION

No matter how well designed and executed an evaluation, the meaning of the results is completely contingent on the appropriateness of the instruments used for measurement. The worth of an instrument depends on its reliability and validity, both of which are functions of the population being measured and the conditions of administration. (See the appendix to part II for more detailed discussion of these issues.)

When well constructed measures are available, the evaluator should use them in preference to developing new ones. Standardization of measures facilitates comparison of findings. This in turn provides the means by which theory is refined and verified. However, given the wide variety of constructs involved in prevention program objectives, and the diversity of approaches for achieving these objectives, appropriate standardized measures might not be available. A brief outline of the steps involved in developing a measure follows (adopted from Struening 1979):

(1) Select a domain of interest, e.g., attitudes toward drug abusers.

(2) Consider the specific uses you intend for the measure, e.g., evaluation of change (pre/post), comparison of groups, assessment of individuals.

(3) Review the literature for theory relevant to your area of interest, for studies in which available measures have been used, and for scales currently being used. Look carefully at the scales to make certain of their psychometric properties, the population on which they have been standardized, face validity of items, etc.

(4) Conceptualize the domain of interest into subareas by identifying, defining, and limiting relevant concepts.

(5) Select a scale model or method, e.g., factor analysis, Likert analysis, Guttman scaling. (The additive probabilistic model and latent trait theory as they apply to measure development will be discussed later.)

(6) Develop an item sample or pool.

(7) Develop the questionnaire from the item pool.

(8) Select a sample from the population of interest and institute a pilot test.

(9) Collect and analyze the resulting data, based on the previously selected scale model.

(10) Make a final selection of items which define the construct, considering reliability (internal consistency and stability) and validity (including face, predictive, concurrent, and construct validity).

(11) Apply the questionnaire, bearing in mind such issues as dimensional invariance of item sets over populations, replicability of the measure, and its relevance to the target population.

For detailed treatment of the issues involved in testing theory and construction, the reader is referred to the American Psychological Association (1974); Cronbach (1960); Cronbach and Meehl (1955); Ghiselli (1964); Nunnally and Durham (1975); Nunnally and Wilson (1975); and Thorndike (1971).

Among the most important principles stemming from the concepts of reliability and validity is that any item (question, rating, test, and so on) is a construct/ method unit. Each item can be viewed as representing a different method, and every condition under which the item is asked may be viewed as a different method. Since any derived "score" is contaminated by the method used to measure the underlying construct, it follows that the variance contributed by the method (that is, the "noise") can be reduced by obtaining multiple measures of the same construct by different methods. For indeed, a composite score based on a large number of highly intercorrelated measures using different measurement procedures is likely to yield the most valid and reliable measure.

The most practical model in general is the additive probabilistic model. Scott and Wertheimer (1962) give a good description of the characteristics of composite measures derived from this model, which can serve as criteria for anyone developing a new measure. Some of these characteristics follow:

- Any respondent's total score on the composite measure consists of the sum of the number of items indicating the construct in question. (Note: although weighting of individual items can be done by a variety of methods, it is not recommended because the time and expertise required doesn't add much to the outcome; further, weights vary across populations.)

- The contribution of an item to the variance of the total score distribution across all subjects is proportional to the items' individual variance.

- Every item in a composite measure correlates positively with every other item.

- Every item correlates positively with the total score. This property, a consequence of the last, becomes an index of the measure's homogeneity. The higher the average item-total score correlation, the more likely is the measure to describe one and only one construct.

- A good measure will be stable from one administration to the next, assuming the attribute has not changed in the interim. This property of test/retest reliability is very important in evaluation of change--if scores fluctuate in unsystematic ways over time, the likelihood of finding true differences between control and experimental groups is reduced.

- Different ways of measuring the same construct should yield similar scores. This property, referred to as convergent validity (Campbell and Fiske 1959), is satisfied when two or more measures of the same construct intercorrelate to a substantial degree.

- Finally, a measure of one construct should not correlate highly with another designed to measure a different construct. This property is known as discriminant validity (Campbell and Fiske 1959). Two measures of the same construct should correlate substantially higher with each other than they do with two measures of another construct (and vice versa). Convergent and discriminant validity must be considered together in evaluating the adequacy of a set of measures.

All of the characteristics listed above may serve as yardsticks for evaluating existing measures, as well as basic guides for instrument development. It must be emphasized that this model is only one of many, and is not appropriate for all situations. Although examina-

tion of the other major models is beyond the scope of this discussion, one other relatively new model is worthy of comment; though it is highly complex, it is valuable because of the promise it holds for measure construction in the future.

This methodology for measure construction is based on latent trait theory, which is discussed in greater detail later in this chapter. Latent trait theory supposes that responses to a test can be predicted or explained by latent traits (or abilities). Since these traits cannot be directly measured, a model is developed which specifies a relationship between observable performance and the underlying trait. These models assume local independence of items. That is, the probability of a given response to one item is not affected by the test responses on any other item. These models have in common the development of item- and test-characteristic curves, which give the frequency of responses for varying levels of trait possession (Hambleton, et al. 1978).

To develop a new measure, one uses a procedure such as that outlined below (Birnbaum 1968; Lord 1977).

- Develop item-characteristic curves for a selected pool of items.
- Decide on the shape of the target information curve, which specifies the desired accuracy of trait estimation at each level of trait possession.
- Select items with item-characteristic curves which fill hard-to-fill areas under the target information curve.
- Cumulatively sum the item curves, continuing to add (or delete) items, until the area under the target information curve is satisfactorily approximated.

Latent trait models have not yet been extensively used in the development of attitude scales, in part due to lack of familiarity with the models and, until recently, lack of availability of computer programs for estimating item and trait parameters. However, this may change as evaluators grow more confident of these models' usefulness.

ISSUES IN DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

Most texts on social statistics (Blalock 1972 remains one of the best) concentrate on two broad goals of quantitative analysis: how to summarize large bodies of numbers (using means, standard deviations, and so on), and how to confirm the results of an analysis using tests of statistical significance (which help protect against sampling and measurement error). These same texts usually have relatively little to say about how to discover the unanticipated or how to expose some new relationships in data (but see Tukey 1969, 1977; Mosteller and Tukey 1977). Graphical techniques, for example, are among the most powerful means of both discovering and informing, yet most texts give only cursory and often misleading advice about displaying data. Similarly, most discussions of how to fit straight lines to data--potentially the most powerful technique of data analysis in most situations--concentrate on how to determine whether a particular regression coefficient is significantly different from zero. Usually little space is given to analyzing deviations (residuals) from a fitted line, or how to transform the variables entering a regression--two basic techniques for discovering patterns in data.

In short, most statistics texts perpetuate the impression that the important uses of quantitative methods in social science and evaluation research are either to summarize large bodies of data or to confirm an observed relationship at the .05 level of statistical significance. However, this view of data analysis in social and policy applications has begun to give way over the past decade in favor of an "exploratory" approach to data analysis (see Tukey and Wilk 1965; Tukey 1962, 1969, 1977; and Tufte 1970). Significance tests are given

a secondary role in this approach, and the distinction between interval and ordinal measurement is usually of little importance. Correlation coefficients are considered often misleading and are used only as a partial first step in analysis, if at all. Instead, analysis begins with the fitting of lines to relationships between variables (transformed variables where necessary), and then proceeds with the examination, often with the aid of scatterplots and graphs of deviations from the fitted line. Each of these new views toward data analysis are briefly examined below.

Significance Testing

Significance tests help protect against the possibility that a relationship arises in a random sample through chance alone (good discussions of significance may be found in Kish 1959 and Kruskal 1968). Such tests are also useful as a rough screening device in the analysis of data collected nonrandomly, where they may help to assess the certainty or uncertainty of results. Responsible evaluation researchers use tests of significance or, usually better, confidence intervals to assess the stability of results, the latter being preferred due to the greater amount of information provided to decision makers. Significance levels are often misused, however, because the dichotomy between "significant" and "nonsignificant" is too sharply drawn, with only those relationships (and all those relationships) that reach the .05 (or sometimes .01) level being accepted as meaningful results. The relevance of a result does not depend on its exact significance level, however, but rather on the substantive judgment of the interpretor. The overemphasis on significance testing may have arisen because of certain abstract reconstructions of what scientists do (see Kaplan 1964), rather than actual scientific practice. The rote of significance testing--assumptions, sampling, distribution, critical region, test statistic, decision--does not provide a guide to data analysis, and may serve to make prevention evaluators feel unnecessarily guilty about violating severe and unrealistic assumptions (Tufte 1970).

Probability levels and test statistics tell little about the strength, and nothing about the substantive significance, of a relationship among variables. The important question is whether the relationship is of substantive interest by virtue of its nature and magnitude; significance tests do not inform on this issue. Even after test statistics are computed, the researcher must face the problem of what to do with significant relationships. There is no guide other than individual judgment for appraising and weighting reports that show significant relationships. Regression methods may aid the researcher, but finally, the extent of real effects of a treatment may remain an open, statistical question.

Converting Ordinal into Interval Data

Although there is an obvious conceptual distinction between ordinal and interval measurement (Stevens 1968), the important question for evaluation research is whether the distinction has any practical meaning for substantive interpretation. The evaluator or prevention professional often knows more about the phenomenon under study than is implied by the mere ordering of observations. When this is the case, numbers ought to be assigned to the ordered categories; this helps to build the additional information into the measurements. Such a procedure will always contain an arbitrary element, of course; the point, as Tukey has put it, is to be wisely arbitrary (for discussion of the problem of being arbitrary, see Nunnally 1967).

There are two major potential gains from assigning numbers to ordered categories. First, this procedure improves measurement by taking advantage of any information in addition to the fact of ordering--numbers put more substance into measurement. Second, considerably more powerful techniques can be used in the analysis of interval as opposed to ordinal data, techniques which thereby increase the chances of learning something new. In contrast to these two potential gains, that is, better measurement and better analysis, there appears to be little in the way of potential costs. (See, for example, Anderson 1961).

Several methods have been proposed for the assignment of numbers to ordered categories (Abelson and Tukey 1959; Shepard 1966). Tufte (1970) summarizes these in three

simple but useful rules: (1) assignment should incorporate the investigator's substantive understanding of the variable measured; (2) simple linear assignment of numbers to categories usually won't do (in any event it is not more statistically conservative than any other assignment); and (3) assignment should often be made so that the distribution of counts looks something like a normal distribution. Above all else, dichotomizing is the poorest strategy that can be imposed on data and should never be considered. Even with the choice of the optimal cutting point, a major amount of information is lost by dichotomizing data.

Avoiding Correlation Coefficients

Correlation coefficients are often used to summarize a relationship between two variables. Such coefficients have serious defects, however, and are probably vastly overused in evaluation research; indeed, Tukey (1954, p. 38) has recommended that "most correlation coefficients should never be calculated." Correlation is a poor way to summarize actual data as revealed in scatterplots, since plots with great variation (markedly linear, curvilinear, cloudlike except for extreme outliers, and the like) can have the same coefficients of correlation. For this reason, any computer program that produces a correlation matrix should also produce scatterplots of the relationships, and correlation coefficients and plots ought to be considered together in analysis and evaluation. The major advantage of scatterplots and other graphical procedures is that they allow the evaluator to decide how much to learn from the data, instead of having the relationship summarized--perhaps unrecognizably so--by a correlation coefficient or other summary statistic. Also, measurement error reduces the correlation between measures, a phenomenon referred to as "attenuation." Although certain reliability estimates can be used to correct for attenuation, such estimations are difficult to make. Attenuation is not a major problem with simple correlations, but can seriously affect results obtained from partial correlations, partial regression weights, and the analysis of covariance or analysis of variance with gain scores.

The shortcomings of correlation coefficients are not mitigated by more complex models using partial correlations, such as causal or path models that depend on standardized regression coefficients (Forbes and Tufte 1968). Similarly, correlations between typical ratios or indexes (for example, welfare expenditures per capita) are also misleading, and the questions they are designed to answer can be more usefully framed as regression problems (Kuh and Meyer 1955; Wallis and Roberts 1956). Scatterplots and regression coefficients are more useful than correlations, especially when increments along the scales of each variable make some sort of substantive sense. In general, the safest procedure is to use both correlation and regression coefficients when they are meaningful.

Fitting Lines to Data

Evaluation researchers have tended to ignore the useful regression procedures that fit lines to data. Fitting lines to relationships between variables (or variables that have been transformed) and then examining the deviations (residuals) off the fitted line by scatterplots and graphs has at least four virtues. First, in order to do regression analysis, the researcher must have a fairly clear idea of just what is to be explained; this must include some notion of causality, with both a dependent (response) variable and one or more independent (explanatory) variables. Second, fitting lines to data generates residuals (those parts of the variation in the dependent variable left unexplained by the describing variables), and analysis of residuals is a major tool for further discovery. Third, the resulting regression coefficients, especially if they are unstandardized, can have substantive meaning and policy implications. And last, a large body of useful experience in the application of regression methods to substantive problems has been built up by econometricians; their advice to evaluation researchers is often more useful than that of other sorts of statisticians.

The analysis of residuals deserves particular attention in prevention evaluation. Usually the independent variables selected a priori do not account for all the variation in the dependent variable of interest. Trying to find something that will explain some of the residual variation helps to discover the unanticipated in a data set.

The initial steps in the analysis of residuals are simply to calculate the deviation from the fitted line for each observation, look at the whole collection of these residuals, and then see what those observations with residuals of the same size have in common. Then the residuals ought to be plotted in the following ways: (1) against fitted or possible values; (2) against variables which have been employed as a basis for the summarizing of fit; (3) against variables which were not used in the fit, such as time; (4) identified according to some meaningful characteristic, for example, according to whether the residual is or is not from an observation which was used in developing the fitted summary; and (5) ordered in probability plots, with empirical cumulative distribution plots, and plots of empirical quantiles against quantiles of reference distributions such as the unit normal. Although all such plots provide indications of the spread of the body of residuals, it is far more important that they combine palatable summaries of individual residuals with sensitive indications of distributional peculiarities of the entire collection of residuals.

Standard mathematical transformations and any other systematic changes in the observed values of variables can play an important role in the process of fitting lines to data. Such transformations are useful for three reasons: they allow the use of linear techniques to fit rather complicated, nonlinear models to data; they often point to substantive results; and they help data to satisfy certain statistically desirable properties, such as normality and stability of variance. The best introduction to transformations is the straightforward essay by J. Kruskal (1968). For converting nonlinear models into linear fit problems, see Draper and Smith (1966).

Transformations are also useful in analysis of percentages. For example, it is not always appropriate to assume that a given difference between two percentages (say a five percent increase) has the same meaning as an equal difference between two other (different) percentages. The data analyst should, when it is appropriate, take into account that a five percent increase starting from a high percentage as the base is actually often a bigger and more important substantive change than the same amount of percentage change beginning at some lower initial level. When this reasoning is correct, the tails of percentage distributions ought to be stretched by transformations of percentages and appropriate tests of significance made. (A number of options are presented in Hovland, Lumsdaine, and Sheffield 1955 and in Kruskal 1968.)

One caveat is in order. As noted by Klitzner (1975), it is often too easy to transform data which should not be transformed. Substantive justification for transformations (for example, converging evidence of response nonlinearity, wildly nonnormal distributions, or theoretical reasons to doubt patterns such as interactions in nontransformed data) should precede data analytic concerns, and nontransformed data should generally be used whenever possible. Data transformations run the risk of masking important effects, and of complicating the relationship between data analysis and the phenomenon of interest.

This, then, is the new view of statistical analysis in the social and policy sciences that has emerged in the past decade under the general rubric of "exploratory" data analysis. With this approach, significance tests are given a secondary role, the distinction between interval and ordinal measurement is often blurred, and correlation is used only as a partial first step in analysis. Furthermore, the fitting of lines and examination of residuals, often with the aide of scatterplots and graphs, is elevated to a central activity. This approach is aimed at discovering the unanticipated in data and exposing new relationships therein, as well as informing users of social and policy research about what is going on in a data set of interest. Given these developments, it is likely that fewer and fewer applications of statistical methods in the future will be used only to summarize large bodies of numbers, or to confirm the results of these analyses using tests of statistical significance.

INTERVENTION DATA STRUCTURES

Data analysis for outcome and impact evaluation involves the tools required to estimate whether particular interventions (treatment or treatments) have an effect, whether this effect is in a desirable direction, the magnitude of the effect, whether it differs from the effects of other interventions, and so on.

In order to appreciate the wide range of diverse statistical techniques available to analyze evaluation data, it is helpful to view such data as necessarily structured along five dimensions with respect to the policy intervention (treatment):

- A time frame (i.e., "before and after" measurements or multiple observations of a time series).
- The possibility of a control group (which might be units of analysis not given the treatment, chosen by randomized design or some type of self-selection).
- Existence of a pretest (i.e., a measurement or measurements prior to the intervention); this might be simply norms for the population or similar populations.
- The possibility of separate samples within treatments, for which observations might be made either before or after treatment (thus introducing something of a control group).
- Random assignment of units to groups (which permits one to assume that the groups are approximately equivalent).

Note that these five features are dimensions of evaluation data, not necessarily of evaluation research designs. Although such designs can be located on the five dimensions (see figure 9), data analysis for prevention evaluation is possible using data sets that have not been designed (or have been poorly designed) to test policy interventions. Indeed, statisticians are routinely called upon to reach evaluative conclusions long after research has been designed and executed. The implications for statistical analysis of each of the five dimensions are outlined here in turn.

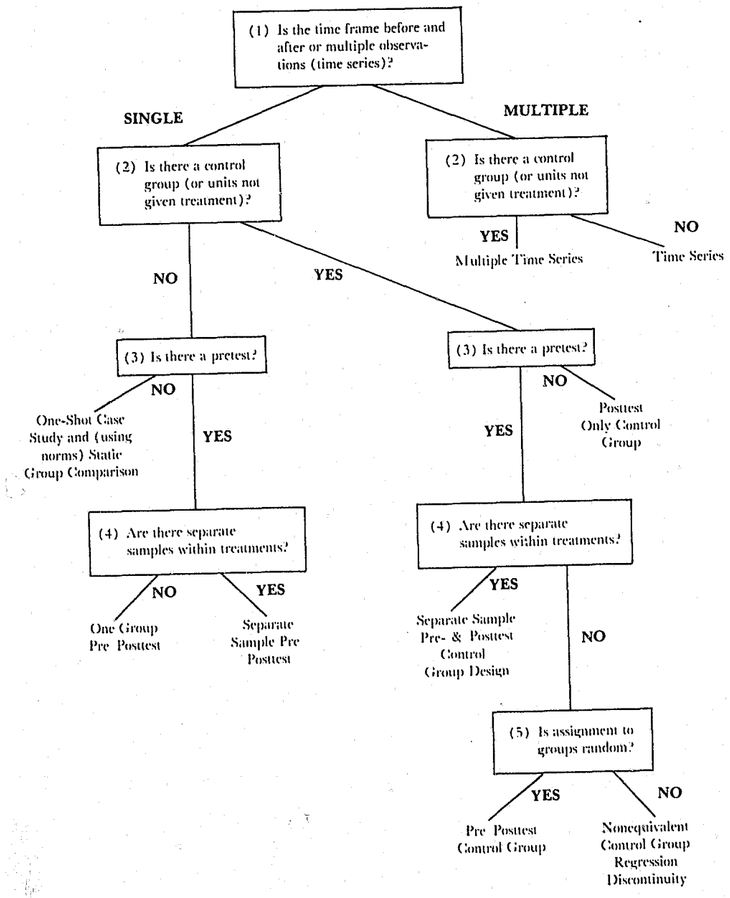
Time Frame

If data include multiple observations, time series analysis is possible. This permits not only identification of possible treatment effects, and measurement of their magnitude and direction, but also assessment of the temporal trends of effects. Trends in pretest observations may be used to counter the effect of maturation on internal validity. Trends in posttest observations are particularly useful in assessing effects which are not static or permanent. Most general linear regression techniques can be adopted to the types of data structures likely to be found in evaluation research involving multiple observations. A newer approach to the analysis of time series is the autoregressive integrated moving average (ARIMA) model developed Box and Jenkins (1976) and discussed in the context of quasi-experiments by Cook and Campbell (1979). The major advantages of this approach include techniques for detrending, deseasonalizing, and correcting for autoregression.

Control Groups

The existence of a control group, or units not given the treatment (whether by randomized design or some type of self-selection), permits assessment of an intervention with the state that would have obtained in its absence (in effect the comparison of two or more treatments, at least one involving no change at all). Various tests for comparison of treatments might be employed when control groups are available, including such parametric tests as ordinary least squares regression and such nonparametric tests as the Wilcoxon rank-sum test and the Siegel-Tukey and Smirnov-Kolmogorov tests. When assignment to the control group is not random, multivariate statistical procedures might be considered as a means to control confounding variables after the fact. Most standard evaluation research designs will include control groups whenever possible.

Figure 9. Standard Evaluation Research Designs Located on Answers to Five Data Analysis Questions



Pretest

The existence of a pretest (observation prior to the treatment), which might involve simply norms for the population or similar populations, is not really essential to true experimental design; randomization can suffice without a pretest. Whenever randomization is not possible, however, or when sample sizes are small, pretests can serve to assure lack of initial biases between groups. Simple slope provides a straightforward measure of change between pre and posttest. Such change data are provided by most standard evaluation research designs.

Separate Samples Within Treatments

Separate samples within treatments provide an opportunity to make observations either pre or posttreatment, thus introducing something of a control group into the statistical analysis. Both the Separate-Sample Pretest-Posttest, and the Separate-Sample Pretest-Posttest Control Group, are examples of quasi-experimental designs that include separate samples within treatments (see chapter 7). As with pretest-posttest analysis, simple slope provides a straightforward measure of change between pre- and posttest observation (which here are for separate samples within the treatment).

Randomization

Randomization is the key to true experimental design, but it is not a central consideration in deciding the particular type of statistical analysis of evaluation research data. Because randomization is used to assure the equivalence (lack of initial biases) among groups, however, its absence is always a signal to the data analyst that multivariate statistical techniques might be considered (measurement of the appropriate variables permitting) as a means of controlling after the fact for confounding variables. The probability and extent to which randomization does in fact render groups equivalent is a function of sample size, of course--a second way in which randomization can enter into the analysis of evaluation data. Randomization is particularly important in distinguishing Pretest-Posttest Control Group design data from that generated by Nonequivalent Control Group designs and regression discontinuity analysis (Thistlethwaite and Campbell 1960).

MULTIPLE POPULATIONS

Evaluation researchers often obtain data from different populations (control versus treatment groups, different cities, regions, schools, and school systems, and so on). Often the desire or need to pool such data raises the question of the degree of homogeneity among various populations, and the statistical validity and shortcomings of pooling. Goodman (1973) presents methods for analyzing the homogeneity and heterogeneity of data cross-classified for several populations. He introduces a threefold classification of models: (1) those that assume "complete homogeneity" among the tables, (2) those that assume "complete heterogeneity" among the tables, and (3) those that allow "partial homogeneity." Goodman (1971) also extends the stepwise procedures originally presented for log-linear analysis of multidimensional contingency tables. He also introduces "guided" and "unguided" selection methods and "multidirectional" methods.

Ecological Inference

When appropriate individual level data are not available, social scientists and evaluation researchers have routinely used aggregate data to make inferences about individuals. Such inference across levels of aggregation, sometimes called "cross level inference," can be either downward (ecological inference) or upward (individualistic inference); the latter refers to

the use of individual level data to make inferences about aggregate level effects (the terms "individual" and "aggregate" refer to units of analysis; an individual need not be a person).

Such practices have been frequently critized in the past three decades, ever since a paper by Robinson (1950) demonstrated that correlations between variables at the aggregate level differ from correlations between the same variables at the individual level. From this finding Robinson concluded that researchers should not use aggregate data to study individuals; those who did were said to be guilty of the ecological fallacy. The analogous shortcoming of upward cross level inference is know as the individualistic fallacy (Alker 1969).

Because of the unavailability of individual level data for many areas of interest to social scientists and policy researchers, there have been many attempts to modify the strict prohibition against downward cross level inference. This problem is of particular interest to prevention evaluators because data on illicit substance use and other illegal activities are often available only as aggregates. The most important conclusion of these discussions has been that aggregate data do not always yield biased estimates of individual level unstandardized regression coefficients (Goodman 1953, 1959). This conclusion has two implications: (1) there are certain cases where, except for possible loss of efficiency (that is, the variance of the coefficient of regression of means on means is usually greater than that for the regression of individual values; see Hannan and Burstein 1974), downward cross level inference can be made with impunity; and (2) if downward cross level inference is made, regression coefficients should be used instead of correlation coefficients.

Following the finding that an aggregate level regression coefficient need not differ from its individual level counterpart, several studies sought to determine the conditions under which the regression coefficients do not differ, that is, the conditions under which cross level bias is absent. These efforts have proceeded along two discernible lines of inquiry: the contextual effects approach, which views such effects as the major source of bias (Hammond 1973; Przeworski 1974); and the structural equations or causal models approach (Blalock 1964; Hannan 1971a and 1971b; Hannan and Burstein 1974), which formulates bias in terms of path models and uses econometric techniques to determine the expected value of the parameters.

Hammond (1973) suggests a link between contextual effects theory and cross level inference. Hannan and Burstein (1974) counsel researchers faced with the question of cross level inference to consider the effects of the variable by which the data are grouped (school districts, counties, and so on); in the bivariate case, at least, aggregate data give unbiased estimates of the individual level relationships when any of the following is true: (1) the grouping variable is uncorrelated with the dependent variable controlling for the independent variable; (2) both the grouping and independent variables are uncorrelated; or (3) the variance of the independent variable equals the variance of its group mean.

Recent work by Firebaugh (1978) combines the two approaches by employing contextual effects models in a structural equation framework. This generates a parsimonious rule--the Group Mean (of the independent variable) Rule--for making inferences about individual level relationships from aggregate data. The Group Mean Rule states: bias is absent when, and only when, the group mean of the independent variable has no effect on the dependent variable controlling for the direct effect of the independent variable itself (that is, when the dependent variable is regressed on both the independent variable and the group means, the regression coefficient for the latter equals zero).

The Group Mean Rule links cross level bias to theory on group effects, which is well known in the social science literature (for example, Blau 1960), and hence provides theoretical leverage to the researcher who must determine whether cross level inference is legitimate in a particular case. The Group Mean Rule is also easily generalizable analytically to the multivariate case, which is important because there are formidable obstacles to analytical investigations of the effects of grouping in regression models containing two or more regressors.

Often, researchers faced with possible cross level bias do not have individual level data, and hence cannot empirically determine whether their data conform to the Group Mean Rule. This is a problem with downward cross level inference, however, and does not differ in principle from the specification problem faced in all causal analyses. In regression anal-

ysis, the researcher must always make assumptions about the data used, and the validity of these assumptions only rarely can be tested empirically, rather than decided on theoretical grounds.

ANALYZING QUALITATIVE OR CATEGORICAL DATA

In many prevention evaluations, the important variables are qualitative in nature. Unlike quantitative variables, qualitative ones (referred to in various literature as categorical, nominal, or discrete variables) pertain to classifications rather than to measurements. They include nominal variables like marital status, for which the categories (single, married, separated, divorced, and so on) are unordered. They also include ordinal variables such as attitudes toward legalization of marijuana (unfavorable, neutral, favorable) or amount of education (grade school, high school, college, postgraduate), in which the categories are ordered. Just as numbers are associated with the "values" of a quantitative variable, categories or classes are associated with the "levels" of a qualitative variable.

Traditional methods of cross-tabular analysis of qualitative variables have proven inadequate for answering many of the questions routinely posed in prevention. At the same time, the recent development of more appropriate statistical methods, together with the availability of associated computer programs, has made it possible (and relatively easy) to approach multidimensional cross-tabular data in totally new ways. Work on discrete multivariate analysis in general (for a comprehensive text, see Bishop, Fienberg, and Holland 1974), and Goodman's work on hierarchical and log-linear models in particular (conveniently drawn together in Goodman 1978), adresses the evaluation researcher's pressing need for a systematic and unified approach to the analysis of qualitative data.

This new approach includes procedures for building statistical models and testing hypotheses pertaining to qualitative data. The general approach has proven valuable for identifying relevant interaction effects among multiway contingency tables, and for revealing relatively simple structures underlying seemingly complex relationships among the variables. Applications of the approach are virtually unlimited, both in evaluation research and in the basic academic disciplines.

Goodman's Work - Hierarchical and Log-Linear Models

Particularly useful has been the system of contingency-table analysis developed by Goodman in a spate of papers between 1968 and 1975 (the best of these are reprinted in Goodman 1978). Goodman's system consists of two logically and practically distinct parts: a scheme for making significance tests by means of "hierarchical models," and a set of techniques known as "log-linear models." Hierarchical models have several important applications: tests for the significance of partial correlations; tests for interactions (specifications) where the control variable has an unlimited number of categories; tests for higher order (three or more variable) interactions; and succinct statements of what is and is not going on in contingency tables (all of these are uses are described clearly and in familiar terms by Davis 1974). None of these tools had been readily available to social scientists and evaluation researchers prior to Goodman's work on hierarchical models, which also affords considerable insight into the general properties of cross-tabulations and the logic of significance tests.

Log-linear models are based on maximum likelihood methods, and provide a more unified system for analyzing cross-tabular data. The general log-linear formulation incorporates all "interactions" in the multidimensional contingency table without the need to designate some of the variables as dependent on the others (applications of this formulation to five data sets are found in Goodman 1978, chapters 3-6). Goodman has developed stepwise procedures and direct estimation methods for multiple classification model building that are somewhat analogous to the classical stepwise regression methods for adding terms (forward selection) and deleting terms (backwards elimination) for quantitative models. (These methods are illustrated in Goodman 1971 and 1973.)

Logit Regression

Goodman's log-linear models subsume his earlier work on the logit model (Goodman 1972), a modified multiple regression approach to the analysis of dichotomous variables. The modified approach is designed especially for use with qualitative data, particularly a simple regression in which the dependent variable and all the explanatory variables are dichotomous. In this case, classical assumptions underlying regression analysis are violated. Particularly troublesome, when the dependent variable is dichotomous, is the fact that the least-squares method of regression estimation can yield probability estimates outside the defined range of zero to one. Goodman's modified approach is to use the logit model in place of the usual linear regression model and to use maximum likelihood estimation procedures in place of least squares.

Because Goodman's logit regression is designed especially for qualitative data, evaluation researchers ought to use this approach in place of traditional regression whenever their dependent variable can be treated as dichotomous. The parameters of the multiplicative version of the logit model have certain similarities to the parameters (regression coefficients) of a dummy variable regression model: just as the latter parameters express changes in probabilities, the former express changes in odds (probabilities are converted to odds by dividing the corresponding probability by one minus the probability; odds are converted to probabilities by dividing the corresponding odds by one plus the odds). The traditional regression approach (using dummy variables) often yields interpretations similar to those of the logit approach, but the traditional model usually does not fit the data as well as the corresponding logit model, and often yields different conclusions (for a clear and compelling example, see Goodman 1978, chapter 2).

Grizzle, Starmer, and Koch (1969) and Theil (1970) have also introduced logit models somewhat similar to Goodman's, with the important difference that estimation methods and analysis are based on weighted least-squares procedures rather than maximum-likelihood estimates. Weighted least-squares estimates have somewhat larger variance than maximum-likelihood estimates (see Rao 1965). It is also more difficult to use the methods proposed by Grizzle, Starmer, and Koch and by Theil than those of Goodman, at least for four-way (Goodman 1972) and five-way (Goodman 1970) contingency tables.

Although logit regression based on maximum likelihood estimation is a worthy replacement for traditional regression techniques, especially for qualitative data, it is not a substitute for more general log-linear analysis. Whenever a single dichotomous variable can be viewed as the dependent variable of interest, logit regression can be employed. When this is not the case, and several variables are to be analyzed simultaneously as a function of each other without designation of independent and dependent variables, logit regression is inappropriate, and the more general techniques of log-linear modeling (Goodman 1978) ought to be considered.

Traditional Regression, Logit, and Log-Linear Models

The logit model is a special case of the general log-linear model where the parameters associated with the explanatory variables are considered fixed. Like the traditional regression model, the logit model expresses a conditional relationship between the dependent (response) variable and fixed values of the independent (explanatory) variables. The more general log-linear model, by contrast, can also be formulated when all variables are dependent (responses), and hence all parameters are free to vary. For this reason, log-linear models are more general than even the traditional regression approach as most generally defined. With log-linear models, variables can be designated dependent on other variables (as in the logit model) or each variable can be analyzed simultaneously as a function of all others without designation of independent and dependent variables. The log-linear model can also be extended to quantitative explanatory variables or to a mixture of qualitative and quantitative explanatory variables (for example, see Nerlove and Press 1973; Haberman 1974).

There are at least two other respects in which Goodman's log-linear approach is more general than traditional regression, to the advantage of evaluation researchers. First, a central place in Goodman's hierarchical system is occupied by interaction effects, which are often important in evaluation research applications, but which can be incorporated in traditional regression models only as products of the independent variables. This is a clumsy approach, at best, and results are often difficult to interpret because they depend on whether the independent variables are standardized, along with other related issues (see Mosteller and Tukey 1977). Although interpretive difficulties are also encountered when Goodman's approach is extended to quantitative variables, for qualitative data his symmetric interactions have a natural interpretation within the context of his hierarchical system. Moreover, his "saturated" model includes all possible interaction effects, while traditional regression models are typically assumed to be linear with no interactions at all, or at most a small number added arbitrarily.

The second way that Goodman's log-linear models are more general than regression models, to the advantage of evaluation research is that the former depend less on classical distribution assumptions. Normality is assumed in the analysis of a quantitative dependent variable using traditional regression techniques. In actual practice, including many prevention evaluation applications, the dependent variable of interest may be far from normally distributed around a regression line (extreme deviations form the classical assumptions occur when the dependent variable is dichotomous, or consists of only a small number of discrete categories). Because of these practical realities, log-linear models are often better suited to evaluation research.

The New Latent Structure and Scaling Models

Latent structure and scaling models are necessary considerations for evaluation researchers because many of the phenomena of interest in evaluation are not subject to direct measurement. For example, classification of a personality type, assessment of degree of alienation, diagnosis of a psychological syndrome or medical disorder, all require indirect observations on a set of indicators or symptoms of the phenomenon to be measured. Because these indicators are often qualitative variables (categorical, nominal, or discrete), recent developments in discrete multivariate analysis in general (see Bishop et al. 1974), and Goodman's work on hierarchical and log-linear models (collected in Goodman 1978) in particular, have revolutionized latent structure and scaling models in the past decade--the culmination of a half century's work in the area.

Likert (1932) first proposed the method of additive scales in which response categories are assigned simple weights (such as zero or one for dichotomous items); each subject's score is the sum of scores on individual items. Guttman (1950) suggested modifying Likert's approach to allow some response categories to have greater weight in a subject's scale score; the Guttman method is to select weights for the categories that maximize internal consistency. Although both the Likert and Guttman approaches assume that all items measure the same phenomenon, neither includes a statistical test to decide the degree to which this assumption holds. Guttman (1944) also invented scalogram analysis, the major goal of which is to determine the extent to which a set of items constitutes a pure scale (as defined by Guttman). His index of reproducibility, however, is drastically affected by the distributions of items, as shown by Festinger (1947), and is sometimes large even when items are relatively unrelated.

A general mathematical model for latent structure was first provided by Lazarsfeld (1950). He related the probability of responding in each category of each item (and the joint probability associated with the response pattern for all items) to an underlying latent variable. This approach, which Lazarsfeld called latent structure analysis, is quite general, and can isolate such response tendencies as the acquiescent response set (the tendency to agree with items) and the extreme response set (the tendency to ignore extreme categories when these number five or more). Unfortunately, the problem of developing an efficient estimation algorithm to use with the general latent structure model has until recently prevented wide application of Lazarsfeld's work.

A breakthrough came in 1974 with publication by Goodman (1974a, 1974b) of two papers which show how latent structure models can be incorporated into the general framework of the log-linear model (Goodman 1978, Part 4). Goodman explored a wide range of causal models which permit both unobserved causative factors and imperfectly measured outcomes. These models can be used to analyze data or to construct tests and indexes for measurement and prediction. To illustrate the wide applicability and flexibility of his models, Goodman (1975) applied them to several data sets previously analyzed by other researchers, with strikingly different results. His basic insight is that Guttman's scaling model can be formulated as a special case of a new latent class model, with appropriate maximum likelihood and chi-square goodness-of-fit statistics; these enable the identification of an additional class of subjects that are "intrinsically unscalable" in data that do not fit Guttman's model. They further provide a corresponding increase in goodness-of-fit.

Goodman considers a wide range of latent-structure measurement models including both identifiable and unidentifiable ones. For each he presents simple methods for determining whether their parameters are identifiable. When they are not, he describes restrictions that might be imposed (if reasonable) to identify the parameters. Goodman's class of latent structure measurement models is somewhat analogous to the quantitative factor analytic models and models hypothesizing linear structural relationships (Jöreskog 1969, 1971; Jöreskog and Sörbom 1977).

LIKELIHOOD INFERENCE

Throughout the development of path-analytic applications in social science (see the collection of papers in Blalock 1971), attention has centered on the logic of the correct choice of and relationship between variables, and on whether the data agree with the model. Few attempts have been made to look at these questions from the opposite direction, that is, to determine whether the parameters chosen to measure the relationships are the best, or if some other perhaps contradictory model is also supported by the data.

In the past two decades, with the rapid development of the modern computer and computing technologies (both hardward and software), statistics have advanced rapidly on several fronts. The two areas of particular interest to evaluation researchers are (1) improvements in the model describing the data, and (2) developments of techniques for drawing inferences from the data given the model. The first text summarizing recent developments in these two areas for social scientists is that of Lindsey (1973).

The traditional approach to data analysis in social science, centered on least-squares regression models, is founded on the assumption that data follow a normal distribution; this is primarily because it was essential to keep statistical inferences simple in precomputer times. One major advance, for this traditional regression approach, was the estimation of transformations which make the data follow more closely a normal distribution (Box and Cox 1964). The more modern alternative, instead of making data fit the model, replaces the normal distribution with distributions better describing the data. For the dichotomous variables that are often included in multiple regression (treatment-control, pretest-posttest, failure-success, and so on), see Cox (1970), in which are described logistic models; this logistic formulation is described in Lindsey (1973, chapter 3), with extensions to cover the polychotomous case.

In the second area, that of drawing statistical inferences from data, given a model, three main schools of thought continue to divide statisticians. These are the classical or Neyman-Pearson school, associated with confidence intervals and hypothesis testing; the Bayesian school, associated with prior probabilities; and the Fisherian school, associated with inference using likelihood functions. Of these, it is the Fisherian approach, originally proposed in the 1920's (see Fisher 1956), that has enjoyed a renaissance since the computer has made feasible its application to the complications of social science and evaluation models.

Fisher's likelihood approach in comparison to the classical and Bayesian approaches--and insofar as social science and evaluation research are concerned--has three main advantages: (1) all of the information in the data about the model is exploited, (2) an absolute minimum of restrictive assumptions is introduced; and (3) the data analyst may proceed from this

approach to either of the others (but not necessarily in the reverse direction). A paper describing likelihood inference in a way most accessible to the evaluation researcher is Sprott and Kalbfleisch (1965); see also the textbooks by Kalbfleisch (1971) and Edwards (1972).

What do these improvements in the models describing data, and in the techniques for drawing inferences from data given a model, mean for evaluation research? The use of improved models means that social processes under study can be described more accurately and that better insight will be provided into how these processes operate. The loss of information in dichotomous data can be avoided, for example. The use of better inference procedures means that the maximum amount of accurate information about a model can be drawn from a given data set. (Determining whether an effect is large or small may often be meaningless, because the data can provide an estimate of a large effect but also indicate that the effect could almost as plausibly not exist, or conversely that a small effect is implausibly zero.)

Even when such analysis requires prohibitively many calculations, logistic models might still be used because the amount of computation is usually about the same as for path analysis using linear regression. In addition, all of the information about effects in the data is used. Methods of making approximate likelihood inferences with a minimum of calculations are described in Lindsey (1973, chapter 3).

NONPARAMETRIC STATISTICS

Nonparametric statistics are sometimes called "distribution free" statistics; actually both terms are misleading. Nonparametric statistical tests do not involve distributions that have no parameters, nor can a population be "distribution free." Both terms refer to a large category of tests which do not require the normality assumption or any other assumption that specifies the exact form of the population. Some assumptions about the nature of the population are required in all nonparametric tests, however, but these assumptions are generally weaker and less restrictive than those required in parametric tests.

The place of nonparametric statistics in social and policy sciences was established by Siegel (1956). A good many of the nonparametric procedures can be easily understood by those with little mathematical training. More recent procedures have been summarized in texts by Bradley (1968) and Pierce (1970); for a more mathematical treatment, see Lehmann (1975). Savage (1953) has compiled an extensive bibliography of nonparametric methods.

What are the advantages of nonparametric tests as compared to tests such as the t-test for the difference of means? In using the t-test, it is necessary to assume an interval scale and a normal population distribution (the latter assumption can often be relaxed). Therefore nonparametric statistics will be most useful whenever either of the two classical assumptions is not met, that is, when it is legitimately impossible to assume an interval scale, or when the sample is small and normality cannot be assumed. Because nonparametric tests involve weaker assumptions than the t-test for the difference in means, they may not take advantage of all available information. If any interval scale can be approximated from the ordering of scores, for example, and if the normality assumption can either be approximated or relaxed--in the case of larger samples--the t-test will ordinarily be preferred to one of the nonparametric alternatives (Blalock 1972).

In more ambiguous cases, the choice between classical and nonparametric tests amounts to a choice between the greater power of the former and the weaker assumptions of the latter. Weaker assumptions are desirable because, if results of a test call for rejection, the null hypothesis is more readily argued to be the single faulty assumption. Unfortunately, tests that require stronger assumptions are usually more powerful in the sense that their use involves a lower risk of Type II error. For this reason, strength of assumptions and degree of power work in opposite directions, and the choice between classical and nonparametric statistical procedures must always be evaluated accordingly. Here, as elsewhere, the choice between power on the one hand and alpha-level stability on the other is dictated by the overall goals of the evaluation effort rather than by purely statistical considerations.

For comparing the locations of two and k independent groups, the Wilcoxon (1945) and Kruskal-Wallis (1952) tests respectively are efficient procedures. Wilcoxon's test requires two assumptions: that the samples have been randomly and independently drawn, and that they come from continuous population distributions. As Walsh (1965) notes, when ranks are assigned at random to tied scores, continuity is imposed. There is considerable evidence (Boneau 1962; Bradley 1968) that Wilcoxon's test is highly efficient, which means that it can be used whenever the assumptions of the t-test are questionable. The Kruskal-Wallis test is a simple extension of the Wilcoxon test to k independent groups, with the same assumptions as the latter and similar strong evidence for high efficiency (Bradley 1968). For making pairwise comparisons of the k independent groups following a significant Kruskal-Wallis test, the Steel (1960, 1961) test is recommended. It entails multiple Wilcoxon tests but with only a single critical test value; tables are available which make it unnecessary to calculate even the single critical value (Steel's original exact tables have been extended by Miller 1966).

For analysis of matched groups, the Sign Test and Friedman's (1937) test are suggested for two and k independent groups, respectively. The Sign Test has a number of advantages: it is completely distribution free, unlike the Wilcoxon (1949), and it has great simplicity; furthermore, it has been extended to permit pairwise comparisons of k matched groups (Miller 1966; Rhyne and Steel 1967), and it is highly efficient (Bradley 1968).

Friedman's test assumes no interactions between treatments and blocks (Miller 1966); its efficiency is still not well known. For making pairwise comparisons of the k independent groups following a significant Friedman test, the Multiple-Sign test (Miller 1966) is recommended (for detailed tables, see Rhyne and Steel 1967). It entails a minimum of assumptions and, unlike the Nemenyi (1963) alternative, significance of a comparison of two populations never depends on other populations within the block. Little is known about the Multiple-Sign test's power and efficiency.

In summary, the Wilcoxon, Kruskal-Wallis, Steel, Friedman, Sign, and Multiple-Sign tests are eminently suited for statistical analysis whenever the data clearly violate the classical assumptions of an interval scale, normality, homogeneity of variance, and so on. It does not follow, however, that the nonparametric methods, being more assumption free, should be used exclusively. Nor should the inference be drawn that poorly designed experiments can be salvaged using nonparametric statistics.

MODELS UNDERLYING NONPARAMETRIC TESTS

One criterion for choosing a statistical test is that it require the weakest possible assumptions for its validity. This feature has been largely responsible for the recent popularity of nonparametric methods (and the one to which they owe their name). Although nonparametric methods have been restricted mostly to testing procedures, they have gradually come to include point and interval estimates as well as various simultaneous inference procedures. However, they still do not have the kind of flexibility and applicability to complex linear models that make least squares and normal theory so widely entrenched.

In order to use nonparametric procedures in an evaluation study, it is not necessary to have a population from which the units in the study have been obtained by random sampling. It is, however, necessary that the treatments being compared have been assigned to the units at random; this is the randomization model.

The randomization model assumes that the units of analysis which are available for observation in a particular study are not chosen but are given, and that they are assigned either to treatment or to control groups at random. Chance enters this model only through the assignment of units to treatment or control.

Unfortunately, because of the simplicity of their randomization assumptions, randomization models do not permit determination of sample sizes and evaluation of the power of tests (which plays an analogous role as the variance of an estimate), two factors which could be used to design a study. Such factors are best discussed in terms of the population from which the units in the study have been sampled; this is the population model.

The population model assumes that the units of analysis in a particular study are drawn as a simple random sample from the population of potential beneficiaries of the treatment(s) to be evaluated, and are then assigned either to treatment or to control groups at random. Chance enters this model, not only through the assignment of units to treatment or control, but also (in a way that can be taken into account) in the selection of the units.

There are situations in which neither randomization nor deliberate sampling are possible (nor are the randomization or population models appropriate). For example, consider an evaluation study of two or more instruments or methods, each making a number of independent determinations of some quantity (distance, speed, temperature, and the like). Here, there is no possibility of random assignment, nor is the population model applicable because the measurements cannot be obtained as purposeful random samples from a population of such measurements. The various sets of measurements can often be assumed to be independent, however, with each set having a common distribution. Then it is possible to test the hypothesis of no difference between the distributions of various sets of measurements; that is the measurement model.

The measurement model assumes that independent sets of measurements are obtained from as many instruments or by as many methods as possible. Chance enters this model through fluctuations in the conditions (such as during school exam periods or two days before a vacation) under which the measurements are taken, fluctuations in the quantity being measured, and possibly physiological and psychological fluctuations in the observer.

NEW APPROACHES TO MULTIVARIATE REGRESSION

Regression has two distinct statistical meanings: column (local) averages, that is, typical values of y for fixed or nearly fixed (and hence local) values of x; and fitting a function or, more humbly, choosing from all possible fits (distinguished by different constants). In the classical style of regression analysis--what John Tukey has labeled the "over-utopian mode"--data were bent to fit assumptions convenient for both mathematical theory and computational practice. These assumptions include most of those commonly found in statistics texts, such as randomness, independency, and normal distribution (for examples, see Draper and Smith 1966). The emerging style of data analysis--what is often called "exploratory data analysis" or "EDA"--reverses the older relationship between data and assumptions. In the new exploratory mode, the object is to make a data set suggest its own analysis. Particular exploratory techniques have, with a few exceptions, resisted institutionalization, so that "exploration" remains conscientiously artifactual: new techniques can be invented for each new data set, and no one technique exhausts any particular problem.

This change in the way data are analyzed, a situation which is likely to continue for some years, demands that evaluation researchers concentrate less on the learning of particular techniques than on the development of flexible philosophies and styles of analysis. These involve at least the following six tenets: (1) the behavior of data, under techniques like guided regression and reexpression, can suggest the data's own analysis; (2) ad hoc indicators and scaling can be more informative than answers to a priori questions; (3) residuals can contain the most important information in a regression analysis; (4) displays and graphs are invaluable for revealing what one could not have expected to see in advance; (5) thanks to modern computer capabilities, iteration can replace the "once-through" calculation of many standard analytic techniques; and (6) data analysis ought to proceed through creative invention and trial and error, with no one technique exhausting any given analysis.

In other words, this new perspective on regression, as codified by Mosteller and Tukey (1977), depends much less on institutionalized techniques, which must be learned and routinely applied, and general use packaged computer programs. It depends much more on the resolve to keep mathematical and computational conveniences subordinate to the understanding of one's data.

Seven of the 16 chapters in the Mosteller and Tukey text are particularly useful for regression in general and for evaluation research in particular. That text's chapter 4 introduces reexpression, the straightening of curves and scatter plots in particular, which provides background for simple linear regression. Chapter 10 introduces a new approach,

grounded in robust and resistant techniques, to certain simple applications (like the measure of location and spread). Chapter 12 offers a new perspective on the meaning and purpose of regression analysis. Regression coefficients and fitting, the latter exploiting iterative techniques to make robust and resistant fits, are discussed in chapters 13 and 14. Chapter 15 discusses the possibility of multiple fits to a given set of data, while chapter 16 tells how to explore residuals for additional fits.

Among the most important problems associated with the above themes are the following: collinearity, measurement error, substantive meaning of regression coefficients, so-called "proxy" variables, alternative weighting of least squares, model selection, and appraisal of multiple variables.

The purpose of regression is to simplify the world, usually in one of five ways: by summarizing data, controlling distracting noise, finding causes, measuring causes, and predicting the future. What is good regression for one purpose may be bad regression for another. To control distracting noises, for example, there are better approaches than the one used for predicting; specifically, one would want to use structural coefficients which are larger than the ones used for prediction. Because evaluation research is necessarily concerned with decision making, there is increased pressure for regressions to be used for causal interpretations. Given this fact, evaluation researchers ought to take warning from the Mosteller-Tukey view of regression, that "every regression is an incomplete description," because some kinds of incompleteness are more damaging to decision making than are others.

Regression lines can be fit by stages, with the effect of each independent variable removed, in turn, from both the dependent variable and from all other independent variables not already fitted. What is a good fit?--especially when choosing among alternative reexpressions, possibly of both x and y? Consistency of description of parallel sets of data (analyzing parallel cases in parallel ways) is the best criterion. This is usually equivalent to another criterion: approximate constancy in the size of residuals (constant spread of the data around the fitted line).

MULTICOLLINEARITY

When two or more independent variables are highly intercorrelated, then it is difficult and perhaps impossible to assess their independent effects on the dependent variable. As the correlation between two independent variables approaches unity, it becomes impossible to tell one variable from the other. This difficulty, called multicollinearity, not only affects the estimates of partial slopes and partial correlations in multiple regression procedures, but it also similarly weakens inferences based on cross tabulations (Johnston 1963; Blalock 1963; Farrar and Glauber 1967). Although it sometimes happens that the use of additional information may alleviate the problem of multicollinearity, it often happens that when an evaluation researcher must rely on so-called "natural experiments" it will be impossible to obtain the independent variation necessary to assess the independent effects of explanatory variables.

Multicollinearity has a number of practical consequences. Some theories that assert the importance of one variable over another, or policy decisions between two or more alternative interventions, while theoretically testable, are actually incapable of being tested if the independent variables are highly interecorrelated. No statistical method known or likely to be developed will break this "multicollinearity deadlock" (Johnston 1963). An additional danger of multicollinearity is that the regression analysis in most cases can be done as usual, with the estimates generated by collinear variables subject to large instabilities. (Farrar and Glauber (1967) discuss some modest palliatives.) Tufte (1970) lists three signs that can alert the analyst to the presence of multicollinearity: (1) high correlations among describing variables, (2) a sizable multiple correlation for the overall regression but with no particular regression coefficient reaching significance, and (3) large changes in the values of the regression coefficients when new variables are added to the regression. In some cases with many variables, collinearity can still present a problem even if there are only modest correlations among the variables.

AUTOCORRELATION, OR SERIAL CORRELATION

In evaluation research, investigators are often interested in answering questions relating to changes which occur over time. Traditional parametric approaches include using multiple regression and multiple analysis of variance to examine differences between groups. These models are based on an assumption of independence between error terms of different measures in the model. In time series designs, measures are often repeated. Thus, the error term from one measure may affect later measurement. Within data which occur naturally over time, there is a tendency for recent information to be more similar to the observation which just occurred than earlier observations. This phenomenon is called positive autocorrelation. The effect of violating the assumption of independence is to reduce the variance estimate for the model and increase the likelihood of finding a statistically significant result.

If the time relationships the evaluator is concerned with occur among the independent or exogenous variables, the problem of autocorrelation may be solved by using lag transformations. When many past observations are used, however, the investigator must again become concerned with multicollinearity. Therefore, theoretical constraints on the relationship among the lag functions should be developed. If the investigator is interested in using previous measures of the dependent variable as a predictor of present scores, the lag function provides an unbiased estimate of the beta's; however, it underestimates the variance, requiring special estimation methods. This may be done using two-stage, least-squares regression or a variety of specially designed time series computer programs (for example, the Box-Jenkins 1976 technique). This is another case where plotting the residual against time will help prevent invalid conclusions or an incorrect choice of a theoretical model. (For further discussion, see Ostrom 1978 and Cook and Campbell 1979.)

STRUCTURAL EQUATION MODELS (PATH ANALYSIS)

Structural equation models might be expected to play an important role in the future of evaluation research, if only as a post hoc, multivariate alternative to designed experiments. A large body of useful experience in the application of regression methods to substantive problems has been built up by econometricians (Goldberger 1964; Malinvaud 1970; Kmenta 1971; Theil 1971; Johnston 1972); more accessible presentations of simultaneous equation models are available in Wonnacott and Wonnacott (1970) and in Wallis (1973).

Path analysis, pioneered by the geneticist Sewall Wright (1921), offers a graphic presentation of the causal interrelationships among variables. It is based on regression analysis, and can provide a more useful picture of relationships among several variables than is possible through other means. Because it assumes that the values on one variable are caused by the values on another, it is essential that dependent, independent, and exogenous variables be distinguished in path analysis.

In several senses, path analysis is an extension of the ideas behind the multivariate elaborate model developed by Paul Lazarsfeld and his colleagues (Lazarsfeld and Rosenberg 1955; Rosenberg 1968). For a concise and elementary introduction, see Babbie (1975, chapter 17). In the elaboration model, partial tables are used to determine the effect of control variables on the initially observed association between two other variables. Path analysis accomplishes the same thing through the use of standardized regression coefficients arranged in logical schematic diagram; the basic logic is much the same.

The most accessible introductions to structural equation models in the social and policy sciences are Duncan (1975) and Heise (1975). Informative articles using structural equation models and path-analytic techniques have been collected by Blalock (1971) and Goldberger and Duncan (1973). A well known and highly controversial policy application of path analysis is that of Jencks et al. (1972).

CAUSAL MODELING OF QUALITATIVE VARIABLES

Simple recursive causal models, in which causal linkages run in only one direction (Duncan 1976), can be developed using simple logit models (Goodman 1972). For more complex recursive systems, in which explanatory variables are nested (that is, one variable or a set of variables is causally prior to another variable or set, which is prior to still others, and so on), any number of nested logit models might be employed in their analysis. Goodman (1973) presents methods for generalizing the analysis of simple logit models to more complex systems, both recursive and nonrecursive, of qualitative variables. For each hypothesized system he presents methods for (1) estimating the parameters in the system, (2) testing whether the hypothesized system fits the data, and (3) partitioning the chi-square test statistic into components that can be used to test individual logit models or other subsystems. This work develops and extends some of the concepts and methods introduced in Goodman (1973).

COMPUTING RESOURCES FOR STATISTICAL ANALYSIS

INTRODUCTION

The remainder of this chapter focuses on computing resources which are available to perform the types of statistical analyses discussed above. The following subjects are adressed: (1) a general overview of the field and the types of programs available; (2) a discussion of criteria for evaluating statistical programs; (3) an evaluative discussion of the BMDP, SAS, and SPSS general purpose programs, which are very widely used; (4) a brief index of special purpose statistical programs which are available; and (5) a limited discussion of other available computing tools, including interactive programs, subroutine libraries, data base management systems, and graphics software.

OVERVIEW OF COMPUTING RESOURCES FOR DATA ANALYSIS

THE AVAILABILITY OF COMPUTING RESOURCES: ADVANCES IN HARDWARE AND SOFTWARE

Computing tools to carry out the types of analyses described earlier in this chapter are now widely accessible, relatively inexpensive, and easy to use.[1] The manufacture and design of the hardware component of computing (that is, the electronic computer and its peripherals) have undergone a technological revolution in the 1970's.[2] Dramatic innovations, largely in electronic circuitry, have led to smaller, faster, and less expensive computers. Researchers are now able to purchase computers, which had been prohibitively expensive, for their own agencies and departments. Also, computer time from outside computing facilities can be purchased at lower rates and with greater ease.

During the 1970's, the software components (the programs which control the computers' computations and operations) have undergone significant, though less dramatic changes. Fifteen years ago, a researcher would typically need a working knowledge of the hieroglyphics of a sophisticated programing language (such as FORTRAN) in order to perform data analyses using a computer. Since then, we have seen a proliferation of statistical programs utilizing well structured and well documented, English-like mnemonics. These programs have significantly decreased the level of programing skills and the amount of learning time required to perform sophisticated data analysis.

One result of this change in computer software is an increase in the number of evaluators, previously unfamiliar with computing, who are now responsible for computing tasks and consequently need to make technical choices regarding software resources. The remainder of this chapter is written to assist this new class of computer user by providing a brief overview of the kinds of computing resources currently available.

CONTINUED

2 OF 3

TYPES OF DATA ANALYSIS SOFTWARE: SOME BASIC CATEGORIES

Programs for data analysis are commonly divided into one of three types: the general purpose program, the special purpose program, and the subroutine library. (Note: in all cases the term "program" will be used in order to avoid the need to differentiate between "program," "package," and "system.") Further, general and special purpose programs are commonly distinguished by two types of processing modes: batch or interactive. The following is a brief discussion of each of these categories.

The general purpose program is typically characterized by the following traits. First, and most importantly, it is general in orientation, that is, it contains a wide array of statistical procedures (defined here as encompassing descriptive, analytical, and display procedures). Such a program generally contains the most commonly used techniques, such as tabulation, correlation, regression, analysis of variance, factor analysis, discriminant analysis, nonparametric statistics, and plotting capabilities. Further, it contains a useful variety of data management tools, including flexible file handling, and data modification/selection procedures. Second, the general purpose program is consistent, that is, all its procedures use a common control "language"--the means by which the researcher instructs the program to perform the task at hand. Usually, this control language uses an easy to understand syntax with English-like mnemonics. Third, the general purpose program is usually integrated, in that its procedures have common data management and data analysis routines. Several general purpose programs will be described later in this chapter, since to the extent that they have the necessary capabilities, they are usually the most powerful tool, especially for the non-programer.

The special purpose program performs one type of statistical procedure, such as time series analysis or cluster analysis. Typically, the special purpose program provides few data management tools. Often, though not necessarily, it is more difficult to use than the general purpose program. Some of the special purpose programs which perform statistical analyses not found in the general purpose programs will be listed later.

The subroutine library is a collection of program modules which are "connected" and "called" by a programing language, such as FORTRAN. An example is a set of subroutines which perform a variety of matrix operations. While useful in cases where a task cannot otherwise be performed in a general or special purpose program, they usually require a great deal more expertise to use.

A batch program processes one set of instructions or control specifications in a "batch" (for example, from cards), and returns the results in a batch (for example, printed output from a line printer). The majority of this section focuses on batch-type programs, since they are the most commonly available.

An interactive program processes a small set of instructions from a time-sharing terminal, and returns the results when requested to that terminal. The researcher can, therefore, perform a set of hierarchically dependent tasks in one "sitting" at a timesharing terminal. Put another way, the researcher can carry on a continuous interrogative "dialogue" with the data via the interactive program. Interactive processing can be a very powerful computing research method. Progress in the area of developing such programs will be mentioned later.

CRITERIA FOR EVALUATING STATISTICAL PROGRAMS

For any given task, the researcher usually has access to several, and often many, programs which have the required capabilities. The choice, therefore, should focus on a number of program characteristics in addition to the specific capabilities needed at the moment. The choice of program can be important, especially for the researcher who intends to do a great deal of computing. The following are criteria which can be used in selecting programming tools. Note that this is a cursory treatment of a complex procedure. The reader is referred for greater detail to the growing body of literature on the evaluation of statistical programs. See Francis et al. (1974); Heiberger (1975 a and b, 1976 a and b);

Kohm and Thomas (1975); Velleman and Welsh (1975); Francis and Heiberger (1975); Forsythe and Hill (1975); Francis and Valliant (1975); Gentle (1975); Thisted (1976); Blashfield (1976); Hardy et al. (1975, 1977); Velleman et al. (1977); Allen and Velleman (1977); Sours (1976); Berk and Francis (1978); Muller (1978); Allen (1976); Ketola (1978); and JASA (1978).

Range of capabilities. It is important to examine the range of statistical capabilities offered by each program under consideration. Choosing a program with a wide range of capabilities may avoid the necessity of learning additional and unfamiliar programs. Further, the range of data management capabilities should be examined. Quite often, the flexibility of a program in handling files, and performing data modification/selection tasks is as important as the statistical capabilities in the overall value of that program.

Accuracy. Any program of significant complexity is prone to error. The degree to which a program is statistically "correct" and numerically stable is of obvious importance.

Efficiency. Some programs consume significantly more computing dollars than others to perform similar tasks. While this is an important criterion, personnel costs should also be considered. Quite often the costs for programers and analysts far outweigh computing costs. In such cases, less efficient, but easier to use programs may well be more cost efficient overall.

Ease of setting up the control language. The degree of difficulty in setting up a program to perform a given task is an important criterion, but is difficult to evaluate. First, it may be a matter of personal preference or style; second, it may be dependent on the nature of the task at hand. For example, a program utilizing flexible grammar and syntax may be quite useful for complex tasks, but may require an undue amount of time to learn for simpler tasks. In general, the following program characteristics will simplify setting up the control cards: (1) English-like mnemonics; (2) free format; (3) consistent and easy to understand syntax; (4) effective checks on errors with informative diagnostics (see below); and (5) good documentation (see below).

Documentation. Two studies (Francis and Valliant 1975; Gentle 1975) suggest that documentation is of primary importance in the effective use of a computer program for both beginning and experienced users. In evaluating the quality of a program's documentation, three aspects are important: a program's tutorial style, which is especially important to the beginner; its usefulness as a reference document, which is important to the experienced user; and its documentation of the statistical formulae and computing algorithms.

Output. The labeling, format, and the extent to which the user can control the form the output takes can help determine the usefulness of a program.

Style of diagnostic reporting. Reporting an error (that is, a control card specification error or a numerical problem) in a clear and prescriptive form is a significant asset to a program user. (Unfortunately, in practice, clear and helpful diagnostic messages are seldom found.)

Extent of program use. A widely used program offers several advantages: researchers will be likely to find it available at the various computing installations they might use; consulting help will be more readily available, since others will be more likely to be familiar with the program; it will be easier to standardize research; and it will be easier to do cooperative research.

How widely a program is used is determined by the preferences of researchers for its capabilities, and the awareness of its availability among potential users. In addition, the extent of use can also be determined by the range of computers and operating systems on which the program can operate--its availability. Some programs are "transportable" or "convertable" to a wide variety of computers and operating systems; others are more limited. (For example, SAS, which will be discussed later, runs only on an IBM 360/370 model under IBM's OS and OS/VS operating systems, and plug-compatible machines.)

Status of the vendor or developer. Programs distributed by a committed, financially sound, and well staffed vendor are more likely to be well maintained, enhanced, and to continue to be widely available and used.

Batch and interactive capabilities. The choice of a program based on its batch or interactive capabilities is usually determined by the researcher's personal preferences and by the nature of the research project.

GENERAL PURPOSE PROGRAMS

The general purpose program will quite often be preferable to a special purpose program which performs the same type of statistical operation, for reasons mentioned previously. The researcher can perform a wide variety of statistical operations without repeatedly being required to learn new and unfamiliar programs. In general purpose programs the syntax, grammar, input-output (I/O), and data handling are usually common to all operations. Such programs usually offer a wider array of data management tools. They are often, though not always, well documented and well maintained. Several general purpose programs are widely available and extensively used. For these reasons, more attention is given here to a discussion of general purpose programs than to other types.

Numerous general purpose programs were developed in the late 60's and 70's. For the most part, they were developed in the academic community. The following is a partial list of programs and their origins:

BMD and BMDP	University of California, Los Angeles
DATA-TEXT	Harvard University
MIDAS	University of Michigan
MINITAB	The Pennsylvania State University
OMNITAB	National Bureau of Standards
OSIRIS	University of Michigan
P-STAT	Princeton University
SAS	North Carolina State University
SCSS	University of Chicago
SOUPAC	University of Illinois, Champaign-Urbana
SPSS	Stanford University and the University of Chicago
STATJOB	University of Wisconsin, Madison

The general purpose programs listed, as well as others, are currently available for use. However, during the mid-70's usage concentrated around a handful of such programs. Concurrently, some of the developers have commercialized their operations, and are becoming increasingly competitive.

Three of the major competitors on the marker are BMDP (Biomedical Package), SAS (Statistical Analysis System), and SPSS (Statistical Package for the Social Sciences). The following paragraphs evaluate these three programs. The discussion is limited to these three because they generally rank the highest in terms of ease of use, range of capabilities, accuracy, documentation, and extent of use. Further, and perhaps most importantly, they are backed by well funded developers committed to their growth and success. BMDP, SAS, and SPSS are under continual development and are likely to dominate the market for some time. Other packages may be as or more useful for some types of applications, and they should not be ignored. However, the extensive use and the growth orientation of these

three packages, coupled with their wide range of capabilities, make them prime candidates for the kinds of research under consideration--research which is ongoing, and which envisions cooperation and replicability as goals.

The remainder of this section will compare BMDP, SAS, and SPSS relative to six of the criteria presented earlier--the six that best distinguish these programs. The attempt is to provide perspective, not to offer recommendations on which program to use.

Capabilities. Table 2 presents the capabilities of BMDP, SAS, and SPSS. The information presented therein is qualified by the following: (1) the list of features is not exhaustive, but includes those of major interest; (2) in cases where a feature is part of the repertoire of all three programs, one must not assume that these programs are "equal" in that respect; (3) these programs are constantly being altered and improved, and this table will quickly become obsolete; and (4) information cited was taken from the manuals referenced (Brown 1977; Helwig and Council 1979; Hull and Nie 1979; Nie et al. 1975; Norusis 1979).

Accuracy. It is beyond the scope of this work to do the extensive primary research required to evaluate the statistical and numerical accuracy of these programs. However, two comments may be helpful. First, with respect to accuracy, SAS and BMDP are generally well regarded by the statistical community (see, for example, an evaluation of regression procedures, Velleman et al. 1977.) SSPS has not been so highly rated. Second, by way of qualification, the SPSS programing group (SPSS, Inc.) has been sensitive to the criticism that its product has encountered in the statistical community, and has taken steps to correct the problems (Nie 1978). For example, they have hired statisticians to examine the code for accuracy and publish an algorithms document. Researchers should, therefore, take into account the fact that criticisms of the accuracy and reliability of SPSS may relate to SPSS versions of the past, rather than to the current and future versions.

Efficiency. Any overall ranking of the computing efficiency of these programs is difficult since efficiency is dependent on many variables, such as: the procedure that is being compared; the structure of the data set; the computer on which the program is run[3]; and the version of the program being used. Not surprisingly, therefore, several studies show somewhat mixed results[4].

The main conclusion to be drawn here is that the programs do show significant differences in efficiency, depending on the factors noted. Therefore, if computer efficiency is a significant factor in the choice of a program, it is recommended that timing tests be made on the computer that is to be used using the particular kinds of computing tasks to be performed. (Again, other programs characteristics which affect personnels costs may be more important in the overall cost of using a program.)

Ease of control card set up. As stated earlier, the ease of setting up BMDP, SAS, and SPSS runs is dependent on the personal preference of the user, and on the nature of the research task at hand. In general, all three packages are "user friendly"--it is easy to specify the programing task. SAS is regarded by many as having the most powerful control "language," which is a distinct advantage in some programing tasks. SPSS is often regarded as being the easiest for the novice to use. (Also, current work plans at SPSS, Inc. include enhancements to make the transformation language more powerful, and the procedure specification language less restrictive.) BMDP has been regarded as the most difficult to use. However, with the advent of the "P" series of BMDP, which utilizes an English-like sentence-paragraph structure, and with the upgrading of its data transformation procedures in the latest version, BMDP has become much easier to use. In short, while each package will have certain advantages for any one particular task, it is expected that they will become more alike in the future with respect to the criterion of ease of programing than they have been in the past. (However, based on SAS's current capabilities and those planned for SPSS, it is expected that SAS and SPSS will have the most flexible and powerful language capabilities. This should be qualified by noting that BMDP's language is sufficient for most purposes.)

Documentation. SPSS is regarded as a good tutorial document for the beginning computer user. The overview of the system, the discussion of the syntax, and the presentation of data management and statistical procedures is extensive. (However, SPSS is oriented

Table 2. A Comparison of the Capabilities of BMDP, SAS and SPSS

	BMDP	SAS	SPSS
Data Management			
1. Input data types			
a. Case by variable	x	x	x
b. Hierarchical records		x	
c. Variable length records		x	
d. Matrices	x	x	x
2. Data definition			
a Variable description			
(1) Variable names	x	x	x
(2) Extended variable labels		x	x
(3) Value labels	x	x	x
(4) Missing values	x	x	x
(5) Valid ranges	x		
b. Position in input record			
(1) Character position		x	x
(2) FORTRAN format	x	x	x
(3) Freefield		x	x
3. Data transformation within cases			
a. Arithmetic recode			x
b. Character to numeric convert		x	x
c. Arithmetic operators	x	x	x
d. Arithmetic functions	x	x	x
e. Conditional statements (e.g., IF, THEN)	x	x	x
f. Branching statements (e.g., GO TO, LOOPS)		x	1
4. Data transformation across cases			
a. Sort		x	x
b. Rank		x	
c. Standardize	x	x	x
d. Aggregate			x
e. Random sample	x	x	x
f. Selective sample	x	x	x
5. Interactive data editing		x	
6. File handling			
a. System files			
(1) Save and retrieve	x	x	x
(2) Create and retrieve subfiles			x
(3) Save and retrieve matrices	x		
(4) Add variables		x	x
(5) Add cases		x	x
(6) Merge		x	x
b. Other files			
(1) Output matrices		x	x
(2) Output cell frequencies			x
(3) Other output capabilities		x	x
(4) Input system files from other two packages		x	
Statistical Analysis			
1. Univariate descriptive measures			
a. Mean, standard deviation, variance, skewness, kurtosis	x	x	x
b. Median Mode	x	x	x
c. Maximum, minimum, range	x	x	x
d. Frequency distributions	x	x	x
e. Robust location estimates	x		
2. Tabulation of data			
a. Single classification	x	x	-x
b. N-way classification	x	x	x
c. Multiple response			x
d. Univariate description cross classified variable	x	x	x
e. Listing of data	x	x	x
f. Aggregation of data			x
g. Report writing		x	x
3. Display of data			
a. Histograms	x	x	2
b. Bar charts		x	
c. Scatter plots	x	x	x
d. Contour plots		x	

4. Contingency table analysis (Categorical data analyses)	BMDP	SAS	SPSS
a. Table statistics			
(1) Observed frequencies	x	x	x
(1) Expected frequencies	x	x	
(3) Row, column, total percentages	x	x	x
(4) Residuals	x	x	x
b. Measures of association/independence			
(1) Pearson product-moment correlation	x	x	x
(2) Spearman rank-order correlation	x	x	x
(3) Other ordered index measures (Tau b, c; Somer's D; Gamma etc.)	x	x	x
(4) Tetrachoric and other 2 × 2 measures	x		x
(5) Chi-square	x	x	x
(6) Quasi Chi-square	x		x
(7) Fisher's and Yate's	x	x	
(8) Empty cells analysis	x		
(9) Stepwise independence analysis	x		
(10) Log Linear fits of hierarchical models	x	3	
5. Analysis of variance			
a. Dimensions			
(1) One-way	x	x	x
(2) N-way	x	x	x
b. Contrasts	x	x	x
c. Posterior comparisons			
(1) Duncan's		x	x
(2) Student-Newman-Kuels			x
(3) Tukey			x
d. Designs			
(1) Unbalanced	x	x	x
(2) Empty cells		x	
(3) Nested	x	x	
(4) Repeated measures	x	4	
(5) Covariates	x	x	x
(6) Random effects	x	x	
(7) Multivariate ANOVA		x	

6. Multiple regression	BMDP	SAS	SPSS
a. Heirarchical	x	x	x
b. Stepwise	x	x	x
c. All possible subsets	x	x	
d. Two-stage least squares		x	
e. Three-stage least squares		x	
f. Nonlinear	x	x	
g. Polynomial	x		
h. Residual plotting	x	x	x
i. Other plots	x	x	
7. Multivariate analysis			
a. Factor analysis	x	x	x
b. Discriminant analysis	x	x	x
c. Canonical correlation	x	x	x
d. Partial correlation	x	x	x
e. Cluster analysis	x	x	
8. Non-parametric analysis			
a. Chi-square goodness of fit	x	x	x
b. Kolmogorov-Smirnov			x
c. Wilcoxan	x		x
d. Runs test			x
e. Kendall's Tau alpha	x		x
f. Concordance	x		x
g. Mann Whitney U	x		x
h. Krushel-Wallis ANOVA	x		x
i. Friedman 2-way ANOVA	x		x
j. Cochran B			x
k. Wald-Woldowirz runs test			x
9. Item Analysis			
a. Reliability estimates			
(1) Cronbach' Alpha			x
(2) Split halves coefficient			x
(3) Guttman's coefficient			x
(4) Maximum likelihood estimate			x
b. Item-to-total correlations			x
c. Tukey test for additivity			x

	BMDP	SAS	SPSS
10. Miscellaneous statistical analysis			
a. Spectral analysis		x	
b. Time-series analysis		x	
c. Probit analysis		x	
d. Life table analysis	x		x
e. Guttman scaling		x	x
f. T-test procedures	x	x	x
g. Matrix algebra procedures		x	
h. Missing data analysis	x		
Miscellaneous Features			
1. Ability to add procedures		x	x
2. FORTRAN code transformations	x		
3. Control available memory	5	x	x
4. Execute procedures from other programs		6	

1 SPSS's DO REPEAT is a rather primitive means of setting up a looping structure.

2 SPSS's histograms are not scaled. They look more like bar charts than histograms.

3 SAS's FUNCAT procedure is more oriented toward testing hypotheses concerning design parameters than testing for independence.

4 SPSS's Reliability procedure will analyze only a repeated measures design with one grouping variable.

5 BMDP requires a FORTRAN subroutine to be partially written and linked into the program through a provided procedure.

6 SAS has a procedure to use a BMDP program under SAS control.

to "hand holding" the novice user, and the manual may be frustrating for the more experienced.) SPSS also provides a separate primer which in some cases is useful for the novice. BMDP does a good job of introducing itself to the beginner, although its prose is more brief than SPSS's and is not written in a "hand holding" style. BMDP is noteworthy for its proliferation of examples of control cards and annotated output. SAS documentation was not designed to be tutorial. A researcher who is inexperienced in computing would probably have a more difficult time getting started with SAS using the manual alone. (It should be noted that the publication of the SAS Introductory Guide in 1978, and the SAS Applications Guide, expected in 1980, may obviate this shortcoming. At this writing, neither manual was available to the writers for review.) For discussions of documentation, see Berk and Francis (1978).

The basic manuals for all three programs serve as useful reference documents for the more experienced user. In addition, all three have reference cards and/or pocket guides.

BMDP and SPSS provide good documentation of the algorithms used. (The lack of such documentation had been a weakness in SPSS, leading to the recent publication of the SPSS Statistical Algorithms Document.) SAS is poor in this area. Its documentation is inconsistent in its treatment of algorithms. Algorithms are sometimes produced in the text; they are sometimes given by general reference to statistics texts; they are sometimes not discussed at all.

SPSS is unique in that it provides expository text on the statistical procedures, giving the researcher an intuitive understanding of the procedure and the motivations for using them. This style of documentation is somewhat controversial, however--some claim that it is quite useful and is in part responsible for SPSS's popularity. Others claim that this text is too simplistic statistically and interferes with the use of the manual as a program document.

Extent of use. SAS is available only for the IBM 360/370 series computer (and plug compatible computers) operating under IBM's OS or OS/VS operating systems, and has been licensed at over 750 installations.[5] SPSS operates on over 25 different types of computing systems, and SPSS, Inc. has about 2,500 licensed installations.[6] BMDP also operates on over 25 types of computer systems. Since BMDP has not been distributed under license, it is difficult to ascertain how many computing installations have received the program; one estimate puts the number at over 3,000.[7] Thus, the likelihood of finding SPSS or BMDP available at any given computer site is higher than for SAS. However, the restriction of SAS to the IBM 360/370-OS-OS/VS systems is not quite as severe as it may seem, because of the popularity of those systems.

It is difficult to say, without conducting a survey, how extensively these programs are actually used at their respective installation sites (that is, how many jobs are run by how many researchers). However, it is the opinion of the writers, based on informal evidence, that at IBM installations SPSS is the most extensively used program, SAS is second, and BMDP third. At other installations, it is believed that SPSS is used quite a bit more than BMDP.

SPECIAL PURPOSE PROGRAMS

A large number of special purpose statistical analysis programs have been developed. Many evolved out of a particular researcher's need to do an analysis for which no program was available. Others have been specifically designed and written for general use.

Computer programs exist to do almost any type of statistical analysis imaginable. Unfortunately, it is not always possible to determine if the resource exists, or how to obtain it if it does. Some work is currently being done which attempts to index statistical programs (Kohm, Ryan and Velleman 1977).

The following is a list of some of the more well-known special purpose statistical analysis programs classified into general topic areas. The list is not exhaustive with regard to either topic or programs listed:

(1) Contingency table analysis (categorical data analysis).

CONTAB Department of Statistics, Florida State University, Tallahassee

CONTAB Department of Statistics, George Washington University

C-TAB International Educational Services, Chicago

ECTA Department of Statistics, University of Chicago

GENCAT Department of Biostatistics, University of Michigan, Ann Arbor

GLIM Numerical Algorithms Group, Oxford, United Kingdom

LOGLIN Harvard School of Public Health, Boston

MULTI-QUAL International Educational Services, Chicago

(2) Item analysis.

LERTAP Department of Education, University of Ontago, New Zealand

LOGOG International Educational Services, Chicago, Illinois

NORMOG International Educational Services, Chicago, Illinois

(3) Cluster analysis.

CLUSTAN David Wishart, University of Edinborough, Scotland

MOCA International Educational Services, Chicago, Illinois

NTSYS Department of Ecology and Evolution, State University of New York, Stoneybrook

(4) Analysis of variance.

MANOVA Clyde Computing Service, Miami, Florida

MULTI-VARIANCE International Educational Services, Chicago, Illinois

RUMMAGE Statistics Department, Brigham Young University, Provo, Utah

(5) Econometric analysis.

ESP Massachusetts Institute of Technology, Boston

Box-Jenkins Time Series Analysis Program Academic Computing Center, University of Wisconsin, Madison

TSP Harvard Institute for Economic Research, Boston

(6) Exploratory data analysis.

EXPAK International Educational Services, Chicago, Illinois

SNAP/IEDA Department of Statistics, Princeton University

(7) Structural equations.

LISREL International Educational Services, Chicago, Illinois.

OTHER COMPUTING TOOLS

SUBROUTINE LIBRARIES

Where the desired capabilities are not otherwise available, subroutine libraries may be helpful. Subroutine libraries can also provide benefits to those researchers who desire to write their own statistical analysis programs. Such programs can usually be written to do an analysis more efficiently (in terms of computing costs) than general or special purpose programs. The programer can control the data input and the printed output of his/her program, and can have access to most of the statistical functions performed by preprogramed routines (that is, the general and special purpose program).

Statistical libraries are normally included along with more general mathematical function libraries, which is beneficial, since a statistical analysis program will usually require both types of routines.

One widely used and highly regarded subroutine library is IMSL.[8] IMSL routines are written to be referenced by FORTRAN programs. APL has a statistical routine associated with it. Also, most hardware vendors distribute mathematical and statistical libraries (for example, IBM Scientific Subroutine Package[9]). In general, these libraries are not regarded to be of the quality of the IMSL routines.[10]

INTERACTIVE PROGRAMS

As in all areas of computing, the trend in statistical computing is toward more use of interactive programs. Interactive programs vary in their abilities to communicate with the user. Some merely wait for the user to tell them what to do and then produce results on the command. Others are more actively involved with the user. For example, a program may give the user help in setting up a command, request additional information, or request the user to correct part of a command that is incorrect. This level of conversation may be primitive at present, but it is an evolving technology which will eventually lead to some very sophisticated conversational programs. Some currently available general purpose interactive programs are MINITAB (Ryan et al. 1979), P-STAT (Buhler and Buhler 1979), and SCSS (Nie and Hull 1979).

OTHER DATA ANALYSIS AIDS

In addition to the use of general or special purpose statistical analysis programs, there are several other computing resources that can be used to analyze data. The following describe some of the major sources of help.

Statistical/Mathematical Programming Aids

Several high level programing languages have been developed which make it relatively easy to program a computer to do mathematical operations. Three widely available and heavily used programing languages are FORTRAN, PL/1, and APL. A fourth, ALGOL, is widely available, but is used minimally in this country. All of these languages have a strong mathematical orientation in their structure.

Graphic Display Programs

A new and rapidly expanding computer resource is software and hardware to perform graphic displays of data. Programs have been developed to present graphic displays on line printers, pen and ink line plotters, and graphics CRT terminals (single and multicolor). These graphics programs can be used to display such things as complex charts, contour line drawings, and two- and three-dimensional perspective plots. Some of these capabilities have been incorporated into special purpose graphics programs (for example, Tell-a-Graph and DISSPLA), making the use of these techniques easier for the statistical analyst. Additional graphic display capabilities are being incorporated into some of the major general purpose statistical programs.

Data Management

Most of the statistical analysis programs discussed here (SAS being an exception) require that the data be structured within a very rigid format. The data must be organized case by variable, with each case having exactly the same number of variables. While it may be necessary to analyze data in this format, it is not always desirable or possible to collect the data this way. For example, data may be obtained from many agencies, each of which organizes it differently; there may be different amounts of data collected at different times for each unit in a particular study (as in clinical studies where each patient has a differing number of visits); or data may have a hierarchical structure (data collected on a particular household where some data pertains to the household as a whole, and other data to the individual members of the household).

A whole field in the computing software industry has developed which creates programs to deal with the storage and retrieval of complex data organizations. These programs are called Data Base Management Systems (DBMS). Recently, some specialized DBMS's have been written which are structured to retrieve data from complex data files and store them directly in a system file compatible for retrieval by one of the general purpose programs.[11]

LOCATING COMPUTING RESOURCES

Normally, the most desirable place to do any computing is on an "inhouse" system. It is probably the fastest and least expensive way to get computing done. However, such systems may not provide the capabilities needed or may not be oriented toward providing timely service. There are usually ways to obtain computing time at a university or governmental computing center. These are more likely to be oriented to providing computer time for doing research rather than administrative work. Commercial service bureaus, many of which are oriented toward providing statistical analysis computing services, also sell computing time.

One problem in finding appropriate computing resources is knowing whether programs are available at that site to do the analyses one needs. Most computing facilities oriented toward providing statistical computing services will maintain a directory of the programs they can provide. (There is usually no problem finding out whether the facility has BMDP, SAS, or SPSS. However, it is not always easy to find the whereabouts of a program like C-TAB.) Occasionally, a consulting service is available to aid both in finding appropriate programs and in using the programs, once found.

Additionally, computing centers may belong to networks which have access to the resources of other computing centers. One such network, EDUNET, is a consortium which links together the resources of computing centers at 22 major universities throughout the country. Membership allows the research to get computer resources which might not be obtainable from other institutions by using TELENET, a nationwide data transmission network.[12]

ENDNOTES

[1]For a layperson-oriented discussion of the industry through the early 1970's, see Nelson, Theodore H., Computer lib. (South Bend, Indiana: 702 S. Michigan Avenue, 1974. (This book/pamphlet is also fun reading.)

[2]For a technical discussion, see the series of articles in the edition of Scientific American devoted to microelectronics. Scientific American, 1977, 237(3).

[3]See, for example, the benchmark tests of similar SPSS runs across several machine types prepared by EDUNET. The results show significant variation in cost and will be presented at the 1979 SIGUCC User Services Conference.

[4]The literature and comments from some researchers in the field suggest that SAS is in general a less efficient program than the other two. This is offered here as a tenative, not definitive conclusion.

[5]Phone conversation with James Goodnight of the SAS Institute, Inc. (July, 1979).

[6]Phone conversation with Wylie Crawford of SPSS, Inc. (July, 1979). Also see: SPSS Newsletter, No. 13, January, 1977.

[7]Phone conversation with Larry Young of the BMDP-77 Computing Group. (July, 1979). Mr. Young also indicated that there were over 500 copies of the latest version, the "77" series, which is now being distributed under license. Also see: BMDP-77 Statistical Software Communications, 10, September, 1978.

[8]IMSL Reference Manual. Houston: IMSL, Inc.

[9]System 360 Scientific Subroutine Package, Version III, Programmer's Manual, Form GH 20-0205. New York: IBM Publications.

[10]The National Algorithms Group, Ltd. (NAG Library), Oxford, United Kingdom, and National Technical Information Services (DATAPAC), Springfield, Virginia, also distribute statistical software libraries.

[11]See, for example, the following:

System 2000 Reference Manual. Austin, Texas: MRI System Corporation, 1976.

OS TOTAL Reference Manual. Cincinnati, Ohio: CINCOM Systems, Inc., 1976.

ADABAS User Reference Guide. McLean, Virginia: Software AG of North American, Inc., 1975.

Robinson, B., Anderson, G., Cohen, E., and Gazdizk, W. SIR Scientific Information Retrieval: User's Manual. Evanston, Illinois: SIR, Inc., 1977.

[12]For more information, write EDUNET, P.O. Box 364, Princeton, New Jersey 08540.

REFERENCES

Abelson, R. P., and Tukey, J. W. Efficient conversion of non-metric information into metric information. Proceedings of the American Statistical Association, Social Statistics Section 2. 1959, pp. 226-230.

Alker, H. R. Jr. A typology of ecological fallacies. In M. Dogan and S. Rokkan (Eds.), Quantitative ecological analysis in the social sciences. Cambridge, Massachusetts: MIT Press, 1969, pp. 69-86.

Allen, I. E., and Welleman, P. F. The handiness of package regression routines. Proceedings of the Statistical Computing Section, American Statistical Association, 1977.

Allen, J. R. Comparison of STATJOB and SPSS and notes on other statistical systems. Working Paper, Madison: Academic Computing Center, University of Wisconsin-Madison, 1976.

American Psychological Association. Standards for education and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anderson, N. Scales and statistics: Parametric and non-parametric, Psychological Bulletin, 1961, 58 (4), 305-316.

Babbie, E. R. Survey research methods. Belmont, California: Wadsworth, 1973.

Beniger, J. R. Sampling social networks: The subgroup approach. Business and Proceedings of the Economic Statistics Section. American Statistical Association, 11. 1976, pp. 226-231.

Beniger, J. R., and Robyn, D. L. Quantitative graphics in statistics: A brief history. American Statistician, 1978, 32(1), 1-11.

Berk, N., and Francis, I. S. A review of the manuals for BMDP and SPSS. Journal of the American Statistical Association, 1978, 73 (361), 65-70.

Birnbaum, A. Classification of ability levels. In F. M. Lord and M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. Discrete multivariate analysis: Theory and practice. Cambridge, Massachusetts: MIT Press, 1975.

Blalock, H. M., Jr. Correlated independent variables: The problem of multi-collinearity. Social Forces, 1963, 62, 233-238.

Blalock, H. M., Jr. Causal inferences in nonexperimental research. Chappel Hill, North Carolina: University of North Carolina Press, 1964.

Blalock, H. M., Jr. (Ed.). Causal models in the social sciences. Chicago: Aldine-Atherton, 1971.

Blalock, H. M., Jr. Social statistics (2nd Ed.). New York: McGraw-Hill, 1972.

Blashfield, R. K. A consumer report on the versatility and user manual of cluster analysis software. Proceedings of the Statistical Computing Section, American Statistical Association, 1976.

Blau, P. M. Structural effects. American Sociological Review, 1960, 25, 178-193.

Bloemena, A. R. Sampling from a graph. Amsterdam: Mathematics Centrum, 1964.

Boneau, C. A. A comparison of the power of the U and t tests. Psychological Review, 1962, 69, 246-256.

Box, G. E., and Cox, D. R. An analysis of transformations. Journal of the Royal Statistical Society, (Series B), 1964, 26, 211-252.

Box, G. E., and Jenkins, G. M. Time series analysis: Forecasting and control. San Francisco: Holden-Day, 1976.

Bradley, J. V. Distribution-free statistical tests. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Brown, M. B., and Dixon, W. (Eds.). BMDP-77, Biomedical Computer Programs, P-series, Los Angeles: University of California Press, 1977.

Buhler, S., and Buhler, R. P-STAT 78 user's manual, Princeton: P-Stat, Inc., 1979.

Campbell, D. T., and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carmer, S. G., and Swanson, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 1973, 68, 66-74.

Cochran, W. G. Sampling techniques. New York: Wiley, 1963.

Cohen, J. Some statistical issues in psychological research. In B. B. Woleman (Ed.), Handbook of clinical psychology. New York: McGraw-Hill, 1965.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Cohen, J., and Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. New York: John Wiley and Sons, 1975.

Cook, T. D., and Campbell, D. T., Quasi-experimentation: Design and analysis for field settings. Chicago: Rand McNally, 1979.

Cox, D. R. The analysis of binary data. London: Methuen, 1970.

Cronbach, L. J. Essentials of psychological testing (2nd Ed.). New York: Harper, 1960.

Cronbach, L. J., and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Davis, J. A. Hierarchical models for significance tests in multivariate contingency tables: An exegesis of Goodman's recent papers. In H. L. Costner (Ed.), Sociological methodology 1973-1974. San Francisco: Jossey-Bass, 1974.

Deming, W. E. Sample designs in business research. New York: Wiley, 1960.

Draper, N. R., and Smith, H. Applied regression analysis. New York: Wiley, 1966.

Duncan, O. D. Introduction to structural equation models. New York: Academic Press, 1975.

Edwards, A. L. Experimental design in psychological research (4th Ed.). New York: Holt, Rinehart and Winston, 1972.

Edwards, A. W. Likelihood: An account of the statistical concept of likelihood and its application to scientific inference. Cambridge: Cambridge University Press, 1972.

Erickson, B. H., and Nosanchuk, T. A. Understanding data. Toronto, New York: McGraw-Hill Ryerson, 1977.

Fairley, W. B., and Mosteller, F. (Eds.). Statistics and public policy. Reading, Massachusetts: Addison-Wesley, 1977.

Farrar, D. E., and Glauber, R. R. Multicollinearity in regression analysis: The problem revisited. Review of Economics and Statistics, 1967, 49, 92-107.

Festinger, L. The treatment of qualitative data by "scale analysis". Psychological Bulletin, 1947, 44, 146-161.

Firebaugh, G. A rule for inferring individual-level relationships from aggregate data. American Sociological Review, 1978, 43, 557-572.

Fisher, R. A. Statistical methods and scientific inference (1st Ed.). New York: Hafner Publishing Co., 1956.

Forbes, D., and Tufte, E. R. A note of caution in causal modelling. American Political Science Review, 1968, 62, 1258-1271.

Forsythe, A., and Hill, M. Design experiments for comparative evaluation of statistical packages. Proceedings of the Statistical Computing Section, American Statistical Association, 1975.

Francis, I., Heiberger, R. M., and Velleman, P. F. Report and proposal of the committee on evaluation of program packages to the section on statistical computing. Proceedings of the Statistical Computing Section, American Statistical Association, 1974.

Francis, I., and Heiberger, R. M. The evaluation of statistical program packages--The beginning. Proceedings of Computer Science and Statistics: Eighth Annual Symposium on the Interface, 1975.

Francis, I., and Valliant, R. The novice with a statistical package: Performance without competence. Proceedings of Computer Science and Statistics: Eighth Annual Symposium on the Interface, 1975.

Frank, O. Statistical inference in graphs. Stockholm: Forsvarets Forskninganstalt, 1971.

Frank, O. Survey sampling in graphs. Journal of Statistical Planning and Inference, 1977, 1 (3), 235-264.

Frank, O. Sampling and estimation in large social networks. Social Networks, 1978, 1(1), 91-101.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 1937, 32, 675-701.

Games, P. A. Multiple comparison of means. American Educational Research Journal, 1971, 8, 531-565.

Gentle, J. E. Comparison of statistical packages by users having some familiarity with computing and statistics. Proceedings of the Statistical Computing Section, American Statistical Association, 1975.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Goldberger, A. S. Econometric theory. New York: Wiley, 1964.

Goldberger, A. S., and Duncan, O. D. (Eds.). Structural equation models in the social sciences. New York: Seminar, 1973.

Goodman, L. A. Ecological regression and behavior of individuals. American Sociological Review, 1953, 18, 663-664.

Goodman, L. A. Some alternatives to ecological correlation. American Journal of Sociology, 1959, 64, 610-625.

Goodman, L. A. Snowball sampling. Annals of Mathematical Statistics, 1961, 32 (1), 148-170.

Goodman, L. A. The multivariate analysis of qualitative data: Interactions among multiple classifications. Journal of the American Statistical Association, 1970, 65, 226-256.

Goodman, L. A. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics, 1971, 13 (1), 33-61.

Goodman, L. A. A modified multiple regression approach to the analysis of dichotomous variables. American Sociological Review, 1972, 37, 28-46.

Goodman, L. A. The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. Biometrika, 1973, 60, 179-192, a.

Goodman, L. A. Causal analysis of data from panel studies and other kinds of surveys. American Journal of Sociology, 1973, 78, 1135-1191, b.

Goodman, L. A. Guided and unguided methods for the selection of models for a set of T multidimensional contingency tables. Journal of the American Statistical Association, 1973, 68 (341), 165-175, c.

Goodman, L. A. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. American Journal of Sociology, 1974, 79, 1179-1259, a.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231, b.

Goodman, L. A. A new model for scaling patterns: an application of the quasi-independence concept. Journal of the American Statistical Association, 1975, 70, 755-768.

Goodman, L. A. Analyzing qualitative/categorical data; log-linear models, and latent-structure analysis. Cambridge, Massachusetts: Abt Books, 1978.

Granovetter, M. S. Network sampling: Some first steps. American Journal of Sociology, 1976, 81, 1287-1303.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. Analysis of categorical data by linear models. Biometrics, 1969, 25, 489-504.

Guttman, L. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Guttman, L. The principal components of scale analysis. In S. A. Stouffer (Ed.), Measurement and prediction. Princeton, New Jersey: Princeton University Press, 1950.

Haberman, S. J. The analysis of frequency data. Chicago: University of Chicago Press, 1974.

Hambleton, R., Swaminathan, H., Cook, L., Eignor, D., and Gifford, J. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48 (4), 467-510.

Hammond, J. L. Two sources of error in ecological correlations. American Sociological Review, 1973, 38, 764-777.

Hannan, M. T. Aggregation and disaggregation in sociology. Lexington, Massachusetts: Lexington Books, 1971, a.

Hannan, M. T. Problems in aggregation. In H. M. Blalock, Jr., (Ed.). Causal Models in the Social Sciences. Chicago: Aldine-Atherton, 1971, b.

Hannan, M. T., and Burstein, L. Estimation from grouped observations. American Sociological Review, 1974, 39, 374-392.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. Sample survey methods and theory, (2 vols.), New York: Wiley, 1953.

Hardy, K. A., Kniefel, D. R., and Reynolds, W. A survey of statistical packages for research and instructional applications in the TUCC computing environment. Working Paper, North Carolina Educational Computing Service, 1975.

Hardy, K. A., Reynolds, C., and Kniefel, D. R. Comparison of statistical packages: A festures matrix approach. Proceedings of Computer Science and Statistics: The Tenth Annual Symposium on the Interface, 1977.

Heiberger, R. M. Activities and plans of the committee on evaluation of statistical program packages. Proceedings of the Statistical Computing Section, American Statistical Association, 1975a.

Heiberger, R. M. A procedure for the review of statistical packages and its application to the user: Interface with regression programs. Proceedings of Computer Science and Statistics: Eight Annual Symposium on the Interface, 1975b.

Heiberger, R. M. Criteria and considerations for computer programs for the analysis of designed experiments. Proceedings of Computer Science and Statistics: The Ninth Annual Symposium on the Interface, 1976.

Heise, D. R. Causal analysis. New York: Wiley, 1975.

Helwig, J. T., and Council, K. A. (Eds.). SAS user's guide, 1979 Edition. Raleigh, North Carolina: SAS Institute, Inc., 1979.

Hovland, C. I., Lumsdaine, A. A., and Sheffield, F. D. A baseline for measurement of percentage change. In P. Lazarsfeld and M. Rosenberg (Eds.), The language of social research: A reader in the methodology of social research. Glencoe, Illinois: Free Press, 1955.

Hull, C. H., and Nie, N. H. SPSS updates, New York: McGraw-Hill Book Company, 1979.

Journal of the American Statistical Association, 1978, 73 (361), 80-98.

Jencks, C. Inequality: A reassessment of the Family and Schooling in American. New York: Basic Books, 1972.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood analysis. Psychometrika, 1969, 34, 183-202.

Jöreskog, K. G. Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Jöreskog, K. G., and Sorbem, D. Statistical models and estimation methods for analysis of longitudinal data. In D. J. Aigner and A. S. Goldberger (Eds.), Latent variables and socio-economic models. Amsterdam: North Holland, 1976.

Kalbfleisch, J. G. Probability and statistical inference. Waterloo, Canada: Department of Statistics, University of Waterloo, 1971.

Kaplan, A. The conduct of inquiry: Methodology for behavioral science. San Francisco: Chandler Publishing Co., 1964.

Ketola, J. A comparison of some features of SAS 76.5, SPSS, and BMDP. Articles in the Newsletter of the University of Texas Regional Computing Center (May and June, 1978).

Kish, L. A. Some statistical problems in research design. American Sociological Review, 1959, 24, 328-338.

Kish, L. A. Survey sampling. New York: Wiley, 1967.

Klecka, W. R., Nie, N. H., and Hull, C. H. SPSS primer, New York: McGraw-Hill Book Company, 1975.

Klitzner, M. D. Hedonic intregration: Test of a linear model. Perception and Psychophysics, 1975, 18, pp. 49-54.

Kohm, R. F., Ryan, T. A., Jr. Preliminary report--Index of available statistical software. Proceedings of the Statistical Computing Section, American Statistical Association, 1975.

Kohm, R. F., Ryan, T. A., Jr., and Velleman, P. F. (Eds.). Index of publicly available statistical software. Stanford: ALCOA Laboratories and SPIRES/ EDUNET Data Base, Stanford Center of Information Processing.

Kmenta, J. Elements of econometrics. New York: Macmillan, 1971.

Kruskal, J. B. Transformations of data. In D. L. Sills (Ed.), International encyclopedia of the social sciences, Vol. 16. New York: Free Press, 1968.

Kruskal, W. H. Tests of significance. In D. L. Sills (Ed.), International encyclopedia of the social sciences, Vol. 14. New York: Free Press, 1968.

Kruskal, W. H., and Wallis, W. A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Asscciation, 1952, 47, 583-621.

Kuh, E., and Meyer, J. R. Correlation and regression estimates when the data are ratios. Econometrica, 1955, 23, 400-416.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer (Ed.), Measurement and prediction. Princeton, New Jersey: Princeton University Press, 1950.

Lazarsfeld, P. F., and Rosenberg, M. (Eds.). Survey design and analysis. New York: Free Press, 1955.

Lehmann, E. L. Nonparametrics: Statistical methods based on ranks. San Francisco: Holden-Day, 1975.

Leinhardt, S., and Wasserman, S. S. Quantitative methods for public management. Washington, D. C.: Urban Management Curriculum Development Program, National Training and Development Service, 1977.

Likert, R. A technique for the measurement of attitudes. Archives of Psychology, 1932, 140.

Lindsey, J. K. Inferences from sociological survey data--A unified approach. San Francisco: Jossey-Bass, 1973.

Lord, F. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14 (2), 117-138.

McNeil, D. R. Interactive data analysis, a practical primer. New York: Wiley Interscience, 1977.

Malinvaud, E. Statistical methods of econometrics (2nd Rev. Ed.). Amsterdam: North-Holland, 1970.

Miller, R. G., Jr. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

Morgan, D., and Rytina, S. Comment on "Network sampling: Some first steps" by Mark Granovetter. American Journal of Sociology, 1977, 83 (3), 722-729.

Mosteller, F., and Tukey, J. W. Data analysis and regression: A second course in statistics. Reading, Massachusetts: Addison-Wesley, 1977.

Muller, M. E. A review of manuals for BMDP and SPSS. Journal of the American Statistical Association, 1978, 73 (361), 71-79.

Nemenyi, P. Distribution-free multiple comparisons. Unpublished doctoral dissertation, Princeton University, 1963.

Nerlove, M., and Press, S. J. Univariate and multivariate log-linear and logistic models (Rand Report R-1306-EPA/NIH). Santa Monica, California: Rand, 1973.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., and Bent, D. H. SPSS, statistical package for the social sciences (2nd Ed.). New York: McGraw-Hill, 1975.

Nie, N. H. and Hull, C. H. SCSS, conversational statistical system, release 3.1, preliminary user's manual. Chicago: SPSS, Inc., 1979.

Nie, N. H. Review of manuals for BMDD and SPSS: Comment. Journal of the American Statistical Association, 1978, 73 (361), 82.

Norusis, M. J. SPSS statistical algorithms: Release 8.0. Chicago: SPSS, Inc., 1979.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Nunnally, J. C., and Durham, R. L. Validity, reliability and special problems of measurement in evaluation research. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research. Beverly Hills, California: Sage Publications, 1975.

Nunnally, J. C. and Wilson, W. H. Method and theory for developing measures in evaluation research. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research. Beverly Hills, California: Sage Publications, 1975.

Ostrom, C. W., Jr. Time series analysis: Regression techniques. In E. Uslaner (Ed.), Quantitative applications in the social sciences. Beverly Hills, California: Sage Publications, 1978.

Pierce, A. Fundamentals of nonparametric statistics. Belmont, California: Dickenson, 1970.

Price, J. L. Handbook of organizational measurement. Lexington, Massachusetts: Heath, 1972.

Przeworski, A. Contextual models of political behavior. Political methodology, 1974, 1, 27-61.

Rao, C. R. and Lahiri, D. B. (Eds.). Criteria of estimation in large samples. In Contributions to statistics. New York: Pergamon Press, 1965.

Rhyne, A. L., and Steel, R. G. A multiple comparisons sign test: All pairs of treatments. Biometrics, 1967, 23, 539-549.

Robinson, W. S. Ecological correlations and the behavior of individuals. American Sociological Review, 1950, 15, 351-357.

Rosenberg, M. The logic of survey analysis. New York: Basic Books, 1968.

Ryan, T. A., Joiner, B. L., and Ryan, B. F. MINITAB reference manual, University Park: Statistics Department, The Pennsylvania State University, 1979.

SAS. SAS introductory guide. Raleigh, North Carolina: SAS Institute, Inc., 1978.

SAS. SAS programmer's guide. Raleigh, North Carolina: SAS Institute, Inc., 1979.

SAS. SAS supplemental library user's guide. Raleigh, North Carolina: SAS Institute, Inc., 1979.

Savage, I. R. Bibliography of nonparametric statistics and related topics. Journal of the American Statistical Association, 1953, 48, 844-906.

Scott, W. A., and Wertheimer, M. Introduction to psychological research. New York: Wiley, 1962.

Shepard, R. N. Metric structures in ordinal data. Journal of Mathematical Psychology, 1966, 3, 287-315.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Sours, K. Statistical package index working paper. New York: Computer Center of the City University of New York, 1976.

Sprott, D. A., and Kalbfleisch, J. G. The use of the likelihood function in inferences. Psychological Bulletin, 1965, 64, 15-22.

Steel, R. G. A rank sum test for comparing all pairs of treatments. Technometrics, 1960, 2, 197-207.

Steel, R. G. Some rank sum multiple comparisons tests. Biometrics, 1961, 17, 539-552.

Stephan, F. Three extensions of sample survey technique. In N. Johnson and H. Smith, Jr. (Eds.), New developments in survey sampling. New York: Wiley, 1969.

Stevens, S. S. Measurement, statistics and the schemapiric view. Science, 1968, 161, 849-856.

Struening, E. L. Measurement. (an unpublished outline), 1979.

Sudman, S. Applied sampling. New York: Academic Press, 1976.

Theil, H. On the estimation of relationships involving qualitative variables. American Journal of Sociology, 1970, 76, 103-154.

Theil, H. Principles of econometrics. New York: Wiley, 1971.

Thisted, R. A. User documentation and control language: Evaluation and comparison of statistical computer packages. Proceedings of the Statistical Computing Section, American Statistical Association, 1976.

Thistlethwaite, D. L. and Campbell, D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Education Psychology, 1960, 51, 309-317.

Thorndike, R. L. (Ed.). Educational measurement (2nd Ed.). Washington, D.C.: American Council on Education, 1971.

Tufte, E. R. (Ed.). The quantitative analysis of social problems. Reading, Massachusetts: Addison-Wesley, 1970.

Tukey, J. W. Causation, regression, and path analysis. In O. Kempthorne (Ed.), Statistics and mathematics in biology. Ames, Iowa: University of Iowa Press, 1954.

Tukey, J. W. The future of data analysis. Annals of Mathematical Statistics, 1962, 33, 1-67.

Tukey, J. W. Analyzing data: sanctification or detective work. American Psychologist, 1969, 24, 83-91.

Tukey, J. W. Exploratory data analysis. Reading, Massachusetts: Addison-Wesley, 1977.

Tukey, J. W., and Wilk, M. B. Data analysis and statistics: Techniques and approaches. Proceedings of the Symposium on Information Processing in Sight Sensory Systems, 1965, California Institute of Technology.

Velleman, P. F., Seaman, J., and Allen, I. E. Evaluating package regression routines. Proceedings of the Statistical Computing Section, American Statistical Association, 1977.

Velleman, P. F., and Welsch, R. E. Some evaluation criteria of interactive statistical program pacakages. Proceedings of the Statistical ComputingSection, American Statistical Association, 1975.

Wallis, K. F. Introductory econometrics (Rev. Ed.). London: Gray-Mills, 1973.

Wallis, W. A., and Roberts, H. V. Statistics: A new approach. Glencoe, Illinois: Free Press, 1956.

Walsh, J. E. Handbook of nonparametric statistics, II: Results for two and several sample problems. Princeton, New Jersey: D. Van Nostrand, 1965.

Wasserman, S. S. Random directed graph distributions and the triad census in social networks. Journal of Mathematical Sociology, 1977, 5, 61-86.

Wilcoxon, F. Individual comparisons by ranking methods. Biometric Bulletin, 1945, 1, 80-82.

Wilcoxon, F. Some rapid approximate statistical procedures. New York: American Cyanamid Company, 1949.

Winer, B. J. Statistical principles in experimental design (2nd Ed.). New York: McGraw-Hill, 1971.

Wonnacott, R. J., and Wonnacott, T. H. Econometrics. New York: Wiley, 1970.

Wright, S. Correlation and causation. Journal of agricultural research, 1921, 20, 557-585.

CHAPTER 10: UTILIZATION AND TRANSFER OF EVALUATION RESULTS

INTRODUCTION

An important theme of the Guidelines is that the final measure of the worth of an evaluation is the extent to which its findings are used by decision makers and other consumers. This presupposes a responsibility on the evaluator's part to try to enhance the use of his/her own product. This responsibility goes far beyond simply presenting a technical written or oral report.

The role of the evaluator must be carefully developed as part of the earliest discussions between the researcher and decision makers. Although this "contract" need not be (and often is not) written, it should be explicit (Twain 1975). Our model of evaluation emphasizes the need for building timely feedback loops into the evaluation plan, thus setting the stage for later use of the findings.

Once the evaluation activity has reached the point where information and implications for change are forthcoming, there are several factors that must be considered in maximizing the potential for realizing that change. These factors are:

- Attributes of the findings
- Characteristics of the evaluation audience
- Characteristics of the organization
- Presentation of the findings
- Dissemination of innovations.

ATTRIBUTES OF THE FINDINGS

Evaluation findings are viewed by those who examine and make decisions based upon them in terms of certain characteristics. Glaser (1973) formulated the acronym CORRECT to represent some of them. These are:

"Credibility." The credibility of a finding stems from the scientific soundness of the evidence which supports its value. It also rests on the reputation of the presenter (evaluator) and the extent to which the finding is supported by respected persons or institutions.

"Observability." Acceptance will tend to follow from findings that can be directly shown. Whenever possible, potential users should have the opportunity to see a demonstration of the finding or its results.

"Relevence." The more a finding relates to the mission of the organization and to specific, persistent problems that interfere with the mission, the greater its likelihood for adoption.

"Relative advantage." Proposed changes may be related to existing practices in terms of cost benefit, cost effectiveness, or other advantages. Change will be adopted if these advantages offset the effort required to adopt them.

"Ease in understanding and installation." This attribute relates in part to observability. Highly esoteric, abstract concepts which cannot be readily understood or demonstrated are less likely to be accepted. Further, even though a proposed change may have relative advantages, it might not be practical to initiate.

"Compatability." Any finding is viewed by the potential user in terms of its relationship to his/her values, norms, current procedures, and the adaptability of existing facilities.

"Trialability, divisibility, or reversibility." Findings involving change are usually viewed as to how easy they may be to pilot test and whether they can be subdivided into components so that the proposed program can be introduced one step at a time. The findings will be viewed with less enthusiasm if adoption calls for an irreversible commitment.

The above attributes cannot be fully separated from the other factors involved in change; in fact, since they are phenomenological in nature, they help to point the way toward preparation of the evaluation report. Remember, though, that findings will be viewed differently by various audiences. The reports should present findings in such a way that the audience at hand is more likely to recognize the potential for positive change.

CHARACTERISTICS OF THE AUDIENCE

Audiences for evaluation findings can be categorized in terms of the major roles they assume in relationship to the program. Keep in mind, however, that although members of each category have at least one role in common, each group is comprised of diverse individuals. This diversity includes both personality and membership in subgroups (cliques, committees), and each member can assume a number of roles.

Table 3 identifies the potential audiences for evaluation findings and their language contexts. While overlap in categories exists (some clients may also be staff, and some staff may play an administrative role), it is possible to consider their unique needs as groups in the reporting of evaluation findings. The "taxonomy" in table 3 may also be useful in establishing relationships during the actual evaluation.

Categorizing audiences as above helps to provide the evaluator with an indication of the expectations of those to whom findings are being reported and to distinguish the language context relevant to each group. For example, discussions which focus on sample sizes, research design, and dependent or independent variables are couched in terms appropriate to an audience consisting of other evaluators, but discussions adressed to program sponsors are better presented in terms of descriptions of recipients, outcomes, and demonstration models (see Table 3).

Table 3. Language Contexts for Evaluation Audiences

AUDIENCES	LANGUAGE CONTEXT		
	Subjects	Instruments	Experimental Treatments
Clients	peers	tests	special classes
Community	children residents	questionnaires surveys	pilot programs
Staff	clients, student members	tests	class loads
Management	assignments	records	scheduling problems
Sponsors	recipients	outcomes	demonstration models
Evaluators	N's	dependent variables	independent variables

Following are brief definitions of the various potential audiences.

Clients. The involvement of clients at all levels of management and staff is a feature somewhat unique to the substance abuse prevention and treatment field. Because of this, evaluators often interact with clients who are peforming a variety of roles. These interactions can complicate the evaluation effort. When the primary objective of such interaction is to provide client (consumer) feedback, it can directly impinge on program activities.

Community. This category may include many segments of the community, such as friends and relatives of program participants, neighbors, business owners, leaders, and others interested in the evaluation. A major way information is transmitted to communities is through the mass media. It is imperative that such efforts be coordinated by the evaluator and program administration. Misinterpretations of ambiguously stated findings can easily destroy community support for prevention activities. It is strongly recommended that all media presentations be planned in advance and cleared with various groups that may be affected by the publicity. It is also advisable to seek help from media personnel in developing press releases.

Staff. From start to finish of the evaluation process, staff should be partners with the evaluator. But they may feel threatened by the evaluation, as these reports have often been used as devices for personnel review. If "playing it straight" with the evalutor could cost a staff member his/her job, then full cooperation cannot be expected. Therefore, during the initial stages of developing the evaluation "contract," the evaluator should make explicit the extent to which individual employe performance evaluation is a component of the plan. It is as important to safeguard the rights of the staff as it is those of program participants.

In presenting findings, the evaluator must keep in mind that the staff are probably strongly committed to the notion that what they are doing is effective, and believe that the nature of their activities is not amenable to description by scientific methods. Further, they might believe (sometimes rightly so) that the evaluator does not know as much about prevention as they do, thus reducing in their minds, the evaluator's credibility. The evaluator should exchange views and feelings with staff as the evaluation progresses, not only to satisfy the overt objectives of the evaluation, but also to minimize non acceptance of findings because of these conflicts.

Management. No matter who sponsors an evaluation, the persons most likely to feel that their own reputations and even survival are threatened by the evaluations are management personnel. The political stakes are high. Negative or ambiguous findings reported to funding sources can signal the demise of the organization or possibly shifts in key personnel. To the degree that management views evaluation as a punitive, fault finding process, management will, consciously or not, discourage an effective evalution from taking place.

Since it is the program managers who are responsible for the legal and ethical aspects of the program, as well as any rearrangement of schedules that some evaluations may require, strict adherence by the evaluator to human subjects procedures and the privacy act will enhance communication and support from the managers. The evalutor also should attempt to schedule tests and other data collection activities to minimize disruptions within a program. Furthermore, the evaluation should be self-supporting; where possible it should tie in to the related costs of the agency (staff time, rent, materials, travel, and the like).

Sponsors. The political ramifications or consequences of program evaluations prompt both demand for and resistance to the evaluative effort.

The following conclusions seem warranted:

- Congress and others are unlikely to make major funds available for prevention until effective models have been demonstrated.
- Effective models can be constructed only on the basis of comprehensive evaluation and analysis.

- These analyses must demonstrate not only that prevention programs have a positive impact on human behavior but do so in a cost efficient and effective manner.

Evaluators. In general, evaluators are able to exchange information using a common language, thus increasing understanding. However, because the field of drug and alcohol abuse prevention has attracted social scientists from a number of different fields, some terminology, preferred analytic techniques, and basic approaches to evaluation research can differ. Thus, any report prepared for an audience of even the most sophisticated social scientists should take into account this diversity.

The format for reporting evaluation results varies with the discipline. Formats often used may be found in the following: the Council of Biology editors style manual, A Guide for Authors, Editors, and Publishers in the Biological Sciences (health-related disciplines); National Education Style Manual for Writers and Editors (education); American Sociological Association Style Manual (sociology); and Publication Manual of the American Psychological Association (psychology).

Several major implications for reporting result from the interdisciplinary nature of the present cadre of evaluators. First, one must review the language context of the specific procedures involved in process, outcome, and impact. Second, in searching the literature it is necessary to consult indexes of various disciplines (for example, Index Medicus, ERIC, and Psychological Abstracts) since no single field covers evaluation.

When presenting reports in a group context, there are certain characteristics of the group and its individual members that are highly relevant in determining the approach to be taken. Even though evaluations are based on scientific methods and empirical evidence, a distinct element of persuasion exists in their presentations. In this regard, the research on small group behavior produced by social psychologists can be of great help to the evaluator in increasing the potential for acceptance of findings.

The considerations which must be taken into account in small group interactions are far too many to cover in the Guidelines. Several examples follow. It has been shown, for instance, that seating arrangements directly affect communications and the development of responses to leadership; those who value their membership in a group are most likely to be influenced by the opinions of other members; and those with low self-esteem are more easily persuaded than those with high self-esteem. One fairly comprehensive text in this field is Group Dynamics: The Psychology of Small-Group Behavior (Shaw 1971).

CHARACTERISTICS OF THE ORGANIZATION

Whether evaluations are sponsored by the prevention program, a funding source, or some other external body, reports on findings will typically be made to several organizations. The characteristics of the organization affect how the findings will be received and used. Davis (1971) developed the acronym, A VICTORY, to provide the evaluator with a checklist of factors accounting for organizational behavior related to adopting change. The factors are listed below:

"Ability." Does the program have the needed resources and capabilities? Specific budgets for starting up and maintaining various alternatives should be provided, as well as detailed descriptions of the proposed alternatives.

"Values." Do the proposed alternatives match the values, style, and philosophy of the program? One should analyze the existing values and perspectives of the organization and assess the conflicts that might exist with regard to the adoption of the most promising alternatives.

"Idea." How is the information required for the desired change to be acquired? Managers should encourage wide participation in the selection of the appropriate alternative, accompanied by a sharing of the necessary information. Typically such sharing may be among policymakers, staff, and even clients and community members, depending on the nature of the program.

"Circumstances." Is the present program at all suited for the alternative under study? Examine the context in which a proposed alternative was first derived and see how closely it matches the setting at hand. Where there are striking contrasts determine what kind of translation is needed to make the innovation appropriate for the present context.

"Timing." When is the right time to make the change? Pick a schedule most likely to result in a successful adoption.

"Obligation." How strong is the obligation to change? Commitment to change can be strengthened by holding group discussions, assigning responsibility for key outcomes, obtaining the endorsement of a recognized authority, and by encouraging public concern about the problem.

"Resistance." What resistance is likely to be encountered? One must anticipate and deal with likely sources of resistance on the part of the staff, participants, and community.

"Yields." What yields may be expected from the change? Provide participants with the rationale for selecting the particular alternative. Then, as the change is being implemented, make available additional feedback and attention to the needs and concerns of participants.

In discussing organization characteristics related to the likelihood of successful innovation, Glaser and Davis (1976) list four groups of variables: goals, structure, communication and decision making, and leadership and staff. A summary of this discussion follows.

Goals. Clarity of goals and acceptance by system members are positively related to the ability to initiate institutional change. Written statements of goals help to reduce anxiety about change and to impart a sense of security while new practices are being introduced. Organizations with sharply defined job descriptions tend to be reluctant to accept innovations. According to Glaser and Davis, "The best work is done when everyone shares the objectives, but each is relatively free to do his share of the common task in his own preferred way." (1976, p. 17.)

Structure. The literature on how organization structure affects the ability to adopt change is inconclusive. There is some indication that neither highly centralized nor greatly diffuse organizations are amenable to change. Inappropriate distribution of power within the organization and conflicting personal motives of individuals can subvert even the most promising innovation.

Citing Thompson (1965), Glaser and Davis point out that "Bureaucratic organizations ... are intrinsically resistant to innovation because they are monocratic, stress conformity rather than creativity, and are conservative in orientation." (1976, p. 19.) They see hierarchical structures as impeding decisions about new programs that are needed, and even if a decision is made to initiate a program, such structures inhibit their generation. Another debated issue is occupational specialization. Whereas some studies have found specialization positively related to the acceptance of innovations, others show that specialization impedes communication between subunits of the organization.

Guest (1962), as cited by Glaser and Davis, relates several factors to the time required by an organization to effectively perform under new behavior patterns. These include staff size, the number of specialized services and groups, the levels of hierarchy, the complexity of technical operations, and the degree of personal insecurity and hostility within the organization.

Communication and decision making. Free channels of communication, formal and informal as well as horizontal and vertical, have been found to be essential for adopting change. Feedback to the staff of surveyed attitudes about such issues as employe-management relations and work conditions helps to increase understanding and communication within the organization. The organization must support, by this method or others, the ability to ask for and give help. That is, colleague and administrator reinforcement and support are crucial to the development of a healthy organization. This leads naturally to the notion that participation in decision making increases the likelihood of general acceptance of decisions to initiate change.

Leadership and staff. Personality, training, and other characteristics of administrators are important to the acceptance of change. Indeed, the leader's personal traits cannot be considered separate from his/her role within the organization. And of course, staff expectations influence the leader's behavior, as do pressures from forces outside the organization. Change can best be introduced if leaders act as guides in focusing staff attention on problems and supporting staff in risk-taking and experimentation.

PRESENTATION OF THE FINDINGS

Lourie (1976) emphasizes that evaluation results must be in the language context of the audience. Neither slang nor technical jargon are recommended. Terms such as "subject," "unequal N's," "quasi-experimental," and so on, may be confusing or even incomprehensible to any audience aside from other researchers.

Reports should be brief and focus directly on those issues requiring decisions. Both primacy and recency have been found to be important in the retention of facts. That is, in different contexts, people tend to remember the first or last pieces of information given to them. Saying the same thing in several ways also increases retention.

Morris and Fitz-Gibbon (1978) present a thorough discussion of the written evaluation report, the major elements of which are:

- Summary. This is a brief overview of the report, why the evaluation was conducted, and its major conculsions and recommendations.
- Background information concerning the program. This section typically describes the origin, goals, and characteristics of the program.
- Description of the evaluation study. This includes the evaluation purpose, design, a description of what implementations and outcomes were measured, and data collection procedures.
- Results. This section is a straight forward presentation of the results of the evaluation.
- Discussion of the results. Here, the researcher interprets each finding stated in the previous section.
- Costs and benefits. This section discusses the methods used and results of cost-effective or cost-benefit analyses.
- Conclusions and recommendations. This part spells out findings introduced in the summary of the report.

In their well constructed discussion of presenting reports, Morris and Fitz-Gibbon also emphasize the use of graphic displays to present data. The authors of the Guidelines agree strongly with this message. Even the most avid "number cruncher" can get lost trying to pull the meaning out of page after page of data when they are poorly presented.

DISSEMINATION OF INNOVATIONS

The major focus of this chapter so far has been on factors which facilitate the use of evaluation results within the organization which has been evaluated. However, it is also important to examine factors that can facilitate the dissemination and utilization of innovative programs. The suggestions made are based on two assumptions: first, a program (for example, on drug education) has already been developed and evaluated with favorable outcomes; and second, the developers (or evaluators) seek maximum utilization of the program.

Because the literature of knowledge transfer and change has semantic inconsistencies, we shall define terms germane to this discussion.

Innovation. This term is used two different ways in the literature. It describes the process by which a "new" product, that is, an idea, program, or practice, is translated into a particular environment, and it denotes the new product itself. The "product" definition of innovation adopted for use here is a plan or practice new to a particular school, agency, group, or site, which by its newness necessitates (or is intended to produce) some observable behavioral or programatic change.

Diffusion. Diffusion refers to the process by which new ideas, products, or programs (that is, innovations) are communicated to the members of a social system. A popular synonym for diffusion of innovations is technology transfer. It refers to a process and is the umbrella under which the following terms can be found.

Dissemination. Dissemination refers to the spreading of information (in this context, details about a new idea or program or the results of an evaluation of a program), and as such can be considered the first step in the diffusion process. To disseminate a program means simply to make information about it known widely, for example, to potential adopters.

Adoption. For a long time, researchers in the education change field considered adoption of an innovation synonymous with its implementation. Today, however, there is growing recognition that adoption more accurately reflects a decision to implement a program rather than implementation of the program (innovation) itself.

Utilization. Utilization of an innovation suggests that the program has been (or is in the process of being) implemented, that is, put into practice in a new environment.

Because most prevention programs include education strategies, the literature on educational change is particularly relevant to the dissemination and utilization of prevention innovations.

Many models of educational change have been developed. Perhaps the best known are the Research, Development, and Diffusion Model (RD&D), the Social Interaction Model (S-I), the Problem Solving Model (P-S), and the Linkage Model. In addition to normative and descriptive models, various administrative programs for change have been suggested, along with conceptual schemes and ideological and instrumental strategies. A discussion of many of these approaches as well as the problems caused by the very multiplicity is found in Sieber (1974).

One of the few things on which there seems to be some consensus among these different approaches to change is that the process of innovation is multistaged. There is no agreement, however, as to the number of stages or how they are defined. Berman and McLaughlin (RAND 1975, 1977) suggest that there are three stages of innovation: support (recognition of need and search for an innovative program), implementation (the change process as an innovation impinges on the organization), and incorporation (routinization of the innovation into the system).

Until the 1970's, educational innovations were developed with little apparent regard for their diffusion; that is, it was difficult to point to any obvious impact of these innovations on the education community. The Federal government, as the primary sponsor of educational research, began emphasizing dissemination of the innovations for which it had paid. Hence, a number of vehicles for dissemination were developed in the 1970's. While their approaches differed, their goal was the same--technology transfer.

The first such vehicle was the Pilot State Dissemination Project (PSDP), started in 1970. This was designed to increase the use of The Educational Resources Information Center (ERIC) by local researchers. The ERIC system, established in 1965, had been virtually untouched by school personnel at the local level. The PSDP utilized a user-choice strategy based on retrieval of options from the ERIC system, and provided personal links between information sources and potential users (for an evaluation of PSDP see Sieber 1972). In 1974, the U.S. Office of Education began more direct participation in the diffusion of innovations by establishing the National Diffusion Network (NDN) under Title III of ESEA.

The strategy of this network is to make available to local education agencies those programs and products whose effectiveness had been demonstrated in school settings. An evaluation of NDN in 1977 by the Stanford Research Institute found it to be highly successful (see Emrick, Peterson, and Agarwala-Rogers 1977).

The National Institute of Education (NIE) has also taken an active interest in dissemination and utilization of educational innovations through many of its regional programs, including the R&D Utilization Project, the R&D Exchange System, and the State Capacity-Building Program. (Several of these NIE programs are currently being evaluated; no results are yet available). A good review of these various Federal programs may be found in Herling (1977).

Based on completed evaluations of several of the Federal dissemination vehicles just noted, it is possible to suggest certain factors which appear to facilitate dissemination of educational innovations in general. The major additional points are as follows:

- The program description should be comprehensive and attuned to the need for adaptability to the particular circumstances of each new location. This flexibility allows the potential adopter to have a sense of ownership of the innovation, thus increasing his/her commitment to its goals.

- The innovative program should include an extensive support system in the form of both materials and services. The materials should include goals and objectives and a full description of the process of the program (including, of course, an evaluation component). Support services should include inperson assistance to potential adopters. The opportunity to observe the program in action should be afforded, if possible, to future implementers.

The potential adopter of a proven program must be attuned to the same issues as the evaluator of an innovation. S/he must assume the role of change agent to ensure support both within the agency and in the local community. It has been shown that some sort of personal intermediary or linker between the program developer and the potential adopter facilitates the dissemination process (Glaser 1976).

For a very thorough exposition of the issues regarding dissemination, it is strongly recommended that the reader consult Putting Knowledge to Use: A Distillation of the Literature Regarding Knowledge Transfer and Change, compiled by Glaser and Davis (1976). This chapter has leaned heavily on this work, including both the dicussions by these authors and the extensive list of references contained in their annotated bibliography.

REFERENCES

Berman, T., and McLaughlin, M. Federal programs supporting educational change Volumes I-VIII. Santa Monica, California: Rand Corporation, 1975, 1977, 1978.

Davis, H. R. A checklist for change. In N.I.M.H. A manual for research utilization. Washington, D. C.: U. S. Government Printing Office, 1971.

Emrick, J. A., Peterson, S. M., and Agarwala-Rogers, R. Evaluation of the National Diffusion Network, Vol. 1. Menlo Park, California: Stanford Research Institute, 1977.

Glaser, E. M. Knowledge transfer and institutional change. Professional Psychology, 1973, 4, 434-444.

Glaser, E. M., and Davis, H. R. Putting knowledge to use: A distillation of the literature regarding knowledge transfer and change. Rockville, Maryland: Human Research Institute, Los Angeles, in collaboration with National Institute of Mental Health, 1976.

Guest, R. Organizational change: The effect of successful leadership. Homewood, Illinois: Dorsey Press and Richard D. Irwin, Inc., 1962.

Herling, R. The tenth national dissemination conference: A report. Washington, D. C.: Council of State Offices, 1977.

Lourie, S. Policy research and decision making in education. In C. Abt (Ed.), The evaluation of social programs. Beverly Hills, California: Sage Publications, 1976.

Morris, L. L., and Fitz-Gibbon, C. T. How to present an evaluation report. Beverly Hills, California: Sage Publications, 1978.

Patton, M. Q. Utilization-focused evaluation. Beverly Hills, California: Sage Publications, 1978.

Sieber, S. D. Trends in diffusion research: Knowledge utilization. In A. R. Jwaideh and B. H. Bhola (Eds.), Research in diffusion of educational innovations: A report with an agenda. Bloomington, Indiana: School of Education, Indiana University, 1974.

Shaw, M. E. Group dynamics: The psychology of small group behavior. New York: McGraw-Hill, 1971.

Thompson, V. A. Bureaucracy and innovation. Administrative Science Quarterly, 1965, 10 1-20.

Twain, D. Developing and implementing a research strategy. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol. I. Beverly Hills, California: Sage Publications, 1975.

APPENDIX TO PART II: INSTRUMENTS AND DATA SOURCES

INTRODUCTION

No matter how creatively designed, well controlled and smoothly executed an evaluation design is, it is only as good as the measures from which data are derived. As a worn needle will impair the sound of the finest hi-fi,so inadequate, noisy, or inappropriate measures will impair the quality of the most elaborate and expensive evaluation. From this perspective, it is difficult to devote too much attention to measurement issues.

The most commonly used measures for all levels of program evaluation are written ("paper and pencil") instruments. Their low cost, relative ease in scoring and administration, and generally high reliability make written instruments attractive data collection devices for prevention evaluation. This appendix reviews a number of instruments which are currently available, and that have, in some cases, been used to evaluate prevention programs. In addition, this appendix adresses a number of issues which relate directly to the selection and use of the instruments presented. These issues, discussed in the following three sections, are:

- Assessment and interpretation of reliability
- Validity
- Reliability and validity vs. relevance

The discussions presented here are in no way intended to be a comprehensive treatment. Rather, they are presented with the hope that they will stimulate thought and perhaps contribute somewhat to wiser use and selection of measurement tools. In each of these discussions a minimal knowledge of psychometric issues is assumed.

ASSESSMENT AND INTERPRETATION OF RELIABILITY

Reliability refers, in the largest sense, to the extent to which an instrument is stable or error free. A major issue concerning reliability is the tendency among practitioners (and some measurement texts) to minimize or ignore the distinction between reliability as stability and reliability as relative internal freedom from error. As elaborated below, choosing a "highly reliable" instrument without reference to this distinction may lead to some unhappy choices.

A second issue concerning reliability is that reliability coefficients are affected by the variance of the scores upon which the coefficient is based. All else being equal, the reliability of an instrument will increase with increasing heterogeneity of scores. Thus, reliability data are most useful when accompanied by supporting information such as sample variances and scatter plots (a condition which, unfortunately, rarely obtains). Other problems of correlational measures (discussed in chapter 9), will, of course, also apply to most measures of reliability.

A final issue concerning reliability is the correctness of speaking of the reliability of an instrument. In truth, a reliability coefficient is as much a function of the population being assessed and the conditions under which the instrument is administered as it is a funct[illegible] the psychometric qualities of the instrument. Such variables as enthusiasm of the [illegible],

motivation of the respondents, or even room color and weather can all affect reliability coefficients. (The score variance issue discussed above is a concrete illustration of this point.)

In sum, the above considerations prevent treating reliability simply as a number between 0 and 1 which should exceed .80 (or some other criterion). Rather, reliability is most usefully conceptualized as a set of statistical, psychometric, and situational variables and conditions which affect the error in the stability of data gathered by a given instrument.

The remainder of the current discussion of reliability points out some peculiarities of three common measures of reliability.

Test-retest Reliability. Calculated as the correlation between scores on an instrument at two points in time, test-retest reliability speaks directly to the stability (as opposed to the freedom from error) of a measure. In fact, test-retest reliability is sometimes referred to as the "coefficient of stability."

A problem arises because "stable over time" may have a number of meanings which cannot be unraveled simply, given the test-retest correlation. For example, a high correlation might mean that respondents' scores remained unchanged over time, or that they changed according to some linear (or at least monotonic) function. Similarly, a low correlation may reflect either essentially random error around a stable group mean or it might reflect some systematic but nonmonotonic change.

Without access to sample means and, ideally, scatterplots, it is difficult to interpret the meaning of either high or low test-retest reliability coefficients. Unfortunately, such information is rarely given in test reports or manuals and may only be available through personal communication with instrument authors. (For a discussion of other drawbacks of test-retest reliability, see Nunnally and Durham 1975).

In general, test-retest reliability coefficients are over used in relation to their worth in selecting evaluation instruments.

Split-Half Reliability. Another commonly used measure is the split-half approach. Here, the items are divided into subsets (typically odd and even items), and the correlation between subsets is calculated to yield an estimate of reliability. A frequently used technique is the Spearman-Brown Formula (Anastasi, 1976). The problem here is that reliability will vary depending upon how the items are divided. How large this variation will be depends upon a number of factors including the true structure of the attitude, trait, competency, and so forth being measured (that is, uni- vs. multi-dimensional), the ordering of items, and the length of the instrument. As with test-retest reliability, split half reliability is probably more commonly used than its worth warrants.

Internal Consistency. A more appropriate estimate of reliability is provided by measures of internal consistency, such as coefficient alpha (Nunnally & Durham, 1975) and Kuder-Richardson techniques (Anastasi 1976). However, because these coefficients were in the past more difficult and time consuming to calculate than a split half correlation, they have been less often encountered in the prevention psychometric literature.

VALIDITY

A number of definitions of validity may be found in the measurement literature. One useful definition is that validity "indicates the extent to which (an instrument) is capable of achieving certain aims" (Issac and Michael 1979, p. 83). These aims may include: (1) providing a representative sample of some universe of behaviors, attitudes, or competencies; (2) substituting for some more expensive, time consuming measure, or (3) inferring the presence or degree of presence of some nonobservable, intrapsychic variable. For further discussions of content, criterion related and construct validity, the reader is referred to Issac and Michael's original discussion (see, also, Anastasi 1976).

A fourth type of validity, currently the topic of some debate among evaluators, is face validity. On the one hand are experts such as Nunnally who bemoan the illogic "of the reluctance on the part of some administrators in applied settings ... to permit the use of

predicator instruments which lack face validity" (1970, p. 149). The other position is taken by Michael Patton, who believes that "one of the best ways to facilitate decision maker understanding of and belief in evaluation data is to place a high value on the face validity of research instruments" (1978, p. 244, emphasis added).

In truth, Patton and Nunnally are probably not disagreeing. No one will argue that face validity alone is sufficient justification for choosing an instrument. However, all else being equal, a face valid instrument seems a strong asset for the evaluation process. More generally, evaluators are urged to consider Patton's advice that consumers be asked prior to any data collection whether a given instrument seems useful and believable.

Issues of utilization aside, critics of face valid instruments sometimes argue that such instruments are too easy for subjects to "figure out." The concern here seems to be that subjects will try to look good, succumb to demand pressures, or otherwise "fool around" with face valid instruments. However, respondents can be equally clever in undermining nonface valid instruments. Also, consistency checks can be built into any instrument which allow detection of false responses (see, for example, Royer Cook's use of a bogus drug, "cadrines," on a drug use survey elsewhere in this appendix). The real issue, however, is not detecting false responses, but rather avoiding them in the first place. Avoiding them has much more to do with the testing environment, particularly the tester's relationship to respondents, than it does to do with measurement theory or practice.

RELIABILITY AND VALIDITY VS. RELEVANCE

The above discussions suggest that obtaining adequate technical information (especially reliability information) about potential instruments may require substantial effort on the part of prevention evaluators. Paradoxically, it is possible to argue that too much emphasis on reliability and validity clouds the real issue in instrument selection, namely relevance.

In a general sense, relevance refers to the fact that the ultimate criterion for choosing (or rejecting) an instrument is not measured in terms of reliability or validity coefficients. Rather, it is measured in terms of the ability of a given instrument to meet the informational needs of a given program. Unless informational needs are met, the instrument is, by definition, useless.

Often, determining the relevance of an instrument is as straightforward as determining the match between program objectives and instrument items and scales. Such a strategy will likely be most useful when program objectives refer to changing or instilling certain attitudes, teaching specific information or competencies, or changing the frequency or probability of specific behaviors. The interested reader is referred to Morris and Fitz-Gibbon (1978) for a step by step method for determining the match between program objectives and instrument items.

Determining the relevance of an instrument becomes more difficult when the instrument is intended to measure a complex, intrapsychic variable such as perceived locus of control or self-esteem. Here, there is an ever present risk that the variable as operationalized by the instrument's author and the variable as operationalized by the program will be similar in name only. A careful examination of the theoretical and empirical bases of an instrument can provide insight into the extent to which the instrument and the program are talking about the same or different variables.

Evaluators are urged to thoroughly research the literature concerning an instrument before making an assessment of its relevance to a given program effort. Failure to pay adequate attention to issues of relevance is one of the surest ways to decrease the probability of finding program effects.

CONCLUSION

Few of the instruments in this appendix adress all of the above issues. Thus, an initiative for the prevention field should be the identification and development of reliable, valid, and relevant instruments for prevention evaluation.

REFERENCES

Anastasi, A. Psychological assessment (4th Ed.) New York: MacMillan, 1976.

Isaac, S. and Michael, W. Handbook in evaluation and research. San Diego, California: Edits Publishers, 1979.

Morris, L. and Fitz-Gibbon, C. How to measure achievement. Beverly Hills, California: Sage Publications, 1978.

Nunnally, J. C. Introduction to psychological measurement. New York: McGraw-Hill, 1970.

Nunnally, J. C., and Durham, R. C. Validity, reliability, and special problems of measurement in evaluation research. In E. Struening and M. Guttentag, (Eds.), Handbook of evaluation research, Vol I. Beverly Hills, California: Sage Publications, 1975.

Patton, M. Q. Utilization focused evaluation. Beverly Hills, California: Sage Publications, 1978.

ORGANIZATION OF THE APPENDIX

Each instrument is coded regarding its potential use in process (p), outcome (o), or impact (i) evaluation. The instruments are grouped in the following categories: multiscale batteries, intrapersonal scales, interpersonal scales, and substance scales.

The appendix also includes descriptions of books which contain similar instruments and data sources that should be considered for process, outcome, and impact indicators.

The reviews of instruments in this section were based on several secondary sources including:

(1) Test manuals or special texts (when available).

(2) Robinson, J. P., and Shaver, P. R. Measures of social psychological attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1973.

(3) Pfeiffer, W. J., and Helsin, R. Instrumentation in human relations training. Iowa City, Iowa: University Associates, 1973.

(4) Lake, D. G., Miles, M. B. and Earle, R. B. Measuring human behavior. New York: Teachers College Press, Columbia University, 1973.

(5) Abrams, A. L., Garfield, E. F., and Swisher, J. D. Accountability in drug education: A model for evaluation. Washington, D. C.: Drug Abuse Council, 1973.

(6) Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1965.

Where appropriate direct quotations are indicated. Any instrument selected for use should be subjected to a current review of the manual, relevant research, and available critiques. Each scale should also be obtained and the items examined. In many instances only selected subscales are appropriate as indicators for prevention programs. If subscales only are used, they should be carefully pretested.

REVIEW OF INSTRUMENTS

MULTISCALE BATTERIES

California Psychological Inventory (p,o,i)

Author: Harrison G. Gough

Ages: Thirteen years and over

Date: 1957, Revised 1964, 1969, 1975

Source: Consulting Psychologists Press, Inc.
577 College Avenue
Palo Alto, CA 94306

Variables: The CPI consists of 18 standard scales designed to assess personality characteristics important for social living and social interaction. The scales are grouped into four broad categories. Each scale serves as an index of interpersonal functioning as described below (Gough 1975).

1. Measures of poise, ascendency, self-assurance, and interpersonal adequacy.

   (a) Dominance--leadership, dominance, persistence, and social initiative.

   (b) Capacity for status--personal qualities which underlie and lead to status.

   (c) Sociability--outgoing, sociable, and participative temperment.

   (d) Social presence--poise, spontaneity, and self-confidence in personal and social interaction.

   (e) Self-acceptance--sense of personal worth, self-acceptance, and capacity for independent thinking and action.

   (f) Sense of well being--self-assured with minimal worries and complaints.

2. Measures of socialization, maturity, responsibility, and interpersonal structuring of values.

   (a) Responsibility--conscientious, responsible, and dependable disposition and temperment.

   (b) Socialization--social maturity, integrity, and rectitude.

   (c) Self-control--self-regulation and freedom from impulsivity and self-centeredness.

   (d) Tolerance--accepting and nonjudgmental social beliefs and attitude.

   (e) Good Impression--capacity for creating a favorable impression, and concern about others' reactions.

(f) Communality--correspondence of reactions and responses to a common pattern.

3. Measures of achievement potential and intellectual efficiency.

   (a) Achievement via conformance--factors of interest and motivation which facilitate achievement in a setting requiring conformance.

   (b) Achievement via independence--factors of interest and motivation which facilitate achievement in a setting requiring autonomy and independence.

   (c) Intellectual efficiency--degree of attainment of personal and intellectual efficiency.

4. Measures of intellectual and interest modes.

   (a) Psychological mindedness--interest in and responsibility to the inner needs, motives, and experiences of others.

   (b) Flexibility--adaptability of a person's thinking and social behavior.

   (c) Feminity--masculinity or femininity of interest.

Description: The CPI is made up of 480 items which the subject answers as true or false.

Administration and Scoring: The inventory is largely self-administering to literate subjects. Time required for administration is 45-60 minutes.

Special answer sheets are available for use with the computer scoring service maintained by the publisher. Handscoring answer sheets are also available. Scoring yields 18 raw scores which can be transfered to a profile form that provides graphic representations of standard scores.

Reliability: Test-retest reliabilities reported in the manual (Gough 1975) include a sample of 200 prisoners retested after 7-21 days with individual scale reliabilities ranging from .49 to .87, and two high school classes retested after one year resulting in correlations ranging from .38 to .75 for males and from .44 to .77 for females.

Validity: Evidence for validity of the scales is summarized in the manual indicating the each of the scales has some validity when compared with specific performance criteria.

Criticism: The scales are not entirely independent and may not measure 18 separate traits and reliabilities. The concepts may be dated and value laden.

Suggestions for Use: A handbook has been assembled by Megargee (1972) explaining the development, characteristics, and scales of the CPI. The published literature is reviewed and uses of the CPI in assessment and research are discussed.

References:

Gough, H. G. Manual for the California Psychological Inventory. Palo Alto, California: Consulting Psychologists Press, Inc., 1969.

Megargee, E. I. The California Psychological Inventory handbook. San Francisco: Jossey-Bass, 1972.

Educational Quality Assessment Inventory (p,o,i)

Authors: Developed by staff of Division of Educational Quality Assessment, Pennsylvania Department of Education.

Ages: Ten/eleven, thirteen/fourteen, sixteen/seventeen (grades 5, 8 and 11)

Date: 1970-1973, with extensive revision in 1977

Source: Division of Educational Quality Assessment
Bureau of Planning and Evaluation
Pennsylvania Department of Education
Box 911
Harrisburg, PA 17126
Note--Persons interested in use of this instrument should send inquiries, stating intended use, to Dr. J. Robert Coldiron, Division Chief.

Variables: The student's self-esteem, attitudes toward others and school, citizenship, health and safety knowledge and practices, creative activities, career awareness, appreciation and knowledge of human accomplishments, and problem-solving skills are examined. Basic academic skills, including reading, writing, and mathematics are also measured.

Description: The inventory is based on ten quality educational goals and includes measurement devices associated with each. Test construction is based on matrix sampling--a method whereby each student responds to one portion of the total test items for each goal. Each portion or form is balanced with the same number of items and with positively and negatively worded items. An appropriate test is available for grades 5 (209 items), 8 (290 items), and 11 (188 items).

Administration and Scoring: The test is self-administered and the forms are randomly dispersed, usually in classroom groupings.

Scoring is done by schools. A school mean score, based on individual student scores, is determined for each goal area of the test. For each goal, a school's percentile state rank is determined by comparing the school's score to the scores of a group of normative schools. Information is collected on resources available to each school, for example, physical facilities, financial resources, teachers, and home conditions. A score range for a school is predicted based on correlation coefficients computed between these quantified conditions and scores for the normative group. The predicted score range is an indication of how other schools score when operating under similar conditions.

A model of criterion-referenced scoring is also provided. A score based on the percent of students in a school completing the inventory and meeting a criterion for minimum positive responses is provided. Criterion-referenced information is given for attitudinal as well as cognitive areas. The school receives a score denoting the percent of students satisfying the criterion and percent of students statewide answering a majority of the items favorably.

Statewide and individual school items response data are also provided, based on the percent of students that gave each of the possible answers for each of the items on the battery.

Reliability: Median internal consistency reliability estimates for 1978 range from .73 to .94 for grade 5, .80 to .95 for grade 8, and .80 to .96 for grade 11. Reliability estimates and item characteristics (mean, s.d., and so on) for all students were essentially the same for all grade levels.

Validity: A comparison of the EQA measure of self-concept with another scale given to more than 300 children in each of the three grades indicated a strong positive correlation for each grade level (r's of .77, .81, and .76; .69, and .79; and .69 respectively). In addition, independent teacher ratings on two components of self-esteem and the EQA measure of self-esteem showed significant correlations. The EQA measure of understanding others was compared in one school district with teacher ratings of students as high or low in their acceptance of those who are different from them. For grades five and eight it was found that the high and low acceptance groups differed significantly from one another with regard to the total score on the EQA measure. No findings in support of the validity of the EQA were found for grade eleven, but the problem may be in the validation process rather than in the instrument itself.

Other comparison studies on the EQA measures of interest in school and learning, health knowledge and habits, and decision making found strong relationships between the EQA measures and other tests. Further reliability and validity information is included in the manual (Division of Educational Quality Assessment 1978a).

Criticisms: The battery is designed for use at a group level (the unit of analysis of all data is the school) and absence of individual student profiles may limit usage. Students may respond to test items the way they believe they are expected to respond, reading skills may affect performance on the test, and the test measures attitudes and intended actions rather than actual behavior. The health practices variable does not adequately cover substance use.

Suggestions for Use: Strengths and weaknesses of a school's performance on goals tested can be discovered by comparing school item data with statewide item data. Schools may classify test items according to when, of if, the items have been included in the instructional programs, and thus compare student performance and test content with local objectives (Division of Educational Quality Assessment, 1978b).

References: Division of Educational Quality Assessment. Getting inside the EQA inventory. Harrisburg, Pennsylvania: Pennsylvania Department of Education, 1978a.

Division of Educational Quality Assessment. Manual for interpreting elementary school reports. Harrisburg, Pennsylvania: Pennsylvania Department of Education, 1978b.

INTRAPERSONAL SCALES

Self-Assessment Scales (p,o,i)

Author: Ardyth A. Norem-Hebeisen

Ages: Junior and Senior High School

Date: 1975

Source: National Humanistic Education Center
110 Spring Street
Saratoga Springs, New York 12866

Variables: The Self-Assessment Scales instrument is designed to measure self-esteem, which it differentiates along 7 dimensions (Norem-Hebeisen 1976):

1. Showing Feeling - Ease in having and showing strong feelings, including the expression and experience of personal closeness.
2. Being Known - Comfort in allowing others to know one's inner self.
3. Performance Sources - Extent to which feelings of personal worth are independent of personal achievement and skill.
4. Social Sources - Extent to which feelings of personal worth are independent of the judgements and approval of other persons.
5. Self-Evaluation - How favorably one compares self with peers.
6. Real-Ideal Congruence - The relative congruence between what respondents think they are and what they think they should be.
7. Well Being - A general sense of personal worth and comfort with self.

Description: The Self-Assessment Scales is composed of 66 items which respondents rate on a seven point, "Completely True Of Me" to "Not True Of Me" Likert type scale.

Administration and Scoring: The Self-Assessment Scales is self-administered with instructions printed on the test booklet cover. Hand scoring sheets are available from the distributor.

Reliability: Cronbach alphas for the seven scales (dimensions) of the self-assessment scales are uniformly high (.77-.88), based on a population of 376 9th, 11th, and 12th grade, white, middle to upper middle class, nondeviant Midwestern suburban youth. (Ahlgren and Norem-Hebeisen 1979.)

Validity: Ahlgren and Norem-Hebeisen (1979) compared the 376 youths described above (Norm group) with four groups of deviant youth (Runaways, n=20, Learning Disabled, n=17, Drug Abusers Pretreatment, n=36, and Drug Abusers in Treatment, n=21) and one group of Posttreatment Ex-Drug Abusers (n=23). These six groups showed distinctly different patterns of responses on the Self-Assessment Scales. Specifically, the Self-Assessment Scales seem able to differentiate the pretreatment and intreatment drug abusers from both the normal and posttreatment groups, and from the two nondrug-abusing deviant groups based upon a number of analyses. A similar study by Norem-Hebeisen (reported in Lettieri 1975) also suggests reasonable discriminant validity for the Self-Assessment Scales.

Criticisms: The seven scales of the Self-Assessment Scales show considerable intercorrelations. The author reports that principal component analysis of the seven

simple sum scales yields four factors: basic acceptance, conditional acceptance, self-evaluation, and real-ideal congruence (Ahlgren and Norem-Hebeisen 1979).

Suggestions for Use: The Self-Assessment Scales is a psychometrically sound tool for assessing self-esteem as conceptualized by the author. Programs operationalizing self-esteem in like fashion should strongly consider this instrument, but may wish to use four, rather than seven scales (see criticisms, above).

References: Ahlgren, A. and Norem-Hebeisen, A. A. Self-esteem patterns distinctive of groups of drug abusing and other dysfunctional adolescents. The International Journal of Addictions, 1979, 14 (6), 759-777.

Norem-Hebeisen, A. A. A multidimensional construct of self-esteem. Journal of Educational Psychology, 1976, 68 (5), 559-565.

Norem-Hebeisen, A. A. Self-esteem as a predictor of adolescent drug abuse. In D. J. Lettieri (Ed.), Predicting adolescent drug abuse: A review of issues, methods, and correlates. Rockville, MD: National Institute on Drug Abuse, 1975.

The Piers-Harris Children's Self Concept Scale: "The Way I Feel About Myself" (p,o,i)

Author: Ellen V. Piers

Ages: Nine - eighteen years

Date: 1976

Source: Counselor Recordings and Tests
Box 6184 Acklen Station
Nashville, Tennessee 37212

Variables: Based on a factor analysis (Piers 1976a), the scale appears to measure six dimensions of self concept including perceptions of one's behavior, school status, physical appearance, anxiety, popularity, and happiness. These dimensions are scored as a single concept index.

Description: The scale presently consists of 140 items which are answered "yes" if a statement is true and "no" if the statement is false (for example, I am a good reader. yes no). The simple format and third grade reading level (Piers, 1976a) allows for group testing as low as age six, and if administered individually, could be given to even younger groups.

Administration and Scoring: The scale takes approximately 15 to 20 minutes to administer and a scoring key based on expert judgements of positive answers is available in the manual (Piers 1976a). The scale is easy to administer and score, but Piers (1976a) suggests that caution be taken with interpretation of results. Percentiles for grades four through twelve are available.

Reliability: Piers (1976a) reports internal consistencies ranging from .78 to .93 and four month test-retest reliabilities ranging from .72 to .77. In a research monograph, Piers (1976b) also reports test-retest reliabilities (three weeks to seven months) ranging from .62 to .96. These reliabilities are certainly at an acceptable level.

Validity: In the test manual, Piers (1976a) reports high concurrent validity correlations with other self concept scales (that is .68 with Lipsitt's Children's Self-Concept Scale and .64 with the SRA Junior Inventory). The research monograph (Piers 1976b) reports similarly good correlations with Tennessee Self Concept Scale (.61) and the Self Esteem Inventory (.85). A factor analysis suggests adequate construct validity and the original item process selection provides content validity (Piers 1976a).

Criticisms: The scale was originally designed for research purposes and Piers (1976a) recommends that other uses of the scale be undertaken with reservations.

Suggested: The Pier-Harris Children's Self Concept Scale is probably the most frequently used outcome measure for self concept changes in the prevention evaluation literature. Its high reliability, however, may make changes more difficult to assess. It remains one of the simplest and most adequate devices in this domain.

References: Piers, E. V. Manual for the Piers-Harris children's self concept scale (The way I feel about myself). Nashville, Tennessee: Counselor Recordings and Tests, 1976a.

Piers, E. V. The Piers-Harris children's self concept scale: Research monograph No. 1. Nashville, Tennessee: Counselor Recordings and Tests, 1976b.

Self-Esteem Inventory (p,o)

Author: S. Coopersmith

Ages: Forms A and B — Nine years and over
Form C — Adults

Date: 1967

Source: Self Esteem Institute: San Francisco, California

Variables: The inventory is a form for self evaluation, specifically in the areas of general satisfaction with self, perceived evaluation by others, social adjustment, physical traits, and self confidence.

Description: The Coopersmith Self-Esteem Inventory is comprised of three different forms. Form A consists of 58 statements, answered by checking "like me" or "unlike me." It provides a general assessment of self-esteem and can be broken down into five subscales. Form B consists of 25 items, and provides one self-esteem score, with no subscales. It takes half the administration time of Form A, and the two forms correlate highly (.86). Form C is like Form B, with 25 items and one self-esteem score, but is reworded to make it more appropriate for adults. Forms A and C also correlate highly (.80).

Administration and Scoring: The inventory is self-administered, takes 10 to 25 minutes depending on which form is used, and can be machine scored.

Reliability: A factor analysis discussed by Robinson and Shaver (1973) revealed four factors: self derogation, leadership-popularity, family-parents, and assertiveness anxiety. Split-half reliability tests on Form A resulted in coefficients of .87 to .90. According to Robinson and Shaver (1973) test-retest reliability is good--.88 over five weeks, .64 over one year, and .70 over three years.

Validity: Convergent validity tests with other measures related to self-esteem and self acceptance gave correlation coefficients between .42 and .66 (Robinson and Shaver 1973). Predictive validity is good, but high discriminant validity scores point to a problem with the inventory.

Criticism: This inventory fairly comprehensively covers self-concept and has the potential to measure sub-areas such as the family and social self-esteem (Robinson and Shaver 1973). A considerable amount of reliability and validity data have been gathered on a variety of sub-populations. One concern raised by Robinson and Shaver (1973) is that the scales correlate highly with social desirability. The inventory is a good candidate for evaluating prevention efforts directed toward self-esteem. Further field validation studies are recommended.

Suggestions for Use: This inventory measures attitudes towards the self in social, academic, family, and personal areas of experience. It was originally used in conjunction with teachers' behavioral ratings of self-esteem, but is now used as a general self-esteem measure.

References: Robinson, J. P., and Shaver, P. R. Measures of social psychological attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1973.

Locus of Control Scale for Children (p,o)

Author: S. Nowicki and B. R. Strickland

Ages: Caucasian children, grades 3 - 12

Date: 1972

Variables: This scale was designed to measure children's beliefs in external-internal control, as defined by Rotter. Internal control refers to regulation of events by personal behavior; external control refers to the belief that events are controlled by forces other than oneself. This scale attempts to assess the degree to which children believe they have control of their lives.

Description: The scale consists of a 40-item test (yes-no answers). Two short forms are available and preferred (Robinson and Shaver 1973). These versions are a collection of the best items from the total scale. Perceptions of control are ascertained in areas such as parental discipline, sports, scholastic success, social relations, and a general belief in luck.

Administration and Scoring: The test was originally administered orally, but has also been self-administered The entire scale should take about 15 minutes to complete, and the format is adaptable to manual or machine scoring.

Reliabilities: Split-half reliabilities range form .63 to .81 (Robinson and Shaver 1973), with older subjects yielding higher correlations. Test-retest reliabilities were lower, but the general level of consistency is acceptable. Feelings of internal control have been found to increase with age and have some tendency to correlate with intellectual achievement in males, yet they do not appear to be importantly related to social desirability of responses (Robinson and Shaver 1973).

Validity: Overall validity is fair: .41 with the Bailer-Cromwell scale, and .31 and .51 with the I+ scale of the Intellectual Achievement Responsibility Questionnaire (Robinson and Shaver 1973).

Criticism: While locus of control has a logical appeal, it may not necessarily correlate with substance abuse in comparison with factors such as peer pressure (Shute 1974).

Suggestions for Use: The Nowicki-Strickland scale is recommended for programs hoping to improve a client's feeling of control. According to Robinson and Shaver (1973), this may be the best test in the locus of control domain.

There were few minorities in the original sample and the evaluator should be aware of possible interactions between ethnicity, socioeconomic status, and test scores.

References: Robinson, J. P., and Shaver, P. R. Measures of social psychological attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1973.

Shute, R. E. Experimental effects of peer pressure on the verbally expressed drug attitudes of male college students. (Dissertation) University Park, Pennsylvania: The Pennsylvania State University, 1974.

Intellectual Achievement Responsibility Questionnaire (p,o)

Authors: Virginia Crandall, Walter Katkovsky, and Vaughn Crandall

Ages: All School Age

Date: 1965; 1968 (short form)

Source: Crandall, V. C., Katkovsky, W., and Crandall, V. J. Children's belief in their own control of reinforcement in intellectual - academic achievement situations. Child Development, 1965, 36, 91-109.

Variables: The Intellectual Achievement Responsibility Questionnaire (IARQ) differentiates the construct of locus of control in two important ways. First, it assumes situational specificity (i.e., school work). Second, it allows the possibility that perceived control of academic successes and academic failures are independent constructs. Hence, the IARQ provides subscores for both beliefs in internal responsibility for successes (I+) and in internal responsibility for failures (I-).

Description: The IARQ is composed of 34 achievement oriented situations, each accompanied by two explanatory statements. Respondents must choose the explanatory statement which best describes their experience. Half the items measure I+ and half measure I-.

Two short forms (20 items), one for elementary and one for secondary, are available, and correlate well (.88-.90) with the longer version.

Administration and Scoring: The IARQ has been administered both orally and in written form. Subscale scores are computed as the number of I+ and I- explanations endorsed.

Reliability: Test-retest reliability coefficients for 304 3rd through 5th graders were .66 for I+ and .74 for I-. Secondary school coefficients for students were .47 and I+ and .69 for I-.

Validity: A number of assessments of convergent and divergent validity are reported. Significant positive associations have been found between IARQ scores and grades, achievement tests, and intelligence tests. These choices of criteria measures are somewhat odd. The purported purpose of the IARQ is to measure taking of responsibility for academic success, not potential for success per se. Indeed, a high I- score requires endorsing a number of statements which admit academic failures.

Criticisms: Additional reliability estimates (e.g., Cronbach alpha) would be welcome. Also, as noted above, validity assessments are somewhat peculiar. Finally, moderate but significant associations have been found between I+ and I- scales for some secondary populations. Such a result calls into question the utility of considering I+ and I- as separate dimensions, at least for secondary school students.

Suggestions for Use: Although additional psychometric information is desirable, the IARQ is recommended for consideration any time locus of control is considered relevant to program goals and school aged youth are involved. Under such conditions it is to be preferred over the Norwicki and Strickland test, owing to its greater differentiation of the locus of control construct.

Note: A refined version of the IARQ will shortly be available from NIDA's NAPA Project (Moskowitz et al 1979).

References: Crandall, V. C. Refinement of the IARQ scale (NIMH progress report, Grant No. MH-02238). December, 1968, 60-67.

Crandall, V. C., Katkovsky, W., and Crandall, V. J. Children's belief in their own control of reinforcements in intellectual-academic achievement situations. Child Development, 1965, 36, 643-661.

Moskowitz, J. M., Condon, J. W., Brever, M., Schaps, E., and Malvin, J. The NAPA Project: Scaling of student self-report instruments (Progress report to NIDA). December, 1979.

Value Survey (p)

Author: Milton Rokeach

Ages: Twelve years - adult

Date: 1967-1973

Source: The Free Press
A Division of Macmillan Publishing Co., Inc.
866 Third Avenue
New York, NY 10022

Variables: The survey takes into account both long and short term changes in values, attitudes, and most importantly, behavior. Terminal and instrumental value differences are also measured within this framework.

Description: The primary focus of the survey is values, rather than attitudes (Rokeach 1973). It was designed specifically to concentrate its evaluation on a cognitive representation. This representation included not only individual needs, but also societal and institutional demands.

Administration and Scoring: Participants are given a two-page list of values (terminal and instrumental) which are placed in alphabetical order, and are asked to arrange items on each page in rank order of importance. The score for an item is its rank.

Sample: Rokeach (1973) includes rankings found from representative samples of adult Americans, as well as grades 7-16. In regard to subgroup differentiations, variables such as sex, income, education, race, age, politics, and religion are included. In addition, cross-cultural comparisons were examined in terms of values for American, Canadian, Australian, and Israeli college men.

Reliability: Test-retest reliabilities were taken and a median correlation of .65 was obtained for both terminal and instrumental values. However, these reliabilities were judged to be too low for satisfactory general use. Thus, some revision were made in the defining phrases of each value and the result was an increase from .65 to .69 (Rokeach 1973).

The data (Rokeach 1973) indicate that the terminal value reliabilities were consistently higher than the instrumental value reliabilities. This finding may suggest that the terminal list is more complete, and thus participants are more certain of their rankings. In Contrast to this finding, individual reliabilities varied considerably. In some instances they were as low as -.30 and as high as .90. These distinct differences may be due to the different sampling groups. In addition, there seems to be a great deal of fluctuation in value rankings cross-culturally (Rokeach 1973). Therefore, the reliability results of the Value Survey could be considered somewhat questionable despite the fact that most reliabilties were above .50.

Validity: Rokeach (1973) cites several studies which deal with value rankings. Since the individual differences in value system stability are somewhat vague, the validity of the Value Survey could be considered weak. Only sex, age, intellectual ability, and liberalism have both terminal and instrumental value stability. Moreover, even this finding is sometimes questioned. As a result, one might want to closely evaluate the validity of such a survey.

Criticisms: Both the validity and the reliability of the Value Survey appear to be somewhat questionable. There is a great deal of fluctuation between and among the different variables which are measured. Concerning the reliability of the survey, there were 86 out of 250 coefficients below .69 (Rokeach 1973). In addition, the various studies which have dealt with value rankings seem to find wide differences in thie conclusions.

Suggestions for Use: Given its low reliability and validity, it is not recommended as a measure of change. Even for descriptive purposes, it must be pointed out that it is only characteristice of an individual or group at one point in time.

References: Rokeach, M. The nature of human values. New York: The Free Press, 1973.

Tennessee Self Concept Scale (p,o)

Author: William H. Fitts

Ages: Twelve years and over

Date: 1964-65

Sources: William H. Fitts, Tennessee Self Concept Scale: Test Booklet 1964, and Manual 1965. Nashville, TN: Counselor Recordings and Tests, Department of Mental Health.

Description: The TSCS is made up of 100 self-descriptive statements, 90 of which are designed to measure self-concept and 10 of which are designed to measure self-criticism (Bentler's review in Buros 1972).

Administration and Scoring: On this test, which is self-administered, the subject marks each statement on a five-step scale ranging from "completely true" to "completely false."

Two different systems are available for scoring the test items: (1) Counseling Form is suggested as most appropriate for providing feedback for the individual, and (2) Clinical Research Form suggested as more applicable when the tester desires "research and clinical assessment" (Bentler's review in Buros 1972). Use of the Counseling Form provides 15 profiled scores: self-criticism, nine self-esteem scores, three variability of response scores, a distribution score, and a time score. Scoring the Clinical and Research Form provides 30 profiled scores: the 15 generated by the Counseling Form (see above) plus response bias, net conflict, total conflict, 6 empirical scales, deviant signs, and 5 scores each of which totals the number of responses for one of the 5 possible categories of responses to each item (Buros 1972).

Sample: 626 individuals of both sexes, aged 12 through 68, and differing in race and socioeconomic status (Bentler's review in Buros 1972).

Reliability: Two-week test-retest reliability coefficients for 60 college students: total self-regard, .92; rows, from .88 to .91; columns, from .85 to .90. These are sufficiently large to warrant confidence in individual difference measurement (Fitts 1965).

Validity: Information concerning validity between self-regard scores from TSCS and other alleged measures of self regard is not encouraging, and there is no evidence to indicate that the separate scores can be discriminantly interpreted. In fact, there may be overlapping among them (Fitts 1965).

Bentler (in Buros 1972) suggests that some of the TSCS's scores correlate highly with other scales that have been developed to measure "personality functioning." He cites the following examples: the Total Positive (TSCS) correlates -.70 with the Taylor Anxiety Scale. Correlations with the Cornell Medical Index and an unpublished "Inventory of Feelings" ranged from .50 to .70. Correlations with MMPI scales are often in the .50's and .60's. These overlaps between TSCS and other established measures could indicate that one might consider using TSCS as an alternative to these measures.

Criticisms: According to Bentler's review of TSCS (see Buros 1972), the scoring method is "cumbersome" and necessitates the use of an involved procedure. Information is lacking concerning the internal consistency of the TSCS scale (Bentler in Buros 1972). Subscores are non-independent (Robinson and Shaver 1973). In addition, Bentler (1972) suggests that while intercorrelations among subscores are presented, the author fails to report any principal components analysis or factor analysis, thus making it impossible to assess the number of different dimensions of self-concept that are being measured by the scale. Bentler posits at most the existence of only two or three different

dimensions. Thus, although the manual describes the test as "multi-dimensional," no data are provided to substantiate this.

Suggestions for Use: The test can be used to differentiate between normal and psychiatric patients and between normals and nonnormals. According to Suinn (in Buros 1972): "the TSCS ranks among the better measures combining group discrimination with self-concept information. The Empirical Scales are useful as a means of screening clients for pathology, while some of the other scales seem to add some intuitive data about self perceptions."

References:

Bentler, M. Review of TSCS. In Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Fitts, W. H. Tennessee Self Concept Scale: Manual. Nashville, TN: Counselor Recordings and Tests, Department of Mental Health 1965.

Robinson, J. P., and Shaver, P. R. Measures of social psychological attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1973.

Suinn, R. M. Review of TSCS. In Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Personal Orientation Inventory (p,o)

Author: Everett L. Shostrom

Ages: Grades 9 - 16 and adults

Date: 1962 - 68

Source: Educational and Industrial Testing Service
San Diego, California 92107

Variables: The POI seeks to measure self-actualization as defined by Maslow (1968). The instrument consists of two major scales: an Inner Support Scale and a Time Competence Scale. The Inner Support Scale attempts to measure the degree to which the respondent tends to act on his/her own principals rather than reacting to outside influences. This scale is composed of five pairs of related, but contrasting, variables which when scored produce ten subscales. The Time Competence Scale attempts to measure the degree to which the respondent lives mostly in the present as opposed to dwelling on past events or being anxious about the future (Bloxom's review in Buros 1972).

Description: Persons taking the test complete 150 comparative value judgement items. They are asked to choose which of two opposing values is closer to what they hold to be true for themselves. The following sample test questions (items 52 and 72) serve to illustrate the format and type of forced-choice items included in the test:

a. I am afraid to be angry at those I love.
b. I feel free to be angry at those I love.

a. I accept inconsistencies within myself.
b. I cannot accept inconsistencies within myself.

Administration and Scoring: Thirty minutes are usually allowed for completing the test. Results can be either hand or computer scored.

Reliability: Reliability coefficients range from a "moderate" .55 to a "good" .85 (Bloxom's review in Buros 1972). According to Robinson and Shaver (1973), there have been no studies of POI's internal consistency.

Validity: According to Bloxom (see Buros 1972) the content validity of the POI scales is "good." Bloxom cites that in five of six studies which administered POI to patients and nonpatient controls, patients' scores increased more from pre-therapy to posttherapy than did those of the controls. Coan (see Buros 1972) also notes that subjects who have been designated as "relatively self actualized" by clinical psychologists tend to have higher scores on POI, and that the scores seem to rise following psychotherapy.

Criticism: According to Bloxom (1972), the manual notes that the various subscales contain overlapping items and that their independence has not been established. Reliance on the subscales could result in overinterpretation (Robinson and Shaver 1973).

According to Coan (see Buros 1972), there appears to be a "bias in favor of extraversion," with emphasis on "overt expression" as opposed to "inner experience." He suggests that the choice of variables is theoretically biased and that the instrument's items are not very sophisticated.

Suggestions for Use: The appeal of self actualization may be strong in prevention programs, but this scale should be used with caution. Pilot testing for process evaluation should be conducted prior to use as an outcome measure.

References: Bloxom, B. Review of POI. In Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Coan, R. W. Review of POI. In Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Maslow, A. Toward a psychology of being. Princeton, New Jersey: Van Nostrand, 1968.

Robinson, J. P., and Shaver, P. R. Measures of social psychological attitudes. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1973.

INTERPERSONAL SCALES

Group Dimensions Descriptions Questionnaire (p)

Author: John K. Hemphill

Ages: Twelve years - adult

Date: 1956.

Sources: The Ohio State University Press
Irene Martin
Columbus, Ohio 43201
Phone: 614-422-6446
614-422-6930

Variables: Characteristics of groups in various situations are studied. The respondents to the questionnaire express their attitudes, perceptions, and impressions or knowledge about specific groups in which they are involved (Pfeiffer and Heslin 1973).

Description: There are 150 statements on the descriptions questionnaire which are designed to specify particular group dimensions. These group dimensions are then arranged into 13 categories which are intended to indicate the degree to which group members perceive characteristics to be true or false. The resulting data were used to develop the following individual categories:

(1) Control--manner in which a group regulates the behavior of individuals during their group functioning.

(2) Stability--group characteristics and dynamics over a period of time.

(3) Intimacy--extent to which group members are knowledgable about each other.

(4) Stratification--degree to which the group is structured in terms of power and influence.

(5) Hedonic Tone--extent of congeniality within a group.

(6) Autonomy--degree to which a group has an independent position from other groups in terms of decision making and activities.

(7) Potency--importance of a group to its members.

(8) Viscidity--extent to which group members function as a total.

(9) Permeability--openness to new members.

(10) Participation--amount of commitment in time and effort by the members.

(11) Polarization--degree to which a group works toward a unitary end.

(12) Flexibility--degree to which informal rather than formal processes are followed.

(13) Homogeneity--degree of similarity in socially relevant characteristics among group members (Pfeiffer and Heslin, 1973).

Administration and Scoring: The administration of the descriptions questionnaire takes about 15 to 20 minutes. 150 statements are answered in degrees, from definitely true to definitely false. Each of these responses is weighted. The scoring takes about ten minutes. Normalized scores can be obtained by converting the row scores on 13 dimensions.

If sufficient scoring keys are available, group members can score the questionnaires for more immediate feedback. In this way, the instrument can be incorporated into the process of the group.

Reliability and Validity: This instrument has been shown to have adequate reliability for process evaluation. Additionally, there is evidence of similarities in group satisfaction and the dimension scores which respondents may give. However, the information in these areas concerning specific statistics must be obtained from the manual (Hemphill 1956).

Criticisms: There appears to be some overlap in the subscales (for example, intimacy correlates .41 with homogentiy, Hemphill 1956).

Suggestions for Use: The questionnaire could be used for both organized and natural groups, to give members some indication of the strengths and weaknesses of the group. In addition, the questionnaire might enable group leaders to evaluate the quality of group participation.

References:

Hemphill, J. K. Group dimensions: A manual for their measurement. Columbus, Ohio: The Ohio State University Press, 1956.

Pfeiffer, J. W., and Heslin, R. Instrumentation in human relations training: A guide to 75 instruments with application to the behavioral sciences. Iowa City, Iowa: University Associates, 1973.

Russell Sage Social Relations Test (p,o)

Authors: Dora E. Damrin, William E. Coffman, and William A. Jenkins

Ages: Children, grades 4 - 8

Source: Educational Testing Service, Princeton, New Jersey 08540 (publisher) and Hillcraft Industries, Route 3, Traverse City, Michigan 49684 (for purchase of building blocks used in test)

Variables: The test was designed to evaluate group problem solving skill: "It was constructed to assess skills, knowledge, and behavior patterns relevant to group planning and action" (Lake, Miles, and Earle 1973). Traits which are assessed include the respondent's ability to be a leader at times and a group member at others, the ability to initiate suggestions, to listen to the suggestions of others, and to accept group decisions (Ibid.). The original design of the test has been modified.

Variables measured in the planning stage include participation, involvement, communication, autonomy, organizational techniques, and final plan. Variables measured in the operations stage include involvement, atmosphere, activity, and success. Variables which limit performance are also considered.

Description: All children in a classroom are instructed to work as a group to build three structures in ascending order of difficulty. Building materials include 36 notched, interlocking plastic blocks (Lake, Miles, and Earle 1973).

Administration and Scoring: An observer records the group work in two stages--the planning stage and operations stage--according to whether a majority or minority of the group is engaged in the particular variables being measured. Fifteen minutes are allowed for construction of each figure. Additional time is allowed for planning and approximately one hour total time is required.

Behavior is coded into content divisions. The examiner then extracts a picture of behavioral trends in group functioning over the specified time period (Lake, Miles, and Earle 1973).

Reliability: According to Lake, Miles, and Earle (1973), there are no test-retest data available on this instrument. They cite the authors' assertion that the instrument measures "dynamic variables that are subject to change over time" and that therefore test-retest methods are not appropriate. Part-time observer reliability had to reach .75 before Damrin considered training of her scorers acceptable.

Validity: Authors apparently did not consider construct and predictive validity to be relevant to this instrument. Analogous scores were obtained for both adult and children's groups, both skilled and unskilled in group planning.

Criticism: Lake, Miles, and Earle (1973) note the lack of investigation of reliability and validity and suggest that until this is provided, the instrument should not be used for evaluation of training or group interaction. The rating system is complex, and requires trained staff to administer, observe, and score the interactions, a factor which may limit use of the test.

Suggestions for Use: This approach to evaluation enables the investigator to assess interaction in a direct manner. However, knowledge of group processes and training in observation and scoring is necessary for correct use of the instrument.

References: Lake, D. G., Miles, M. B., and Earle, R. B., Jr. (Eds.). Measuring human behavior. New York and London: Teachers College Press, 1973.

Chapin Social Insight Test (p,o)

Author: F. Stuart Chapin (test) and Harrison G. Gough (manual)

Ages: Thirteen years and over

Date: 1967 - 68 (originally developed by Chapin in 1942)

Source: Consulting Psychologists Press
577 College Avenue
Palo Alto, California 94306

Variables: The Social Insight Test is designed "to assess the perceptiveness and accuracy with which an individual can appraise others and forecast what they might say and do" (Gough 1968).

Description: The test is composed of 25 "problem situations" in interpersonal relations or personality dynamics. Subjects are asked to select the one of the four given options which best describes the most likely reason for the behavior described or consequences it will have (Lanyon's review in Buros 1972).

Administration and Scoring: The test is self-administered and requires approximately 30 minutes to complete, although there is no time limit. Items are scored with differential weights of one, two, or three. Scoring is done by hand.

Reliability: According to Lanyon's review (1972), internal consistency reliabilities for this test are in the .68 to .78 range. Orr (1972) cites the reliability results for several different groups--a sample of 100 males (corrected odd/even coefficient of .78), a sample of 494 males (item/test correlations as projected by the Guilford method of .71), and a sample of 215 females (item/test correlation of .68). Test-retest reliabilities were not computed according to the test's manual.

Validity: Validity coefficients provided by the manual are generally low. Lanyon (1972) points out that "correlations with the scales of the CPI, the MMPI, and the SV are low, with very few correlations reaching .30." In a comparison of the test with psychologists' ratings, there were "modest correlations" with such qualities as good judgment and ability to communicate (Lanyon, 1972). Q-sort data reveal that the items which are most highly related to scores on the CSIT include "ability for getting the cooperation of others, and for being both a good leader and a good listener, and responsiveness to the subtleties of other people's behavior" (Lanyon, 1972). When mean scores were computed for a variety of occupational groups, the orderings were "reasonable in terms of the social insight which might be expected to be required" (Orr, 1972).

Lanyon (1972) suggests that low validity scores may result because the CSIT appears to be assessing "a relatively specific personality or social attribute." Orr (1972) suggests that rather than dealing with a unitary dimension, the SCIT may be assessing a multifaceted, heterogeneous concept.

Criticism: Lanyon (1972) cites what he considers several "minor" points: the correct response options for some of the questions appear rather "arbitrary;" and the wording of some of the questions and response options seem "outdated." While the manual includes means and standard deviations from a number of sample groups, there are no norms provided in percentile or standard score form (Lanyon, 1972).

Lake, Miles, and Earle (1973) point out the possibility that the items simply measure "the subject's learned ability to 'psychologize' about interpersonal relations, or to apply conventional categories of explanation." Orr (1972) suggests that the reading load demanded by the CSIT needs to be reduced. He also suggests that while the test's reliability is too low for individual work, it is probably acceptable for group work.

Suggestions for Use: The CSIT has been cited as "the most promising available instrument for assessing social insight" (Lanyon, 1972). Orr (1972) states that it represents "an interesting and useful attempt to measure potential interpersonal behavior as it might occur in real-life situations." While the instrument is not recommended for individual or clinical use, it is deemed appropriate for experimental purposes.

References:

Gough, H. G. The Chapin Social Insight Test: Manual. Palo Alto, California: Consulting Psychologists Press, 1968.

Lake, D. G., Miles, M. B., and Earle, R. B., Jr. Measuring human behavior. New York and London: Teachers College Press, Teachers College, Columbia University, 1973.

Lanyon, R. I. Review of CSIT. In Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Orr, D. B. Review of CSIT. In Buros, O. K. The seventh mental measurements yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Family Environment Scale (p,o,i)

Author: Rudolf H. Moos

Ages: Families--no age limits or reading levels suggested

Date: 1974

Source: Consulting Psychologists Press, Inc.
577 College Avenue
Palo Alto, CA 94306

Variables: Based on Murray's theory of environmental press, this instrument is designed to evaluate several key variables of family life. The three major scales of the instrument reflect interpersonal relationships within the family, nature and direction of personal growth emphasized within the family, and the organizational structure of the family unit.

Description: The final form of the instrument, Form R, contains 90 true-false items, nine items measuring "press" toward each of the ten subscale variables. Questions measuring the relationship dimension evaluate press within the family toward cohesion, expressiveness, and conflict. The personal growth scale evaluates press toward independence, achievement orientation, intellectual-cultural orientation, active-recreational orientation, and moral-religious emphasis. The final major scale, system maintenance, evaluates press toward organization and control within the family unit. A thorough description of the instrument subscales can be found in the instrument manual (Moos, Insel, and Humphrey 1974).

Form R, "real," is designed to measure family members' perception of what the family is actually like. A parallel form, Form I, "ideal," is available to evaluate family members' perceptions of what the family should be like. There is also a short form of this measure (Form S) and an Expectations Form (Form E) (Moos, Insel, and Humphrey 1974).

Administration and Scoring: The scale can easily be administered to individuals or groups. Test booklets are reusable and answers to true-false questions are marked on an answer sheet which is readily scored using a plastic template. It is not a timed measure; however, completion of the measure should require less than 30 minutes. Scores are computed for each of the ten subscales and a family incongruence score. This score reflects the degree of agreement between family members on characteristics of the family.

Reliability: Test-retest reliability (over an eight week period) ranged from .68 to .83. Internal consistency coefficients are between .64 and .79. Item to subscale correlations are between .45 and .58. The reliability of the scale appears adequate.

Validity: Validity information is lacking.

Criticisms: Issues related to the reliability and validity of the instrument have not been thoroughly evaluated. Reliability information, although encouraging, is based on a relatively small number of subjects. Reliability and validity information should be collected as part of a pilot project.

Suggestions for Use: The scale may be an acceptable measure of family environment. Family incongruence scores, development of a family profile, as well as differences between Real and Ideal Scores (Forms R and I) could prove to be useful information in family counseling. The instrument also has potential for research studies. For example, Moos, Insel, and Humphrey (1974) have collected information relating patterns of alcohol use to various characteristics of the family as measured by the FES.

Reference: Moos, R. H., Insel, P., and Humphrey, B. Combined preliminary manual: Family, work and group environment scales. Palo Alto, California: Consulting Psychologists Press, Inc., 1974.

The Work Environment Scale (p,o,i)

Authors: Rudolf H. Moos and Paul M. Insel

Ages: Adult

Date: 1974

Variables: Based on environmental press theory, this instrument is designed to measure the social environment of work settings. The three major scales of the instrument measure relationships within the organization, the degree of emphasis on personal growth in the work setting, and system maintenance or basic system organization and characteristics.

Description: There are various forms of this instrument. Form R is the most widely used, and contains 90 true-false items, nine items measuring press toward each of ten subscale variables. Relationship dimension items measure press toward involvement, peer cohesion, and staff support. Items in the personal growth dimension measure press toward autonomy and task orientation. System maintenance and system change dimensions measure press toward work pressure, clarity, control, innovation, and physical comfort. A more complete description of the instrument subscales can be found in the WES manual (Moos, Insel, and Humphrey 1974). Form R "real" measures what employees believe their work environment is actually like. Form I "ideal" measures what the employees would like their work environment to be. Form E (expectations) assesses what the employee expects the work setting to be like. A short form of the instrument (Form S) is also available.

Administration and Scoring: The scale can be easily administered to individuals or groups. Forms R, I, and E are parallel and can be scored using the same scoring template. Test booklets are reusable and answers to true-false questions are marked on an answer sheet which is readily scored using a plastic template. This is not a timed measure, but completion of the 90 item forms should be possible in less than 30 minutes.

Raw or standard scores are obtained for each of the ten subscales, which result in descriptive profiles for various employee groups and allow for comparisons between work units or personnel groups.

Reliability: Internal consistency ranged from .70 to .91. Item to subscale correlations varied from .48 to .63. The reliability of this instrument appears adequate.

Validity: Moos, Insel, and Humphrey (1974) state that the Form R subscale intercorrelation average of .25 indicates that the subscales measure different but related characteristics of the work setting. It would appear that adequate validity information is lacking. Statistics relating to the alternate forms of the instrument are available in the WES Manual (Moos, Insel, and Humphrey 1974).

Criticisms: The validity of the instrument has not been clearly supported.

Suggestions for Use: The WES could be used as a process instrument for determining staff dynamics within an agency. Differences between personnel units can be compared. The instrument has potential for use in varied work settings.

References: Moos, R., Insel, P, and Humphrey, B. Combined preliminary manual: Family, work and group environment scales. Palo Alto, California: Consulting Psychologists Press, Inc., 1974.

Evaluation of Classroom Climate (p,o)

Author: John Withall

Ages: Classroom students of any age

Date: March 1969

Sources: Association of Childhood Education International
3615 Wisconsin Avenue, N. W.
Washington, DC 20016
or
University Microfilms
A Xerox Company
P.O. Box 1346
Ann Arbor, Michigan 48106

Variables Measured: There are specific clues which indicate behaviors, interactions, and confrontations within the classroom climate. Some of the variables which focus on these behaviors include the professional stand which is taken by the classroom teacher. The teachers' verbalizations are divided into two major subgroups. The behaviors are viewed as either learner or teacher oriented in context. Within the seven individual categories are behavior differences, including commendatory, acceptance, problem structuring, neutral directive, reproving, and teacher supportive behaviors (Withall 1969).

Description: The primary concern of the instrument is with the socio-emotional interactive behaviors between the teacher and his/her students. More specifically, the affective tone that accompanies communications between individuals in groups is of prime importance (Withall 1969). Therefore, the evaluation of classroom climate focuses on clues which indicate general emotional factors. It appears as though climate probably affects the degree of freedom, as well as spontaneity and range of roles available to each individual within the limits set by the group.

The first category is learner supportive statements or questions. The major intent of these statements is to encourage, bolster, and praise the learner. The second category is acceptance or clarifying statements or questions. The intent of these responses is to help the learners gain insight into their problem. The third is problem-structuring statements or questions. These statements may enable teachers to further increase their understanding of what the learner has said. The fourth category includes neutral statements which evidence no supportive intent. These statements may simply be repetitions of statements that the learner has just made. The fifth includes directive statements or questions which persuade the learner to take the teacher's point of view. The sixth category is reproving, disapproving, or disparaging statements or questons. These statements intend to chastise the learners for unacceptable behavior and prevent them from repeating the behaviors. The last category is teacher-supportive statements or questions. These statements reassure the teacher of his/her position (Withall 1969).

Administration and Scoring: As assessment of the classroom climate index is computed by dividing the total number of statements that fall into categories one, two, and three by the number of statements that fall into categories five, six and seven. Using this climate index, one can determine whether the atmosphere is liberating for the learners and enhances their inquiry, testing, and coping skills. Likewise, the index can determine whether a climate is hindering learning and coping activities.

Reliability and Validity: Withall reported that one can reliably and validly categorize the teachers' verbal behaviors through the use of the Classroom Climate Index. In fact, there is research evidence which concludes that the social-emotional climate in any teaching situation is related to the quality of the problem solving, inquiring, and coping activities of the learners. However, substantiating data supporting these statements are lacking in Withall's Evaluation of Classroom Climate. There is no indication that reliability and validity coefficients were analyzed in this particular instrument.

Criticisms: Although Withall reports that the Classroom Climate Index is both valid and reliable, more substantial evidence supporting this claim is necessary.

Suggestions for Use: The Classroom Climate Index could be used with caution in some teaching-learning situation.

References: Withall, J. Evaluation of classroom climate. Childhood Education, 1969, 45, 403-408.

Hill Interaction Matrix - Group (p)

Author: William Fawcett Hill

Ages: Six year - adult

Date: 1963

Sources: Youth Studies Center
University of Southern California
Los Angeles, California 90007
Phone: (213) 746-6292

Variables: The variables measured in the HIM-G include group interaction content, group interaction style, and therapist activity. The content dimension includes topic, group, personal, or relationship variances, while the style dimension includes the categories of responsive, conventional, assertive, speculative, and confrontive behaviors (Pfeiffer and Neslin 1973; Lake, Miles, and Earle 1973).

Description: The Hill Interaction Matrix-Group focuses on group and leader behavior, rather than on the style of individual members. The HIM-G is intended to measure group behavior. Its processing is faster then the content analysis method which was the original purpose of the matrix (Pfeiffer and Neslin 1973).

The matrix consists of 72 descriptions of group behaviors. An observer, leader, or member of a particular group rates the group on each item. Four statements are given within each cell of the matrix. These statements describe given behaviors with four different emphases: trainer sponsored behavior, trainer encouraged behavior, member behavior (number of members), and proportion of time needed to fulfill group needs. The rating scale ranges from "not-at-all" to 40-100 percent of the time for "frequency of member participation." Similarly, statements coded as "member participating" have a rating scale ranging from no members to seven or more members. An example of the type of statement which might be included within a cell matrix is, "Members express negative or hostile feelings or delusional ideas about certain conditions, institutions, or events" (Pfeiffer and Neslin 1973).

Administration and Scoring: The administration of the HIM-G takes roughly 20 minutes, while the scoring takes only 10 to 15 minutes. The key for the matrix is in the supplement to the HIM Monograph. A number of indices can be constructed from the matrix by recording member silence, amount of leader participation, risks taken, and the therapist's activity. In the actual scoring it is necessary to convert cell, column, and totals to percentages. This can be achieved by dividing each total in the columns by the "Overall Total" (Pfeiffer and Neslin 1973).

Reliability and Validity: Primary statistics on reliability and validity are available through the Hill Interaction Matrix Monograph. The validity correlations for the HIM-G are, however, recorded as being over .90, with scores from the content analysis system. This figure indicates a relatively high degree of validity for purposes of implementing the matrix. However, reliability testing of the scale is lacking so that even the high validity coefficient may be somewhat questionable.

Criticisms: The manual to the Hill Interaction Matrix is somewhat confusing. This can pose problems in scoring the matrix as well as interpreting it. In addition, there is an absence of normative data. This phenomenon is acknowledged as being a problem; however, there is some indication that the matrix has validity. There is also a lack of available information concerning the reliability of the instrument.

Suggestions for Use: The Hill Interaction Matrix-G appears to have several possible uses. It can be used to sensitize members to their transactions by having the leader or some member give feedback to the group every few minutes. In addition, the matrix enables both leaders and members to complete a scale after selected meetings or at the end of a laboratory.

Possible uses for the HIM-G might include determining staff dynamics within an agency. It might also determine the content of interaction between staff and participants.

References:

Lake, D. G., Miles, M. B., and Earle, R. B., Jr. (Eds.). Measuring human behavior. New York: Teachers College Press, 1973.

Pfeiffer, J. W., and Neslin, R. Instrumentation in human relations training: A guide to 75 instruments with application to the behavioral sciences. Iowa City, Iowa, University Associates, 1973.

Reactions to Group Situations Test (p)

Author: Herbert A. Thelen and Dorothy Stock Whitaker

Ages: Twelve years - adult

Date: 1967

Sources: John Wiley and Sons, Inc.
605 Third Avenue
New York, New York 10016
Phone: (212) 867-9800

Variables: Aspects of individual behavior in group settings are measured by this test. There are five scores which indicate preferences for each of five kinds of behaviors. The five include inquiry mode or work mode, fight mode, pairing mode, dependency mode, and flight mode.

Description: The Reactions to Group Situations Test is useful in sensitizing participants to important dimensions of group relations. It is a way of introducing them to certain assumptions concerning people in therapy and similar groups.

According to Pfeiffer and Neslin (1973), the inquiry mode focuses on task oriented behavior, group oriented responses, and problem solving orientation. The indicated preference in the fight mode is an angry response. The pairing mode is designed to see if members support another person's idea. It also looks at an individual's expression of intimacy, warmth, and commitment towards the whole group. In the dependency mode, preferences are for support and direction with reliance on "a definite structure, rules, regulations, reliance on leader or on outside authority, and expressions of weakness or inadequacy." Preference for the flight mode includes withdrawal or lessened involvement, joking, fantasy, daydreaming, inappropriate theorizing, generalizations, irrelevancy, and excess activity in busywork.

Administration and Scoring: The instrument takes 10 to 15 minutes to administer. Statements are made and two answers are given; the respondents are to answer in the manner which they feel is most similar to their own behavior. Their answers indicate which mode they may be most closely related to.

The scoring of the reactions to group situations takes about 12 to 18 minutes. According to Pfeiffer and Neslin (1973), the instrument is "pleasant to take and short and easy to score."

Reliability and Validity: There is no indication that reliability and validity testing has been completed for the Reactions to Group Situations Test.

Criticisms: Both the validity and reliability of Reactions to Group Situations are unclear.

Suggestions for Use: This instrument could be a useful way of measuring important relationships within a group setting. Appropriate settings might include any type of group gathering including academic, therapeutic, and social situations. However, researchers are urged to conduct pilot tests of reliability and validity before using this scale.

References:

Pfeiffer, J. W., and Neslin, R. Instrumentation in human relations training: A guide to 75 instruments with application to the behavioral sciences. Iowa City, Iowa: University Associates, 1973.

Youth Perception Inventory (p,o,i)

Author: Fred Streit

Ages: Early Adolescents

Date: 1977

Source: Fred Streit Associates
168 Woodbridge Avenue
Highland Park, N.J. 08904

Variables: The Youth Perception Inventory (YPI) assesses the respondent's perception of his/her parents' behavior. The YPI measures 26 concepts representing eight dimensions as described below (Streit 1978).

Dimensions	Concepts
Autonomy	Extreme autonomy, lax discipline
Autonomy and Love	Moderate autonomy, encouraging sociability, encouraging independent thinking, equalitarian treatment.
Love	Positive evaluation, sharing, expression of affection, emotional support.
Love and Control	Intellectual stimulation, child-centeredness, possessiveness, protectiveness.
Control	Instrusiveness, suppression of aggression, control through guilt, parental direction.
Control and Hostility	Strictness, punishment, nagging.
Hostility	Irritability, negative evaluation, rejection.
Hostility and Autonomy	Neglect, ignoring.

Description: The YPI is made up of 108 statements which the respondent indicates are true of both parents, true of mother or father only, or not true of either parent.

Administration and Scoring: The YPI is self-administered and hand scored with score sheets available from the distributor. Computation and analysis (Means, standard deviations and t-tests) of derived "substance abuse proclivity" scores are available from the distributor on a for-fee basis (proclivity scores will not be provided for individual cases).

Reliability: A split-half reliability coefficient of .91 is reported for the data from mother and father together (Streit 1978). The size and characteristics of the population upon which this coefficient is based are not given.

Validity: Concurrent validity data suggest the ability of the YPI to "predict" self-reported substance use (Streit 1978). In one such study, phi coefficients (derived from X^2 analysis of correct vs. incorrect identification of users) ranged from .26 to .86 with the majority being under .50. Significance levels for the X^2 statistics are not given, nor are n's.

Criticisms: Although the author refers to the "predictive validity" of the YPI, no prediction was involved in the studies reported (a fact the author himself notes).

Streit's (1978) attempt to infer predictive validity for the YPI based on evidence concerning the Strong Vocational Interest Blank is weak as best.

In addition, further reliability information would be welcomed.

Suggestions for Use: The YPI should be considered by programs whose objectives include altering participants' perceptions of their parents. Additional psychometric information is highly desirable.

Reference: Streit Fred. Technical manual: Youth Perception Inventory. Highland Park, New Jersey: Fred Streit Associates, 1978.

SUBSTANCE SCALES

National Survey on Drug Abuse: 1977 (o,i)

Authors: Hebert L. Abelson, Patricia M. Fishbourne, and Ira Cisin

Ages: 12 years - Adult

Date: 1977

Source: Abelson, H. I., Fishburne, P. M., and Cisin, I. National Survey on Drug Abuse: 1977. Rockville, MD: National Institute on Drug Abuse, 1977.

Variables: The instruments from the National Survey on drug abuse measure a number of variables including use of licit and illicit drugs and attitudes towards and beliefs about marijuana.

Description: The National Survey on Drug Abuse instruments consists of two interview schedules. Both forms may be used with adults and youths and both assess use of marijuana, hashish, inhalents, cocaine, hallucinogens, opiates other than heroin, heroin, alcohol, tobacco, and caffeine. In addition, both forms assess attitudes towards and beliefs about marijuana. The forms differ in that one contains questions on the nonmedical use of psychotherapeutic drugs (both prescription and nonprescription), while the other contains questions on friends' use of heroin.

Adminstration and Scoring: The National Survey on Drug Abuse interview schedules are designed to be administered in a face to face interview by a trained interviewer. A number of interesting devices are employed to increase the validity of self-reported drug use. One such device involves giving respondents pictures of psychotherapeutic drugs when questioning them about the use of these substances. The pictures allow respondents to identify substances which they have taken, but know by a local street name. Another such device involves having respondents answer the interviewer's questions on answer sheets which the interviewer cannot see.

No scales are derived and scoring is direct.

Reliability: No information given.

Validity: As noted above, a number of devices are built into the National Survey on Drug Abuse instruments which minimize respondents perception of risk. While these devices may be expected to increase the validity of self-reported substance use, no independent assessments of validity are reported in the Abelson et al (1977) report. However, a known groups validity study, done as a pilot test, provided supportive data.

Criticisms: Some programs may find these interview instruments overly long and cumbersome. In addition, it is likely that a limited number of programs will require all the information collected. Adequate psychometric information is lacking.

Suggestions for Use: Because of the wealth of information provided, the National Survey on Drug Abuse instruments seem particularly well suited to impact evaluations.

Programs using these instruments will need to conduct thorough pretests in order to obtain reliability and validity estimates. Pretesting is especially important if some of the questions are deleted.

References: Abelson, H. I., Fishburne, P. M., and Cisin, I. National survey on drug abuse: 1977. Rockville, Maryland: National Institute on Drug Abuse, 1977.

Senior Survey (o,i)

Authors: Lloyd D. Johnston, Jerald G. Bachman, and Patrick M. O'Mally

Age: 17-23 years

Date: 1979

Source: Institute for Social Research
University of Michigan, Ann Arbor, Michigan 48109

Variables: The Senior Survey measures variables in six main categories. The categories and variables are given below.

(1) Drug Usage Variables:

- Cigarettes
- Alcohol
- Marijuana/Hashish
- Hallucinogens
- Cocaine
- Stimulants
- Sedatives
- Tranquilizers
- Heroin
- Other Opiates
- Inhalents
- Marijuana Only/Annual Prevalence
- Illicit Drug Use (Other than Marihuana)/Annual Prevalence

- Probability of Future Use of Drugs
- Grade of First Use of Drugs
- Degree and Duration of Feeling High

(2) Background and Demographic Variables:

- Sex
- College Plans
- Region
- Population Density

(3) Attitude and Belief Measures:

- Perceived Harmfulness of Drugs
- Disapproval of Drug Use
- Attitudes Regarding Legality of Drug Use
- Attitudes Regarding Marijuana Laws

(4) Attitudes and Beliefs of Parents and Friends:

- Parents' Disapproval of Drug Use
- Friends' Disapproval of Drug Use

(5) Exposure to Drug Use:

- Exposure to Drug Use
- Friends' Use of Drugs

(6) Perceived Availability of Drugs

Description: The Senior Survey consists of five forms, each containing about 400 items (a given respondent completes only one form). Forms two through five have a

common core section (Part B) containing key drug and alcohol use questions. Part A of each form adresses a different set of attitudenal, behavioral, or demographic variables. Form one has slightly different, but comparable questions.

Taken together the five forms provide an extremely complete profile of the sample.

Administration and Scoring: The Senior Survey is administered in groups and takes respondents about 45 minutes to complete. When originally used by the Institute for Social Research, the Survey was completed in pencil for optical scanning and computer scoring.

Reliability and Validity: Johnston, Bachman, and O'Malley (1979) provide no reliability or validity data; however, a reasonably persuasive argument for the general validity of scales of this type is made.

Criticisms: Because five forms are required for complete information, larger samples are needed to provide stable data. This criticism obviously does not apply to Part B of forms two through five.

Suggestions for Use: The completeness of data provided by the Senior Survey makes it especially well suited to impact evaluation. With proper pretesting, portions of the complete survey will be useful at all levels of prevention evaluation.

References: Johnston, L., Backman, J., and O'Malley, P. Drugs and the class of '78: Behavior, attitudes, and recent national trends. Rockville, Maryland. National Institute on Drug Abuse, 1979.

Drug Education Evaluation Scales Part III: Incidence (p,o,i)

Author: John D. Swisher and John J. Horan

Ages: Twelve years and older

Date: 1973

Source: Abrams, A., Garfield, E., and Swisher, J. Accountability in drug education: model for evaluation. Washington, D.C.: Drug Abuse Council, 1974.

Variables: The purpose of the Drug Education Evaluation Scales is to assess the extent of drug abuse in a given area, and/or determine the behavioral effects of experimentally oriented drug programs. It evolved out of a number of projects attempting to determine the effects of various approaches to drug education. Essentially, it is an inventory for assessing the extent of drug use behavior.

Description: The format of the test lists substances in a column down the right hand side of the page and lists extent of use options in a row across the top of the page. Additional substances may be added to the scale, which allows for adaptation to a particular setting. The advantage to the scale is that it has a very simple format, extensive flexibility, and it is easily answered.

Administration and Scoring: It is strongly recommended that this scale be given anonymously and in small groups. The examiner should be someone whom the students trust, such as a school counselor or a respected peer. This scale has typically been included last in a set of two or three other instruments (for example, an attitude scale and an alternatives scale). A use score can be derived by simply totaling the numbers checked.

If a program's objective is to reduce the use of the most dangerous drugs, then it would be appropriate to consider weighting the more dangerous drugs.

Reliability: No reliability data are available at this time. However, the scale includes a dummy drug which provides one form of consistency check. It is also advisable to ask subjects (on separate forms) about the relative honesty of their answers.

Validity: No validity studies have been conducted. The scale does correlate with attitudes toward drugs at approximately .65 and with peer use of drugs at .64 (Warner and Swisher 1976).

Criticism: The instrument is only useful when subjects cooperate fully, since self-incriminating durg use data is requested. Moreover, reliability and validity information is not available.

Suggestions for Use: The behavior subtest of the Drug Education Evaluation Scales has been used in planning and evaluating a number of drug education projects. Additional psychometric data are needed to assess the reliability and validity of this instrument.

References:

Abrams, A., Garfield, E., and Swisher, J. Accountability in drug education: Model for evaluation. Washington, D.C.: Drug Abuse Council, 1974.

Warner, R. W., and Swisher, J. D. Alienation and drug abuse: Synonymous? National Association of Secondary School Prinicpals Bulletin, 1976, 53, 55-62.

Drug Use Checklist (p,o,i)

Author: Roy Cook

Ages: Nine years and older

Date: 1975

Sources: Roy Cook
Institute for Social Analysis
11800 Sunrise Valley Drive
Reston, Virginia 22091

Variables: Six categories of frequency of use for 12 substances.

Description: This measure elicits information about recent drug use, is anonymous, quick to complete, and requires a minimum of instructions. It contains a bogus drug item as well as items about caffeine, alcohol, and tobacco.

The Drug Use Checklist was constructed after reviewing a variety of drug use measures (Hays' Questionnaire, Althoff's Drug Use Scale, the Penn State Scale, and a scale developed for use in the military) to meet the above requirements. It is similar to a checklist used in military research which proved reliable (test-retest estimates) and valid (compared to other indices, including urinalysis data) in a world-wide drug prevalence survey.

Administration and Scoring: The questionnaire can be reproduced on one page, and each individual only answers one question per substance. Scoring is simply a matter of tallying extent of use for each substance, or summing the tallies for all substances. It is more appropriate to weight the scoring; for discussion of this procedures, see Chapter 4: Outcome Indicators and Measures.

Reliability: In order to aid in the identification of false responding the checklist includes one bogus item ("Cadrines"). The data on individuals who indicate the use of this non-existent drug should be discarded. If a high percentage of respondents (that is, above 20 percent) check this category, the entire data set is suspect. Other reliability data are not available at this time, and it is recommended that an evaluator establish his/her own reliability for this scale.

Validity: The only validity data available was a comparison of scores for troubled and church-affiliated youth. The drug use patterns of the two groups were very different, indicating that the subjects were answering honestly and that the scale can discriminate between populations. Other validity data are not available, and it is recommended that the evaluator establish his/her own validity for the scale.

Criticism: The major problem with this scale is its lack of use in program evaluation. Other data may be available from the author, and he should be consulted in interpreting results.

Suggestions for Use: This simple format allows easy administration of the scale as an indicator of changes in use of various substances. It allows evaluators to focus only on items relevant to the evaluation.

Attitudes Toward Drugs (p,o,i)

Author: Roy Cook

Ages: Ten - eighteen

Date: 1975

Sources: Roy Cook
Institute for Social Analysis
11800 Sunrise Valley Drive
Reston, Virginia 22091

Variables: This drug attitude scale includes items covering feelings about how acceptable drug use is, how much the respondents would like to use drugs, what they feel the effects of drug use are or can be, why they feel people use drugs or do not, and what kinds of people are users or non-users. There are also statements expressing attitudes about relevant laws and several statements which may show a differential acceptance according to which particular drug or type of drug is under consideration.

Description: Most drug prevention programs deal with attitudes toward drugs, whether they set out specifically to do so or not. They are, by definition, in the business of shaping their clients' feelings about drug use and policy. It therefore seems desirable to determine what attitudes are held by the participants in a program and how they may have changed after a period of time.

Administration and Scoring: The scoring of each item on the present scale goes from one to four, with one for "Strongly Agree" on half the items and the reverse (one for "Strongly Disagree") for the rest. Under this system, a high score represents liberal or pro-drug attitudes and a low score shows conservative or anti-drug attitudes. The possible range of scores for the whole instrument is from 33 to 132.

Reliability: No reliability data are available; however, reliability for this type of questionnaire is easy to establish, and should be completed by the evaluator.

Validity: The only validity data available was a comparison of scores for troubled and church-affiliated youth. The drug use patterns of the two groups were very different, indicating that the subjects were answering honestly and that the scale can discriminate between populations. Other validity data are not available, and it is recommended that the evaluator establish his/her own validity for the scale.

Criticism: The scale items may present reading problems for some 10-18 year old subjects. This scale has also not been widely used in program evaluation.

Suggestions for Use: One appropriate use for this scale is to administer it in conjunction with a drug use scale, which provides a validity check for both scales. In some situations where actual drug use data are not possible to collect, it could be used a more direct correlate of use; however, in doing so one must acknowledge the problem of assuming equivalence between attitudes and use.

Drug Evaluation Scales Part I: Substance Knowledge Scales (p,o,i)

Author: John D. Swisher and John J. Horan

Ages: Twelve years and older

Date: 1973

Source: Abrams, A., Garfield, E., and Swisher, J. Accountability in drug education: A model for evaluation. Washington, D.C.: Drug Abuse Council, 1974.

Variables: The knowledge subtest of the Drug Education Evaluation Scales is designed to assess information gained from various drug education programs.

Description: Based on separate item analyses, the knowledge subtest of the Drug Education Evaluation Scales has undergone at least four revisions. The current form consists of 41 multiple choice items focusing on five types of commonly abused drugs: marijuana, hallucinogens, stimulants, depressants, and opiates.

Administration and Scoring: There are forty-one items and the score is taken as the total number be correct.

Reliability: Internal consistency reliability coefficients on the instrument have exceeded .80. An instrument of this type is highly susceptible to error as a function of changes in recent research findings and/or legislative revision. Accordingly any user of this scale would be well advised to make minor adjustments in a few items or to include parallel subscales for nicotine, caffeine, and/or alcohol. However, the addition of subscales requires new item analyses and checks on reliability and validity.

Validity: Content validity was attempted by including approximately the same number of items from each category of drug. Construct and criterion-related validity are suggested by higher scores in user than in nonuser groups (p .01) and by a slight but significant correlation (r = .26, p .05) between test scores and grade-point averages.

Criticism: The scale may be outdated and must be revised for any current use. It would also improve the scale if items for alcohol, nicotine, and/or caffeine were included. Additional reliability and validity data should be obtained.

Suggestions for Use: The Drug Education Evaluation Scales have been employed in a variety of descriptive, correlational, and experimental studies. It has been found, for example, that users are generally more knowledgeable about drugs than nonusers and that increased knowledge about drugs is directly related to liberal drug attitudes. Combined with attitude and behavior scales, the knowledge subtest has also served as an outcome measure in a number of experimental drug abuse prevention programs (for example, Swisher and Crawford 1971; Swisher and Warner 1971).

References:

Abrams, A., Garfield, E., and Swisher, J. Accountability in drug education: A model for evaluation. Washington, D. C.: Drug Abuse Council, 1974.

Swisher, J. D., and Crawford, J. A. Evaluation of a short term drug education program. The School Counselor, 1971, 18, 265-272.

Swisher, J. D., and Warner, R. W. A study of four approaches to drug abuse prevention. Research Report, The Governor's Justice Commission, Pennsylvania, 1971.

Drug Knowledge (p,o,i)

Author: Roy Cook

Ages: Twelve years and older

Date: 1975

Source: Roy Cook
Institute for Social Analysis
11800 Sunrise Valley Drive
Reston, Virginia 22091

Variables: A drug knowledge scale is basic to an evaluation of any program which has education as a goal. Such programs are designed to equip their clients with the information necessary for them to make responsible decisions about drugs, both licit and illicit, and to examine their own social position and motives relevant to drugs. With these goals in mind this scale was designed to discover how much the respondent knows about the characteristics of a variety of drugs, including their physical and psychological effects, their origins, and their uses.

Description: This scale is similar in some ways to Althoff's Drug Knowledge Scale and to the Penn State Drug Knowledge Scale. Like them it has a multiple choice format and an emphasis on the effects various drugs may have on those who take them. Terminology related to the drug field, both slang and otherwise, forms a large part of this scale. Only a few items deal specifically with street names, because it is intended for use over as broad a spectrum of U.S. geography and culture as possible. This scale, unlike Althoff's or the Penn State Scale, covers not only illicit drugs but also legal over-the-counter drugs like tobacco, caffeine, and alcohol. It therefore covers a larger number of products and conditions of use.

Administration and Scoring: A total score is obtained by adding up the number of correct answers. Test scores may range from 0 to 35. It might also be of interest to look at some groups of related items. The largest of these would cover information about the effects of drugs on the user. Another possible category of questions to be reviewed separately would be those that have to do with terminoloy, including origins, definitions, and slang.

Reliability: No data available.

Validity: No data available.

Criticism: As a research tool, the user should plan a pilot test to establish reliability and validity. Additional data may be available from the author.

Suggestions for Use: The Drug Knowledge Scale is readable and understandable by adolescents. However, further field testing is desirable to gather validity and reliability data. The Scale's general use in drug prevention program evaluation is dependent on the program objectives--the content of the drug education should match item content.

University of Pittsburgh Youth Alcohol Survey (o,i)

Author: Howard Blaine

Ages: 12 years and older

Date: 1977

Source: Howard Blaine
University of Pittsburgh
5K01 Forbes Quadrangle
Pittsburgh, PA 15260

Variables: The U. P. Youth Alcohol Survey consists of seven main scales:

1. A quantity frequency index for wine, beer, and liquor.
2. A measure of the amount of supervision or control present in the situations in which drinking usually occurs.
3. Shortened versions of the Williams' temperate and intemperate use scales (Williams, et al 1968).
4. An alcohol knowledge scale.
5. An attraction of drinking scale.
6. A drawbacks to drinking scale.
7. A consequence of drinking scale.

In addition, the survey contains one item on smoking, three items on marijuana, and six demographic items.

Description: The U. P. Youth Alcohol Survey consists of approximately 180 multiple choice, true-false, and Likert-type items.

Administration and Scoring: The survey is self-administered and takes about 50 minutes to complete. Keypunch numbers and codes appear on the survey form. Data from factor and cluster analyses are currently being used to construct derived scales. Inquiries concerning these analyses should be directed to the author.

Reliability and Validity: Data are currently being gathered and are available from the author.

Criticisms: No psychometric data are available at this writing, although the short forms of the Williams' intemperate and temperate use scales seem to correlate well with the originals.

Suggestions for Use: The U. P. Youth Alcohol Survey should be appropriate to alcohol abuse prevention programs targeted at teens. Psychometric information should be obtained either from the author, or ideally in pretests with the program population.

References: Williams, A., D. Cicco, L., and Unterberger, H. Philosophy and evaluation of an alcohol education program. Quarterly Journal of Studies on Alcohol, 1968, 29 (3) pp. 685-702.

Adolescent Drinking Behavior and Attitudes Scale (o,i)

Authors: J. Valley Rachal, Jay R. Williams, Mary L. Brehm, Betty Cavanaugh, R. Paul Moore, and William C. Ekerman.

Age: 7th - 12th Grade

Date: 1977

Source: Valley Rachal
Research Triangle Institute
Post Office Box 12194
Research Triangle Park, North Carolina 27709

Variables: The Adolescent Drinking Behavior and Attitudes Questionnaire gathers data in the following 10 areas:

1. Demographic, attitudinal, and personality variables.
2. Drinking variables including a quantity-frequency index, a brief consumption history, and the respondents estimate of how many students his/her age drink in a drinking setting.
3. Circumstances surrounding drinking, including where and with whom drinking occurs, and attitudes towards peer drinking.
4. Consequences of drinking, including a variety of outcomes, the majority which are negative (e.g., being criticized by a date, getting into school trouble).
5. Perceived environment.
6. Availability and opportunity to obtain alcohol.
7. Use of drugs other than alcohol.
8. Deviant behavior which is not (necessarily) alcohol related.
9. Antisocial behavior including impulse or antisocial behaviors which are both related and nonrelated to alcohol.

Description: The questionnaire is based upon the approach of Jessor et al. (1968). It consists of 104 multiple choice, Likert-type, and checklist items in booklet form, with directions printed on the cover.

Administration and Scoring: The questionnaire is self administered and takes about 50 minutes to complete. Scoring is direct.

Reliability: Data across two independent national samples (1974 and 1978) are similar enough to suggest good reliability (Rachal 1980).

Validity: Rachal et al. (1976) report significant differences in the expected direction on a variety of social and behavior variables (as measured by the questionnaire) among 13,122 minimal, nonproblem, and problem drinkers (as measured by the questionnaire).

Criticisms: Since each item produces one datum, the questionnaire generates over 100 separate data points. Factor or cluster analysis is thus desirable.

Suggestions and Use: The questionnaire seems well suited to outcome or impact evaluation of alcohol prevention programs.

References:

Jessor, R., Graves, T. D., Hanson, R. C. and Jessor, S. L. Society personality, and deviant behavior: A study of a tri-ethnic community. New York: Holt, Rinehart and Winston, 1968.

Rachal, J. V. Personal Communication, 1980.

Rachal, J. V., Hubbard, R., Williams, J. and Tuchfeld, B. Drinking levels and problem drinking among junior and senior high school students. Journal of Studies on Alcohol, 1976, 37, (11), 1751-1761.

A REVIEW OF INSTRUMENT AND DATA SOURCES

INSTRUMENT SOURCES

Measures of Social Psychological Attitudes, by John P. Robinson and Phillip R. Shaver, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106. This book reviews more than 100 measures of social-psychological attitudes, including categories such as self-esteem, locus of control, alienation and anomia, authoritarianism, values, attitudes toward people, religious attitudes, and methodological scales. In addition to psychometric properties of the scales, key references, along with the scale itself are presented.

Each section contains a review of literature relevant to the construct (for example, values) and then critiques several instruments for each construct. The instruments generally are aimed at adolescents and adults, but not exclusively. A review of a particular instrument covers basic information and particular attention was given to reliability and validity.

Scales for the Measurement of Attitudes, by Marvin E. Shaw and Jack M. Wright, McGraw-Hill, 1967. This book of 600 pages includes chapters on the nature of attitudes and methods of scale construction, followed by eight chapters on attitudes toward social practices, social issues and problems, international issues, ethnic and political groups, social institutions, and others. More than 500 references are found on pages 571-592. While somewhat dated, and perhaps not as relevant to the drug investigator as to others, this book was planned carefully and should prove useful to anyone studying attitudes.

Drugs and Attitude Change, National Institute on Drug Abuse. U. S. Government Printing Office, Washington, D.C., 20402, 1974. This volume, edited by Ferguson, Lennox, and Lettieri is the third in a series of volumes devoted to research issues related to drug abuse. It contains 152 pages of reviews of studies of drug related attitudes, most of which focus on attitude change. A review article by William J. McGuire (1969) entitled, "The Nature of Attitudes and Attitude Change," including 840 references, is recommended to the reader.

Handbook of Research Design and Social Measurement, by Delbert C. Miller. David McKay Co. Inc., 3rd edition, 1977. Among the five chapters of this 158 page volume is a 248 page chapter entitled, "Selected Sociometric Scales and Indexes." It includes descriptions of a wide variety of indexes, indicators, measures, and scales under headings such as social status, social indicators, social participation, community, morale, job satisfaction, and others. Each instrument is printed completely and described in terms of reliability, validity, correlations with other measures, key references, and other pertinent information. This softcover book is highly recommended.

Drug Abuse Instrument Handbook, National Institute on Drug Abuse. U. S. Government Printing Office, Washington, D.C., 20402, 1977. This book is designed to aid evaluators in identifying, acquiring, and developing valid and reliable instruments related to psychosocial drug use and abuse. Intended to serve as a basic reference guide, it categorizes more than 2,000 items from 40 instruments and suggests additional items for the creation of new instruments. Instruments were selected on the basis of their ability to discriminate between drug users and nonusers, and to identify different drug user types. The items themselves are organized into four major divisions and then into specific subdivisions within these sections, with repetition of items being kept to a minimum.

Section one includes items concerned with demographics. The second section covers interpersonal variables including group affiliations, family/parental relationships, peer relationships, family vs. peers, and interpersonal adjustments. Issues assessed in the interpersonal section involve feelings about education, religion, marriage/sex/parental role, career/life goals, personal values, socio-political orientation, and world personality (general). The fourth section contains information relevant to drug usage. The fifth section contains comprehensive descriptive summaries of each instrument and how it has been used in drug research. This includes information regarding target drugs, age range, assessment areas, design features, and administration. Abstracts describing reliability or validity, history,

availability, and reports referring to the scale are included. A guide indicating where specific items from the scale are located in the preceding four sections concludes this highly comprehensive and useful anthology.

A Guide to 3,000 References on Instrument Construction and Selected Applications. Chun, Cobb, and French (Eds.), Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106, 1975. This 664 page volume is an excellent guide to a variety of measures covering a range of psychological variables relevant to drug researchers. Each study of an instrument and selected applications of that instrument are indexed by author and subject matter.

Instrumentation in Human Relations Training, by J. William Pfeiffer and Richard Heslin. University Associates, La Jolla, California 92037, 1976. This volume provides brief descriptions of 92 instruments which demonstrate wide application to the behavioral sciences and, more specifically, to actual training settings. The authors provide an introduction to instrumentation and its uses in affective program which, depending on the level of the reader's expertise, can serve as review or basic instruction.

Annotations of selected instruments and suggestions for appropriate use are given, however, no discussion of the reliability and validity data on the scales is provided. Instruments were selected on the basis of their relevance to human relations groups and their general availability.

A Sourcebook of Mental Health Measures. Comrey, Backer, and Glaser (Eds.). Human Interaction Research Institute, 10880 Wilshire Boulevard, Los Angeles, California, 90024, 1973. This volume contains a rich variety of references under titles such as Crime, Drugs, Family Interactions, Juvenile Delinquency, Mental Health Attitudes, Occupational Adjustment, and many others. Each reference includes a brief description of the instrument including what it measures, available psychometric properties, descriptions of samples used to construct it and, in most cases, where a copy of the instrument can be obtained. Other salient references sometimes are provided.

Psychotherapy Change Measures, by Irene E. Waskow and Morris B. Parloff (Eds.). National Institute of Mental Health, U. S, Government Printing Office, Washington, D.C., 20402, 1975. This book focuses on measures of sufficient quality to detect change related to psychotherapeutic intervention. Chapters of this volume, written by experienced investigators, cover a variety of measurement areas, including patient, therapist, relevant other measures, and independent clinical evaluations. Procedures and standards for selecting a battery of measures for detecting change due to therapeutic intervetion conclude this volume. Each chapter presentation includes instrument descriptions and a set of key references on the subject being covered.

Socioemotional Measures for Preschool and Kindergarten Children, by Deborah Klein Walder. Josey-Bass, San Francisco, 1973. This volume provides a comprehensive listing and description of the 143 measures available to professionals as of 1972, for assessing the affective growth of the young child, aged three to six. Information including reliability and validity is provided for each socioemotional measure, thus making the anthology a useful reference guide as well as a state of the art review. An extensive bibliography is provided, but the scales themselves are not included. Part one of the book reviews the definitions and classifications of measures for the young child. Procedures for locating available measures are noted. Six main measurement techniques--projective techniques, unobtrusive measures, observational procedures, rating scales, self-report measures, and situational measures--are described along with critiques of the advantages and disadvantages of using each with young children. Finally, recommendations for future work are suggested.

Part two, the bulk of the text, classifies the measures on a convenient schema comprised of six categories: attitudes, general personality or adjustment, interest, personality or behavioral traits, self-concept, and social skills or competence. The eleven measures of attitude are primarily self-concept inventories and structured interviews designed to assess reactions or feelings about racial and ethnic concerns and are not available commercially. Of the 38 measures of general personality and emotional adjustment, 80 percent are projective measures, and most are commercially available.

A System of Assessing Affectivity, by Robert Bills. University of Alabama Press, 1975. This book was designed to fill the need for a well-validated, reliable, and systematic approach to assessing affectivity, which would both reflect a cohesive point of view and report results in a form usable in a school's self-improvement program. The system, developed over two decades, including five principle instruments, applicable to grades seven through twelve and some lower grades as well. The purpose of this package of instruments is to assess the affective learning of students and clients. It also contains considerable technical and statistical information and research reviews. The scales themselves are appended. It was designed to be useful in coordinating a set of recommendations for improvement of a program and for advising administrators regarding the effectiveness of innovative practices.

Accountability in Drug Education: A Model for Evaluation, A. Abrams and J. Swisher (Eds.). Drug Abuse Council, Washington D.C., 1973. This volume was one in a series of handbooks provided by the Drug Abuse Council. It offers a discussion of the research fundamentals which are crucial to identifying effective drug abuse education activities. Its purpose is to aid project facilitators in clarifying specific goals and outcomes and to gauge the effectiveness of the research guidelines and resources now available for assessing emerging methodologies and their older counterparts. The handbook can be a useful source for those involved in drug abuse prevention and drug education research and for others seeking to enhance their sensitivity to programmatic concerns and dilemmas confronting drug educators and youth alike. In an effort to provide readers with a set of fundamentally reliable instruments, which can be used as a point of departure in conducting research, an aid in locating measures, and a standard of comparison for other scales, particular examples are included or mentioned in the text.

Section One discusses program development and evaluation issues. Section Two, "Measures for Drug Education," contains two articles describing two Stanford University and four Pennsylvania State University Evaluation Scales, all measures geared to evaluate the effects of drug education programs on audiences of various age groups regarding attitudes toward drugs and use of drugs. Samples of the instruments and reliability and validity data are provided also.

Ralph G. Connor Alcohol Research Reference Files (CARRF), Center for Alcohol Studies, Rutgers Univerisity, Allison Road, Piscataway, New Jersey 08854. The CARRF is a collection of questionnaires, interview schedules, and survey forms used in research on drinking and alcoholism. Instruments are available at minimal cost in for such topics as drinking among young people, drinking history, drinking and driving, attitudes, drinking in industry, screening, and evaluation. A bibliography of relevant research reports is provided with each instrument. An inventory of available instruments may be obtained from the Center for Alcohol Studies.

Measures of Occupational Attitudes and Occupational Characteristics, by J. P. Robinson, A. Athansiou, and K. B. Head. Institute for Social Research University of Michigan, Ann Arbor, Michigan, 1968. This volume focuses on the effects of worker's attitudes in the business sector. However, the instruments might be adapted to assess the relationship between attitudes and performance in other environments. The conclusions from the analysis of empirical studies suggest that a higher degree of job satisfaction is achieved when the interests and needs of the individual are matched with the job. They suggest further that the effects of worker turnover and absenteeism are related to the degree of job satisfaction. Instruments used to measure these relationships include, among others: the Index of Job Satisfaction (Kornhouser 1965); Factors for Job Satisfaction and Job Dissatisfaction (Dunnette 1966); Job Satisfaction Scale (Johnson 1955); Job Satisfaction (Hoppock 1935); and Work Satisfaction and Personal Happiness (Noll and Bradburn 1968). Each instrument is reproduced in the text. Where known, its description, design, administration, use, and reliability and validity are summarized and critiqued.

Handbook of Organizational Measurement, by J. L. Price. D. C. Heath and Co., Lexington, Massachusetts, 1972. Price intends his handbook to promote the standardization of measures used in the study of organizations. He identifies twenty-two concepts that are used to describe or analyze organizational behavior and the behaviors of individuals within organizations. Each concept is defined in general terms and differentiated from similar concepts. Dimensions of the concept appropriate for measurement are discussed. This

general information is followed by a description of at least one study that measured the concept and the instrument that was used to measure it. Sample size, data collection, computation, validity, and reliability are summarized. Among the concepts Price reviews, which appear to be most appropriate for an analysis of program processes as is proposed in these Guidelines are: absenteeism, alienation, communication, consensus, effectiveness, motivation, and satisfaction.

DATA SOURCES

The Client Oriented Data Acquisition Process (CODAP), sponsored by NIDA, collects data from all drug abuse treatment and rehabilitation facilities receiving federal funds. In addition to NIDA funded units, this includes those funded by the Veteran's Administration and the Bureau of Prisons. About 2,200 clinics are currently included in CODAP reporting system, approximately 60 percent of all treatment units in the United States. The types of data available from CODAP include: demographic characteristics of clients, such as race, age, sex, education and employment status: drugs of abuse; patterns of abuse, such as age of first use of a drug, time interval between age of first use and entry into treatment (a calculated variable); number of prior treatment experiences; and treatment related data, such as modality and environment at time of admission and discharge, weeks in treatment, reason for discharge, and so forth.

Some of CODAP's limitations stem from reporting errors, which vary by clinic, and clerical errors. Historical analyses are hampered by the proportion of discharge files which cannot be matched to admission files, although this problem is less severe than it has been in the past. To date, CODAP has been used effectively to describe client characteristics and their changes over time, but it has not yet been shown to be useful in predicting success in treatment.

Drug Abuse Warning Network (DAWN). The data collection is sponsored by the Drug Enforcement Administration and the National Institute on Drug Abuse to monitor drug use patterns. DAWN collects data from 26 large standard Metropolitan Statistical Areas (SMSA's) throughout the U.S.

Its purpose includes the identification of drugs and substances associated with abuse, providing data for control and scheduling, assessment of health hazards associated with drug use, and provision of data for program planning. The statistics are collected from reports of drug abuse episodes in hospital emergency rooms, and medical examiners. These statistics include 18 data elements relating to the facility reporting, the person involved, and the drugs abused. The data include age, race, sex, employment, reason for taking drugs, reason for contact, disposition, dosage form, route of administration, and patients' clinical status. No names are included. The output includes a hardcopy quarterly report, monthly statistical summary, monthly tabulations and reports, and monthly computer tape. All of the reports are available publicly upon request from the Forecasting Branch, NIDA.

DAWN is the most frequently cited and used data source for drug program evaluation. The data have several limitations. DAWN covers only 26 SMSA's, thus it is not nationally representative. DAWN is heavily collected from areas with serious drug problems. Secondly, hospitals are not compelled to report their emergency room information to the DAWN system. In many cases, hospitals' emergency rooms do not have adequate record files. Thirdly, drug-related data from medical examiners suffer from a time lag problem. The length of time from the date of death, confirmation as a drug related death and the actual reporting to DAWN vary among SMSA's. These difficulties in interpreting DAWN data are discussed by William A. Barton in DAWN, An Operational Analysis and Evaluation (Rockville, MD: NIDA November 1975).

DEA-Heroin and Cocaine Retail Price/Purity Index. The index was developed to help estimate the availability of illicit narcotics. Specifically, the retail price and purity of heroin and cocaine are measured to provide a means of estimating the changes in the illicit drug supply over an extended time period. Evidence is acquired through a retrieval system. Selections are made so that only samples meeting the retail purchase can be examined in the laboratories. The Herion and Cocaine Retail Purity Index is computed quarterly. Through this index, one can estimate an average dollar price for a pure milligram of either herion or cocaine. The data are computed nationally, by geographic region, and for selected metro-

politan areas. The information is available publicly. Reports are available in hard copy, or are published alone with an analysis on the DEA Performance Measurement Report, The Quarterly White House Drug Briefing, and the DEA Statistical Report.

Geo-Drug Enforcement Program. The system is designed specifically to define the procedures for the classification, reporting, and compilation of the drug environment. The more specific elements of this sytem include the importance of location, resources expended, and lastly, the direct results of arrests and drug seizures. The input is acquired internally with the data pulled at DEA headquarters from personal history reports on arrestees. Still other ways of acquiring information are through vouchers for payments, the System to Retrieve Information from Drug Evidence (STRIDE), and from the DEA foreign regions office. The data indicate domestic drug removals, expenditures for purchase of information, and manpower utilization. The Geo-Drug Enforcement Program is the key source of output for this information. They publish their findings quarterly in hard copy. The report is intended for internal DEA management and government use, provided by Office of Enforcement., DEA.

National Drug and Alcoholism Treatment Utilization Survey (NDATUS) This is a national survey of all drug abuse and alcoholism treatment facilities in the country. It is specfically designed to collect data from treatment services units, whether publicly or privately funded. NDATUS collects the following types of information from each known drug abuse, combined drug abuse/alcoholism, and alcoholism treatment unit in the United States regardless of funding source identification information, drug abuse treatment population, alcohol abuse treatment population, funding information, treatment unit staffing, and methadone and LAAM treatment unit information. The survey is conducted by NIDA and NIAAA through the State alcoholism and drug agencies. Data from the survey are shared with the States and State data tapes are available. All data from the system are available to the general public through the Drug Abuse Epidemiology Data System.

The major problems with data collection stem from the inability of some reporting units to correctly distinguish sources of funding at the unit level. The result of this is that many treatment units are forced to almost arbitrarily split funding with other units within the overall program. Because of problems with financial reporting, one should exercise caution when interpreting this data.

Prosecutor's Management Information System (PROMIS). PROMIS is a method for tracking litigation from the arrest of a defendant to the disposition of the case. The program also aids in furnishing an objective process for identifying serious crimes and recidivists rates, as well as giving information on the effect of policies and procedures. PROMIS creates files of litigation which can be accessed by criminal justice personnel. The input is acquired through various means including indictment lists, sentencing records, daily trial schedules, daily court action reports, preliminary hearing calendars, grand jury disposition cards, misdemeanor calendars, breakdown cards, and arraignment cards. The defendant's history for other pending cases, the status of cases, defendant's address, time, date, and charge are all included within the system. The data are collected from the master file. The major sources of output are through the One Day Misdemeanor Calendar, the One Day Preliminary Hearing Calendar, the Misdemeanor Specially Assigned Cases, and the Felony Specially Assigned Cases. The general inquiry and management report packages are available to the public. However, all other data are for internal use only. PROMIS is only available in several locations at present, but its use is growing as more jurisdictions develop the capability to adopt it.

Uniform Crime Reports. This system produces criminal statistics on a national basis for use in law enforcement administration, operation, and management. However, since there are other professionals who are interested in the crime problem, the data are also intended for their use and public use as well. Data are collected from monthly and annual reports which are submitted by 135,000 local, county, and state law enforcement agencies throughout the nation. The types of criteria which the law enforcement agencies report are: Drug arrests, property stolen, value of property stolen, homicide data, law enforcement officers killed, age, sex, and race of persons arrested. The output includes Crime in the United States (annual, hardcopy), Uniform Crime Reports (quarterly, hardcopy), Law Enforcement Officers Killed (annual, hardcopy), Analysis of Assults on Federal Officers (annual, hardcopy), and Bomb Summary (annual, hardcopy). Data are provided at various levels, including local,

county, state, and national aggregations. Reports are available to the public from the Federal Bureau of Investigation.

Alcohol Epidemiologic Data System (AEDS). The Alcohol Epidemiologic Data System (AEDS) is under development to provide a national, centralized, epidemiologically oriented data repository of alcohol related problems, consumption, and epidemiologic summaries. Development of the AEDS was initiated in July 1977 by the NIAAA's Laboratory of Epidemiology and Population Studies and Special Studies Branch, and its contractor, General Electric Company, to identify, locate, describe, classify, catalog, and, where appropriate, acquire data, data bases, and data resources that are relevant to epidemiologic descriptions of and investigations into alcohol abuse and alcoholism. When developed, AEDS is designed to provide data oriented services to the alcohol research community, alcohol program planners, and other concerned agencies.

The concept of AEDS operation is that one centralized activity will maintain a file of information about alcohol epidemiology data that is comprehensive, current, readily available, and easily used. Agencies seeking data that relate to the epidemiology of alcohol abuse and alcoholism can acquire these data from one coordinated resource, rather than duplicating the efforts of identifying, locating, and acquiring such data bases.

Answers to inquiries from recognized agencies and researchers will be provided by the AEDS staff. An automated retrieval system is under development to permit users to search for and retrieve specifica data files that match a set of inquiry index descriptors. In most cases, the actual data file itself will be archieved in machine readable form at AEDS; arrangements can be made for copies of data sets upon formal request to the NIAAA.

In some cases of large online data base systems, linkage to AEDS will be arranged rather than physical acquisition and archiving.

The AEDS also collects and maintains selected hard copy versions of data-containing reports, documents, and other publications. However, the primary focus is on machine readable data so that from these tables users may develop the specific information, tables, or statistics they need.

Note: The above description was provided by the Laboratory of Epidemiology and Population Studies, NIAAA, and is reprinted with their permission.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1983 O 410-948

END