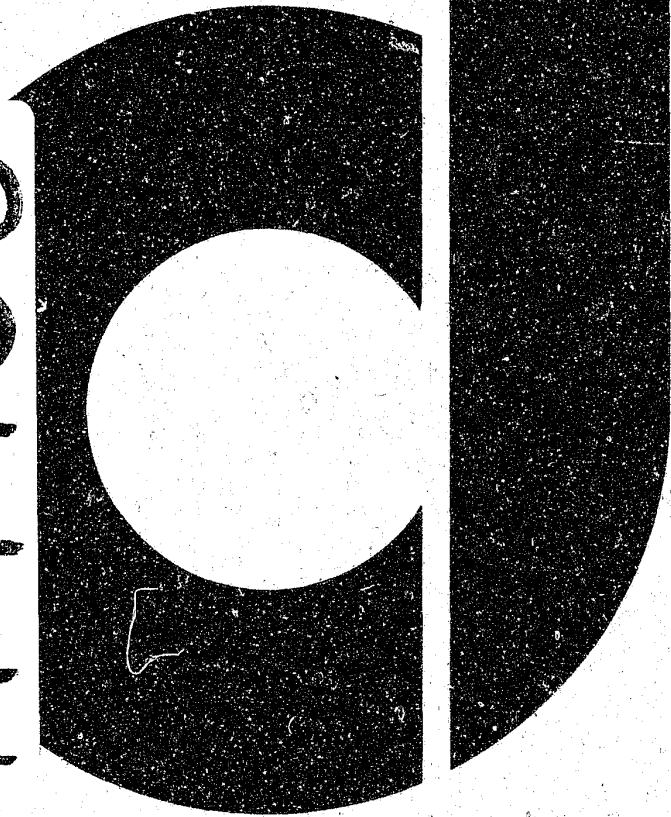




ROUTINIZING EVALUATION:
Getting Feedback on
Effectiveness of Crime
and Delinquency Programs

Mainly Corrections
BY
Daniel Glaser

147758



147758

BLUJ

CRIME AND DELINQUENCY ISSUES

A Monograph Series

**ROUTINIZING EVALUATION:
Getting Feedback on
Effectiveness of Crime
and Delinquency Programs**

by

Daniel Glaser, Ph.D.

University of Southern California

National Institute of Mental Health
Center for Studies of Crime and Delinquency
5600 Fishers Lane
Rockville, Maryland 20852

This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the National Institute of Justice.

Permission to reproduce this [redacted] material has been granted by

Public Domain/National
Institute of Mental Health

to the National Criminal Justice Reference Service (NCJRS).

Further reproduction outside of the NCJRS system requires permission of the [redacted] owner.

This monograph is one of a series on current issues and directions in the area of crime and delinquency. The series is being sponsored by the Center for Studies of Crime and Delinquency, National Institute of Mental Health, to encourage the exchange of views on issues and to promote in-depth analyses and development of insights and recommendations pertaining to them.

This monograph was written by a recognized authority in the subject matter field under contract number HSM 42-72-119 from the National Institute of Mental Health. The opinions expressed herein are the views of the author and do not necessarily reflect the Department of Health, Education, and Welfare.

DHEW Publication No. (HSM) 73-9123

Printed 1973

FOREWORD

Professionals in the field of crime and delinquency agree that selected social and behavioral science concepts and procedures should be used to a greater extent in planning and testing the effectiveness of crime and delinquency programs. They also agree that researchers with an evaluation perspective should make greater use of crime and delinquency programs for theory building and study. These agreements suggest that both the basic and applied interests can be served by collaborative efforts.

It is true that there are substantial problems in putting these agreements into operation. Problems stem from a variety of sources including: (1) traditional orientations that practitioners and program administrators in the crime and delinquency field are only concerned with fulfilling their mandate to prevent, treat, and control delinquency and crime as effectively as possible, while researchers with an evaluation perspective are only interested in developing and testing highly specialized research methodologies and contributing to a body of knowledge having esoteric theoretical import; (2) traditional perspectives that practitioners and program administrators are disinterested in, threatened by, and resistive to evaluative research and/or theory building in their area of work, while researchers do not and cannot tolerate working in a practice setting where exigencies of service overshadow typical research concerns; and (3) traditional practice of evaluative research (systematic program assessment) by short-term commitment and project.

To substantially alleviate these problems, various changes will have to be made by agencies and the research community. Such changes will require shifting philosophies of operation, restructuring organizations, adapting operations and methodologies, and generally adopting a commitment for collaboration.

Adapting current operations and methodologies, like the other required changes, apply to both agencies and the research community. A strategic place to begin is developing improved methodologies for conducting evaluative research. This is the task which

has been undertaken by Dr. Daniel Glaser. With the overriding aim of providing a document to assist agencies and the evaluation researchers working in them to routinize evaluation, Dr. Glaser has written a unique book. Drawing upon the most applicable research designs and procedures and the richness of his professional experience as both a researcher and agency administrator, he has set forth usable sets of designs and procedures that can be applied to a variety of evaluation problems. Moreover, he has illustrated the application of these designs to a variety of agency settings. Yet Dr. Glaser has been able to maintain a general perspective in his writing so the reader can use the ideas for application in settings other than those to which he specifically refers. In sum, it is a thorough, exceptionally thoughtful work that balances both abstract concepts and concrete practices. It reflects the complexity and potential of evaluative research for improving both agency practice and the body of knowledge in the crime and delinquency field.

In order to provide the author complete freedom to develop the various issues of this topic, no detailed specification or outline were set in advance and no substantive changes have been made by the National Institute of Mental Health. The views expressed, therefore, are those of the author.

SALEEM A. SHAH, PH.D.
*Chief, Center for Studies of
Crime and Delinquency*

CONTENTS

	<i>Pages</i>
Chapter 1	The Problem: Can Science Guide Us? 1
Chapter 2	The Unnecessary Stumbling Block: Defining "Success" 4
Chapter 3	Measuring Success: Poor, Good, and Better Criteria 16
Chapter 4	The Payoff: Cost-Benefit Analysis . . 26
Chapter 5	Combating Spurious Criteria: De- fenses Against Snipping 48
Chapter 6	Experiments, Pseudoexperiments and Quasi-Experiments: Who Should be Compared? 55
Chapter 7	Crossing the Time Barrier: Obtain- ing Outcome Data 84
Chapter 8	Automating Input Data: The Integra- tion of Research and Operations Records 103
Chapter 9	Combining Input Data: From Score Systems to Word Systems 137
Chapter 10	Evaluation Responsibilities: Who Should Evaluate Whom? 156
Chapter 11	Making Knowledge Cumulative: Perpetual Inventories of What We Think We Know 176
References 183
Appendix A	A Precoded Form as Initial Replace- ment for Narrative Report 189
Appendix B	Revised Precoded Forms 198
Appendix C	Forms From the U.S. Bureau of Pri- sons RAPS-2 Project 203

CHAPTER 1

THE PROBLEM: CAN SCIENCE GUIDE US?

This book is written primarily for organizations that try to change people adjudged delinquent or criminal. It may also prove useful to establishments for persons regarded as addicted, psychotic, retarded, or any other designations of deviance, provided their clients are considered modifiable, so that they may be helped to merit such labels as "reformed," "cured," "rehabilitated," "normal," "educated," "trained," or, minimally, "improved." Our concern is with organizations—for example, prisons, probation offices, treatment centers, hospitals, clinics, and training schools—which proclaim that one of their objectives is to make their clients no longer deviant, or less deviant than previously.

Before further discussion it will be acknowledged that often the most effective way to reduce the extent to which people are labeled deviant is not to change their behavior, but to change labeling practices so that they are no longer considered deviant. For example, instead of trying to change people so they will cease the moderate use of marijuana, we can cease regarding this practice as warranting their being changed, as we did in the 1930's with alcohol consumption. In other contexts the author (1971, 1972) and others (e.g., Becker, 1963; Schur, 1971) have stressed this orientation to much that is considered deviant, although the author has argued that some behavior which clearly victimizes others ("predation") has always been and probably always will be regarded as deviant in every society. In the present volume, however, the question of what behavior should be regarded as warranting change is not addressed. Given the concern of some organizations with changing the behavior of people, the problem is to determine how their efforts may be continuously guided by research. Illustrations will be supplied from a variety of such organizations without discussing here whether or not they *should* endeavor to change their clientele.

As societies become affluent and their science and technology

grow, there is an increasing concern with measuring objectively and precisely the relative effectiveness of people-changing endeavors. For economic, humanitarian, scientific, and other reasons, many wish to know which of the alternative possible forms of organization, policy, or procedure is most effective in altering a particular type of client. The effort to provide such knowledge is what is referred to here as "evaluative research."

Despite a tremendous increase in money and personnel presumably devoted to evaluative research on people-changing endeavors, this research remains sporadic. Scientific assessment of efforts to change people are undertaken primarily in temporary projects yielding results that are inconclusive and not readily comparable with the results of other evaluative efforts. In part because of this, officials can rarely demonstrate conclusively the effectiveness of their programs, as compared with the effectiveness of alternative programs or of no programs.

When a research staff is made a permanent part of a people-changing organization, with the proclaimed objective of providing evaluative knowledge to guide the organization's choice of programs for particular types of clientele, the researchers frequently drop their original concern with evaluation. Instead, they are kept busy gathering other information desired by management (e.g., precise data on the number and variety of services that the agency provides, whether effective or not), or they pursue academic or arcane research of more interest to themselves than to management. In the latter case they may be retained as "window dressing," to suggest that the organization recognizes the importance of supporting scientific inquiry, but they are then regarded as a luxury by top management and are likely to be eliminated or assigned nonresearch tasks whenever there is an economy drive.

Despite the transient nature of many evaluative endeavors, the institutionalization of such research as a continuous source of policy guidance is asserted repeatedly to be a necessary goal. Decisions are made daily on the denial or restoration of liberty to the clientele of many people-changing organizations, and billions of dollars are spent annually on efforts to change them as well as others who come voluntarily for help. These decisions and expenditures are made by organization officials or staff with little precise information on the validity of the innumerable individual diagnoses, prognoses, and prescriptions and recommendations that they issue, or on the consequences of the services they provide subsequently.

Evaluative research in corrections has been called "an elusive

paradise" (Glaser, 1965a), because, though it has been promoted and initiated by leading criminologists for over a century, it has not been established securely. Therefore, when the sponsors of this publication proposed a "how-to" manual on evaluation research in corrections, it was appropriate to respond that the most crucial need was not information on how to conduct evaluative research. Rather, the need is for advice on how to assure that it will be conducted continuously in a manner which influences treatment. This volume addresses both of these problems, that of how to do evaluative research and how to make it routinely guide policy and practice in people-changing organizations. At times, however, the focus will be on only one of these problems and at times it will be on the other, or both will be addressed together.

It is postulated here that science can guide us. For many people-changing problems it can readily be shown that science already has guided us. Some treatment methods have been scientifically demonstrated to be more effective than others for achieving some kinds of change in some kinds of people. For many problems in people-changing, however, science now provides little or no demonstrably useful advice. This book is concerned with showing how scientific guidance can be increased, and especially, how it can be provided in a manner that makes it routinely sought and adopted by those entrusted with people-changing.

Each of the 11 chapters of this book deals with a separate aspect of evaluation in people-changing:

- (1) grounds for concern with evaluation;
- (2) defining success or failure;
- (3) choosing among alternative measures;
- (4) assessing effort and attainment in monetary terms;
- (5) resisting spurious evaluations;
- (6) determining what clients to compare;
- (7) obtaining data on treatment consequences;
- (8) obtaining data on subjects and programs;
- (9) combining measurements of subjects and programs;
- (10) determining who should do the comparing;
- (11) making the knowledge gained by evaluative research cumulative.

Several recommendations are made in each of the following chapters to enhance the prospect that science will guide organizations created to change people. Most of these suggestions are independent of each other, and the effects of following more than one are believed to be additive in increasing the probability that evaluative research will be routinized.

CHAPTER 2

THE UNNECESSARY STUMBLING BLOCK: DEFINING "SUCCESS"

Two good definitions of evaluation are "the procedure by which programs are studied to ascertain their effectiveness" and "measurement of accomplishment with respect to a program's particular target" (Greenberg, in Caro, 1971: 155). These terms are relatively easy to apply in assessing a business operation, where the objective is to earn money and effectiveness is measured by profit-and-loss statements. Such definitions are difficult to apply, however, to the multiple goals of a people-changing agency, when measurement is not precise or certain. It is this stumbling block—often expressed as "How can we measure how successful we are if we can't define 'success'?"—which is most often raised when correction officials are asked why they rarely conduct evaluative research.

Multiple Goals

Dr. Russell H. Levy, the Director of Research and Long-Range Planning of the Illinois Department of Corrections, observes:

A failure to distinguish among different types of correctional program goals (humanitarian, managerial, and correctional) leads us into research projects whose results cannot be accepted. The question "how successful is this program?" may be a question about its consistency with societal values, the efficiency of specific objectives of the program, or its contribution to an altering of the frequency or intensity of law violating behavior. (Personal communication.)

Every organization, like every individual human being, pursues a variety of objectives, but the aims of people-changing establishments are especially diverse and often subtle. Dr. Levy's three categories—humanitarian, managerial, and correctional—provide a useful beginning for classification of goals, but each of these three labels designates a cluster of different interests. Often the

pursuit of one aim impedes attainment of another, and the result is then a conflict among goals. For example, humanitarian interests include minimizing client suffering, but correctional concerns may require restriction of freedom, or may be pursued through imposition of emotional pain by breaking down defenses in an encounter group. Managerial goals, such as keeping costs within a prescribed budget, impede expenditures for many services to clients that could be justified from the standpoint of other goals.

A partial solution to the multiple goals problem in evaluation is to assess achievement of each of the different goals separately. Thus, a prison's operation can be graded by its low escape rate, small number of mass disturbances, few inmates receiving disciplinary reports, high proportion of prisoners completing education or vocational training programs, and high productivity of institution farms and industries, as well as low recidivism rates of its inmates. Compared to the last item, the others are relatively easy to tabulate, and are frequently presented in annual reports or other assessment documents.

While evaluation of such items increases the completeness with which officials know the consequences of their activities, there are three limitations to such efforts. The first is that identification of goals in an agency, especially after it becomes long-established, is itself a research problem. The second is that because multiple goals are rarely of equal importance to an agency, they must be assessed in terms of their relative significance. The third problem, closely related to the second, is that the interaction of objectives must be investigated to determine if attainment of one goal impedes achievement of another.

Manifest and Latent Goals

One characteristic of organizations and of individuals is what Merton (1957: 199) has called "displacement of goals." An agency or program originally created for one purpose frequently acquires additional functions, often unofficially, and its operation may then be guided more by the acquired objectives than by the purposes which first motivated its establishment.

Official goals may appropriately be called *manifest*, since they are proclaimed in the legislation, directives, or formal announcements under which programs are created or policy is publicly justified. Actual goals must be inferred from the behavior of functionaries within an organization, in terms of the objectives they seem to have. Those interests and objectives which seem to account for policy and practice, but are different from the publicly pro-

claimed objectives of an agency or program, may appropriately be called its *latent* goals.

Sometimes officials are consciously aware of their latent concerns and admit them, at least off the record, even when they do not advertise them. At other times the directors or staff of agencies drift into the pursuit of unofficial objectives at the expense of their proclaimed primary purposes, and are either unaware of this shift or loathe to admit it.

The supplementation or even the replacement of manifest goals by latent goals is readily observable in the activities of many correctional and other people-changing agencies. Such multiplication of objectives will be illustrated here by the example of parole boards. The official purposes of a parole board usually are stated in terms such as those of the National Conference on Parole (1957:66): “. . . the protection of society on the one hand and the rehabilitation of the offender on the other,” and “. . . helping the youth or the adult offender solve his personal problems in an orderly and acceptable manner.” Observation of parole board activity, however, frequently reveals the following four latent goals, which, while independent of the official objectives, are nevertheless highly influential in many parole decisions:

(1) *Reduction of disparity in sentences.* Often parole boards are confronted with persons convicted of the same crime, with similar criminal records, who have markedly different periods of confinement before they are eligible for parole. This can happen when a judge has great discretion in determining the penalty for a particular offense, or when prosecutors drop charges if the accused will plead guilty to a charge carrying a lesser penalty. The contrast between rural and urban areas is usually great in this respect, with penalties dispensed in rural areas more severe. But there usually is much disparity among cases from any metropolitan area. In many instances it is customary for board members to “retry the case,” to determine what is an equitable penalty. They may assert that one person has done enough time for this crime or that others must wait in prison much longer before they will have had the usual penalty for their crime and criminal record. Evidently, their objective is to assure that a fair retribution is exacted by the State—an aim that may or may not result in the same parole decisions that would follow if they considered only the protection of the public and rehabilitation of the offender.

(2) *Maintaining order in correctional institutions.* Parole board members, being nominated or appointed by the elected State or Federal executive, are likely to have much loyalty to this executive

and to the correctional administration as a whole. This is often the case even when the board is legally independent and its terms extend beyond those of the persons to whom the members owe their appointments. Apart from these sources of loyalty, of course, they identify much more with the general citizenry than with the prison population. One of the many ways in which such loyalties frequently are manifested is in the members' sense of obligation to "back up" the disciplinary authority of institution officials by denying parole to those who are disobedient while confined and by giving highly favorable consideration to conforming inmates. This is often done independently of the relevance of such behavior to the inmate's prospects of returning to crime when released. In the second half of the 20th century, after many prisoners rioted to protest arbitrariness, rigidity, or inconsistency in parole granting, some parole boards became more concerned that their reputation among inmates as generous and predictable may help in maintaining order at prisons.

(3) *Balancing the State budget.* Since it costs about 10 times as much to confine an offender as to supervise him on parole, any sudden shrinkage in the granting of parole creates a strain on institution budgets, even though many of their costs are inflexible. There have been times in the history of a number of States when correctional institutions were overcrowded or the State government was operating in a continuous financial crisis, or both. In such circumstances juvenile and adult parole-granting agencies have frequently responded to this financial pressure by facilitating the closing of an institution through reduction of inmate populations, or as one board member once expressed it to me, by "opening the back door wider so that more can be let in the front door." This practice is widespread among juvenile and youth parole authorities because they often consist of boards or commissions responsible for both parole and institution administration.

(4) *Maximizing public support.* When comparison is made between the attention given by parole boards to decisions on highly publicized cases and the attention they give to more ordinary cases, it becomes evident that they are concerned with public reaction to the parole of notorious offenders. Concern with public opinion is much less pronounced in their routine decisions. The parole board members, for the sake of their own careers as well as the careers of those to whom they owe their appointments, cannot confine their attention to the protection of the public and the rehabilitation of the offender when they know that the public is stirred up by a prospective parole decision. This was especially

evident to the author during his several years of employment by the Illinois Parole and Pardon Board at Joliet prison when Nathan Leopold and other highly publicized murderers received their first parole hearings.

All latent goals of parole boards or other correctional agencies may be justifiable. My concern at this point is not with evaluating the relative merit of different goals. Rather, it is with stressing the need to be aware of all of them, so that one may guide agency action effectively with respect to any one of them. *It is in the public interest that latent goals be made manifest, by determining what they are and stating them explicitly.* Only if a goal is recognized can the effectiveness of efforts to achieve it be evaluated, and the consequences of pursuing one goal for attainment of others be measured.

Returning to the example of the parole board, in recent years there has been increasing criticism of the ostensible arbitrariness and inconsistency of parole decisions. There have been many proposals to reduce the power of these boards to determine the duration of confinement. All this reduction can accomplish, however, is the transfer of some of the parole board's influence on penalties back to the courts, with no guarantee that court penalties would be any less arbitrary or more consistent than parole board decisions. There is ample ground for expecting less defensible sentencing decisions from courts than from parole boards due to the diversity of sentencing policies and competence among judges (especially elected ones), and because of the plea bargaining process by which the crime charged is altered in order to change consequent statutory sentences (see, for example, data on court decision determinants in Glaser, 1972: 92-93). The best method of achieving checks and balances in the sentences imposed on offenders would appear to be continued division of penalty-fixing authority among legislatures, courts, and parole boards, with a limited range of discretion for each, but with more possibilities for review of decisions than now exist.

If people-changing agencies are to be made more responsive to the public interest, they must make the purposes of their case decisions explicit, and the consequences of their decisions must be evaluated to determine the extent to which they accomplish their purposes. In our example of the parole board, attainment of each of their four latent goals can be measured objectively with reasonable precision. A board's reduction of sentence disparity can be assessed by comparing the *diversity* of time served before parole by particular types of offenders (e.g., first conviction armed rob-

bers or recidivist auto thieves) with the *diversity* of time they would be confined under their minimum sentences before they are legally eligible for parole, or with the time they would serve under their maximum sentence. The merit of diversity can only be debated if the diversity is known.

All other latent goals are also measurable. Parole's contribution to maintenance of order in correctional institutions could thus be inferred by noting fluctuations in disorders with changes in parole-granting practices, and perhaps by investigating inmate views on what they must do—if anything—to acquire a parole. The economic functions of parole can be measured by standard bookkeeping procedures. The maintenance of public support by the board, for itself and for the administration by which it was appointed, can be measured by opinion polling.

The consequences of stress on one parole board objective for the achievement of others can also be investigated. For example, research on Federal youthful offenders some years ago indicated that a good behavior record in prison was associated with a lower parole violation rate only for those who had been previously confined. On first confinement, inmates with both favorable and unfavorable postrelease prospects were equally likely to have difficulty avoiding a rule infraction record; only after a prior institutionalization were those who still "had not learned to do time" worse risks on parole than those whose disciplinary record suggested adjustment to prison regulation (Glaser, 1964: 296).

Therefore, excessive stress on the prison disciplinary record in parole decisions to promote obedience to prison rules might well impede attaining the objective of reducing recidivism, especially if the stress is applied to youthful inmates during their first institutional confinement.

The foregoing discussion implies that if evaluation of a correctional agency's activities is to affect its practices, the evaluation must measure achievement of both manifest and latent goals. This can be done only if latent goals are first identified and made explicit, thereby becoming manifest goals. There is no easy and certain procedure for identifying latent goals, but the following are a few broad guidelines: *First*, assess the relevance of actual practices at any agency to its official precepts. *Second*, whenever this first step suggests that some practices are inappropriate, try to determine why they are used. Those who make decisions will give reasons for them which may reveal that they have latent goals supplementing their officially stated goals. Frequently the reasons given will not explain sufficiently the decisions which are

made. When this happens, investigators will have to formulate other explanations and test them through systematic observation and analysis.

Personal versus Organizational Goals

Perhaps the most widespread latent goals are those private work or career preferences of employees which are incompatible with the manifest purposes of the organizations that employ them. While such goals are not readily admitted by those who pursue them, they must be revealed if evaluation research is to have a continuous impact on practices in people-changing agencies. Changes, either in personnel or in work incentives and procedures, to make the pursuits of employees (and volunteers) more compatible with agency objectives can be made only after latent goals are exposed.

There are both obvious and subtle sources of discrepancy between organizational and personal concerns of employees in people-changing enterprises. The obvious sources include an interest in making work easy, pleasant, and secure. This may be evident in staff making only minimal effort, or more subtly, in their emphasizing services such as providing fun and games, which keep clients contented at the expense of activities that require more staff effort but are more relevant to goals of people-changing. Personal interests of staff which are independent of the official purposes in organizations may be evident also in decisions that suggest political aspirations and obligations or prejudice and bias.

A much less obvious source of discrepancy between precept and practice is the generally unintended differential reinforcement that the staff's alternative activities receive from agency directors. An example, elaborated in Chapter 8, is found in jobs such as that of probation officer, and many other types of caseworker, in which the employee is expected to submit diagnostic and prognostic reports and recommendations on his clients, and to provide them such assistance as counseling, therapy, or technical advice, or to monitor their activities. In this type of job there often is more immediate and pronounced reward or penalty for the ability to write reports than to perform the other duties of the job.

A written report is a tangible product of which the employee can be proud. It is the product most visible to supervisors—they usually cannot see the relatively private services of counseling, assistance to, or surveillance of a client. The supervisors—whether chief probation officers, judges, wardens, parole board members,

agency directors, or other officials—make case decisions mainly on the basis of the written reports they receive from their subordinates. They can assess the clarity and plausibility of a report much more readily than its validity. Indeed, staff who are most proficient at report writing frequently have a background that makes them feel alien and uncomfortable in dealing with their clients.

Differential reinforcement for diverse requirements of a job often may be revealed by time studies which find a distribution of staff time different from that implied in job descriptions. Contrary to a supervisor's expectations, there may be more time devoted to paper work than to field work or to investigation than to supervision. Before-and-after time studies may demonstrate that when caseloads are reduced, time devoted to paperwork expands at a greater rate than time spent in other activities (cf. Glaser, 1969: 299-303).

Still another source of discrepancy between personal and organizational goals is an employee's primary concern with a career outside the organization. This has been a serious impediment to the routinization of evaluative research in people-changing agencies, especially during the 1960's, when college faculties were rapidly expanding. Repeatedly members of the research staff were much more concerned with completing theses, or with preparing academic publications which would facilitate their obtaining faculty appointments than with producing research that could guide agency practice. Occasionally research was undertaken that could serve all of these goals simultaneously, and the personal objectives of the researchers supplemented the motivation provided by organizational objectives. Time study analyses of assigned staff activity, in conjunction with assessment of the relevance of alternative activities, should reveal when employees' interest in other careers interfere with their contribution to their employers. Research employees should be assessed as much or more by the kinds of research they have *completed* as by their academic degrees or their research proposals.

The solution to problems of discrepancy between personal goals and organization goals will vary greatly with the agency, and with the particular case in an agency. The optimum solution always is to make staff activity both gratifying to employees and valuable to the organization. Sometimes this can be done by dividing responsibilities that were formerly the concern of one person and providing different types of persons for each. For example, separating the investigatory from supervisory tasks in the job of probation officer. Sometimes this is done by replacing specialists

with a team, as in casework teams, which may include one or more paraprofessionals who share a common caseload with a professional. The professional's skill at writing reports and at negotiating with supervisors and other agencies is combined with the paraprofessional's capacity for better rapport with clientele. In such teams each member can make his contribution to the others, especially if no rigid caste differentiation exists among them. The risk of intrastaff barriers may be reduced by making it possible to move from one team role to another, as from paraprofessional to professional.

Sometimes discrepancies between personal and organizational interests are reduced by sharply curtailing or eliminating tasks of little value to the organization despite staff interest in them, such as the preparation of long narrative reports, and by increasing the rewards specifically linked to tasks of greater value to the organization.

Hierarchies of Goals and Types of Success

While organizations have multiple goals, manifest and latent, not all of them are equally important. Indeed, when some latent goals become manifest, such as repaying a political obligation, they are promptly disavowed. Others, such as keeping inmates contented, are often justified by officials only as means to other ends, such as increasing custodial security or fostering rehabilitation.

Both individuals and total organizations rank their goals, considering some more important than others. There is variation in the time perspective towards different goals: some goals have highest immediate priority for an individual, such as paying bills or acquiring a driver's license, but others are considered more important in the long run, such as having a happy marriage or a satisfying profession. The term "salience hierarchy" has been used to designate a person's ranking of goals by their time priority, and "prominence hierarchy" to designate ranking by overall importance (McCall and Simmons, 1966).

Presumably the most prominent objective of any people-changing agency is to change its clientele so as to eliminate the reason for their coming to the agency. Whether such changes are achieved is assumed to be the most important question for evaluative research. Are correctional establishments influencing delinquents and criminals to stop committing offenses? Are addiction-treatment organizations helping addicts become abstinent? Are

vocational rehabilitation centers converting unemployables to employees?

Unfortunately, these most prominent questions seldom are investigated by agency researchers, who concentrate instead on what is most salient to the administrators of the agency. Usually this means giving first priority to routine tabulation of the number of major actions accomplished in various categories, for example, the number of admissions, releases, transfers, and discharges. Frequently even this first priority is replaced by a crash push for answers to questions raised by public criticism, legislators, trustees, or others who control the agency's income. It is the kind of work that the director of research of one of the country's largest local correctional agencies calls "helping the boss put out fires."

In the absence of such outside pressure, second priority is given in many agency research offices to developing new types of management information. These innovations often consist of elaborate forms for detailed recording by staff of their own activities. The directions to make entries on these forms at regular intervals—whether they be hourly, daily or weekly—are widely violated; and the forms are completed late. Entries are made hastily and often describe events and activities so that they appear to have been done "properly" rather than describing them accurately. More significantly, there is little investigation of whether recorded information is related to achievement of the agency's primary goal of changing the client, such as terminating criminal activity or drug abuse.

One of the concerns of this manual, especially in Chapter 8, is with limiting recording and reporting to that which is clearly useful, both for operations and for evaluative research. Determining what is most useful for both these purposes is a research task, also discussed in Chapter 8, and in Chapter 9.

"Putting out fires" probably will never cease to have high priority in people-changing agencies, but making evaluation research an asset in fire-quenching is a concern of this manual. This means making evaluation accessible and relevant so that agency researchers have a salience hierarchy appropriate to their agency's prominence hierarchy. It means that their highest priority should be evaluation of their agency's achievement of its most important goals. The first step toward this objective is to recognize that an agency has multiple goals, manifest and latent, and to identify them.

An important second step in evaluative research is to discern that accomplishments of a goal may be matters of degree rather

than all-or-none phenomena. For example, getting clients to stop committing crimes or to terminate drug use may be partially rather than completely achieved. Thus, strategic assistance may result in offenders committing crimes less persistently, even when it does not markedly alter their gross arrest or parole violation rate (e.g., the PICO Project, as reported in Adams, 1961). It appears from many recent reports that a larger proportion of addicts can be shifted from heroin to methadone than can be made completely abstinent from opiates. It seems that if methadone maintenance is readily available and supplemented by employment or subsidized training, there may be a marked reduction of crimes committed by addicts who are not reached by abstinence programs. Such examples illustrate the need to consider alternative kinds of success, ranking some as more important or desirable than others but not neglecting any that have appreciable importance. As the next two chapters show, it is sometimes possible to develop scales of overall accomplishment which take different degrees of change into account, but are more greatly affected by major changes than by minor ones.

The point to be stressed here is that there are different kinds of success for any agency. Sometimes all are related to the agency's primary goal, but they represent different degrees of accomplishment of that goal. Sometimes different kinds of success are related to different goals (such as reducing drug dependence, reducing crime, and increasing employability). Ranking them by their relative importance may be difficult, but it is useful to assess them all because accomplishing one or two kinds of success may be preferable to accomplishing none. Finally, it is important to recognize different kinds of success because the kind that can be measured validly or usefully varies greatly from one people-changing circumstance to another.

Conclusion

In summary, defining success can cease to be a stumbling block if one recognizes that:

- a. There can be multiple kinds of success, each with respect to a different goal.
- b. One must try to determine what behavior of officials and staff in people-changing agencies is rewarded, because that which is gratifying tends to become the actual goals in their work. Those goals that are not proclaimed are the latent goals, and they may influence behavior of personnel more than the manifest goals of

the agency. Only when all actual goals are made manifest and justified can research to determine the extent to which goals are achieved be relevant to officials, and hence have strong prospects of becoming routinized.

c. It follows that one should seek definitions of success that are useful rather than sacred, and this also applies to definitions of recidivism, abstinence, employment, and other concepts in terms of which success may be defined for specific goals.

d. The usefulness of a definition of success is not just a function of its implications for goal attainment, but also of the feasibility of measuring it, both of which are problems discussed in our next chapter.

CHAPTER 3

MEASURING SUCCESS: POOR, GOOD, AND BETTER CRITERIA

To determine usefully the relative success of different people-changing programs for a variety of kinds of success, it is of course desirable that each kind of success be clearly defined. It is also appropriate that the various types of success be ranked as to their long-run importance or prominence, as well as to their immediate utility or salience. Finally, no definition of success can be useful unless methods of measuring its attainment are sufficiently precise, valid, and reliable to warrant confidence that they improve the quality of knowledge available for guiding policy makers.

Following conventional usage in psychometrics and actuarial research, any measure of program outcome will be referred to here as the *criterion* by which the program is evaluated. Our problem is how to select the most useful *criteria* for different evaluation circumstances.

A variety of concerns in criterion selection will be discussed: that criteria be "hard" or objective, rather than "soft" or subjective; that they be relevant to attainable goals; that they involve continuous rather than discrete statistical variables; that they be useful in the determination of appropriate financing for a program. None of these concerns should be regarded as indispensable. Sometimes one must be neglected and another emphasized, although all are desirable. But all these concerns usually can be pursued simultaneously to make evaluative research more clearly useful, and hence more likely to be routinized.

The Most Objective Criterion

By their very nature, subjective evaluations tend to be biased, no matter whether people try to be unbiased. In the absence of evaluative research, the effectiveness of people-changing endeavors usually is assessed by the subjective impressions of those who administer, support, or oppose particular policies, practices, or

programs. Frequently they have devoted large amounts of time, training, or money to get one practice adopted rather than another. Indeed, their jobs, careers, and reputations may depend upon favorable evaluations for one approach and unfavorable judgments of others. That is why, to achieve the greatest degree of objectivity, independent evaluative research personnel are needed. (See Chapter 10 for who should conduct research.)

Subjective assessments tend to be biased not only by the interests of the evaluator, but by being based upon observations of an unrepresentative sample of the events or cases to which the evaluation is applied. We tend to be especially impressed by dramatic cases that come to our attention, whether favorable or unfavorable, and we therefore generalize about programs from such cases when they are not typical of most of the experience in the program.

Most important, perhaps, is that subjective impressions are based on private feelings rather than externally observable events. Our strictly personal impressions are the most easily and quickly obtained evidence of treatment outcome; they develop spontaneously, and therefore they frequently become a component of program evaluations, even when the evaluations are made by independent researchers striving to be objective.

Problems of obtaining a representative sample will not be discussed extensively, since there is an ample literature on this aspect of research and some sampling problems distinctive of evaluative research are discussed in Chapter 6. Even when assessments of people-changing efforts avoid sampling bias, however, they frequently retain largely subjective criteria of outcome. This is most glaring when the criterion of effectiveness consists of statistics on the personal judgments about a program made by a representative sample of observers, whether outside observers, agency staff, or the program's clients. Perhaps equally questionable are subjective data that consist not of direct judgments on program effectiveness, but of client responses to questions about their personal sentiments and opinions. Agencies that have as their goal the reduction of delinquent or criminal *behavior* frequently assess their effectiveness not by the subsequent acts of their clients, but by paper-and-pencil tests on the clients' self-conceptions, feelings, and beliefs.

If the aim of a program is to reduce some type of deviant behavior in the community, the ideal criteria of the program's effectiveness are objective data on the post-treatment behavior of the program's clientele. If, instead, one measures deviant behavior reduction by questionnaires on the clients' subjective opinions and

feelings, one leaves very problematic the correlation between their questionnaire responses and their actual behavior. As a matter of fact, there is evidence that questionnaire responses frequently are completely irrelevant, or even grossly misleading as predictors of behavior (Seckel, 1965 and Jesness, 1971, for comparisons of responses on one of the most carefully developed questionnaires—the Jesness Inventory—and actual postrelease behavior). Sometimes this discrepancy between word and deed occurs because subjects wish to curry favor to gain freedom or other benefit, and sometimes this discrepancy has more elusive causes.

Fiedler and Bass (1959) compared self-esteem scores for confined offenders, offenders under correctional supervision in the community, and nonoffenders. They made such a comparison twice, once for juveniles in the community and in institutions for delinquents, and once for military personnel on duty and confined in disciplinary barracks. Both of these studies revealed that non-offenders had more favorable self-conceptions than offenders, regardless of whether the offenders were confined, but that offenders confined in correctional institutions had distinctly more favorable self-conceptions than those in correctional programs of the free community.

Two explanations for the latter contrast were offered by Fiedler and Bass. The first, from psychoanalytic theory, is that those who are incarcerated are relieved of guilt feelings by being punished, and thus view themselves more favorably than do offenders granted probation. The second explanation, from reference group theory, is that institution inmates compare themselves with the other prisoners, against whom they do not seem so bad, while offenders under correctional supervision in the free community compare themselves and are compared by others with the non-offenders there, and thus more have an unfavorable self-conception in the community than they would have in a penal institution.

Regardless of which of these interpretations one accepts, the research findings show that self-concept tests would tend to make any community correctional program seem less successful than a program of incarceration. The evaluation of counseling programs may be similarly misleading when it is accomplished by tests of insight into psychological principles or by personality inventories. This is especially true when applied to counseling programs in an institution. The invalidity of such assessments develops because clients learn how to take the tests. Counseling instructs them in a "vocabulary of adjustment" needed to score favorably by verbal evidence, and in an institution they learn that "showing insight"

may hasten parole. That is why many of these programs, especially in institutions, have little or no impact on postrelease recidivism rates, although they foster favorable subjective evaluations by clients or by staff (Harrison and Mueller, 1964; Seckel, 1965; Kassebaum, Ward and Wilner, 1971).

Though the foregoing implies that one should avoid "soft" or subjective indicators of the outcome of people-changing efforts, and seek the "hardest" data available on changes in behavior, some compromise frequently is unavoidable. When prompt assessment is demanded, soft data are more readily obtained. In these circumstances it cannot be overemphasized that conclusions must be tentative and qualified until knowledge is available about the validity of the use of subjective responses as indices or predictors of objective performances. Preliminary conclusions on even a few dimensions of measurable behavior in the community, derived only from short-term followups on small samples of early releasees, may often warrant greater confidence than subjective data from larger samples of clients for whom no data on behavior in the community are available.

The Most Attainable Criterion

One useful approach to deciding upon a criterion is to consider first the kind of change that can be most readily accomplished by an agency for a particular type of client. This change may be thought of as a subgoal, or means, to the major goal. The subjects for whom such a subgoal can be achieved may then be evaluated separately, with respect to the major goal, comparing those who had help with those who had not in achieving this subgoal. For example, good health and employment are essential to achieving self-sufficiency in most other pursuits. Therefore, an agency with the primary objective of terminating crime or addiction, probably will achieve this objective most readily in those of its clientele who have health or employment problems if it first helps to relieve these problems. A few examples will illustrate this.

A large proportion of skid-row alcoholics are undernourished and debilitated. Therefore, any agency that makes more readily available to them such things as food, shelter, and medical services, including vitamins and other prescribed food supplements, in addition to such social services as Alcoholics Anonymous and other counseling, should be more successful in changing them than agencies that cannot supplement these counseling approaches with immediate physical assistance. That this is the case was suggested

rather dramatically by evaluation research in a St. Louis alcoholic detoxification center (Law Enforcement Assistance Administration, 1970).

Similarly, any group of felons and accused misdemeanants who have long sought employment unsuccessfully and are not chronically alcoholic or drug-addicted, should commit fewer crimes if they are released to subsidized job training and job placement than if they are kept in pretrial incarceration. That this is the case was dramatically demonstrated by experiments with these services for such clients in Washington, D.C., and in New York City (Leiberg, 1971: 36-46; Vera Institute of Justice, 1970).

To avoid frustration in finding attainable criteria where hardly anyone is successfully changed, it is often appropriate to subdivide a high failure rate effort into sequential problems, each of which has close to a 50-percent favorable outcome. For example, in New York City, where addicts were granted welfare payments provided they entered a treatment program to which they were assigned by a Central Referral Service, it was found that less than 20 percent were in the program a few weeks after their referral. The task of predicting which services would increase the proportion staying in these programs, however, could be usefully subdivided into two problems, each of which had closer to 50-percent success rates but different criteria of success. The first problem was getting the addict to go from the Central Referral Service to the program to which he was referred, and the second was retaining him in the program once he got there.

Part of the problem in the referral service appeared to be that the ex-addict staff, who were from total abstinence programs such as the Phoenix Houses, were particularly hostile to other types of programs, especially those providing methadone maintenance. Therefore, they violated guidelines as to the agencies to which they were to refer particular types of addict. Other variables apparently were significant also, such as client expectations of discomfort and humiliation at any unfamiliar agency. This apprehension suggested the need for an escort and orientation service at the referral center. The problem of retention involved not only the initial reception of new entrants, but other matters which varied with the agency to which referral was made. Each of the separate problems of referral and retention could be lightened more markedly by a single change in policy or procedure than the total problem of getting addicts into some treatment program.

A somewhat different strategy is appropriate when there already is a high success rate, as in many adult probation services.

For such clientele one can usefully search out homogeneous subgroups with almost a 50 percent failure rate and sharing a common problem that a special program might alleviate, such as lack of employment skills or of appreciable employment experience. Dual evaluation criteria, success at procuring and holding a job, and success at not acquiring a further criminal record can then usefully be applied to comparing the records of such clients with and without a stipend-paying vocational training and placement program, or with and without placement in a residential community correctional center pending the development of economic self-sufficiency from legitimate earnings. These programs would presumably change the careers of a larger proportion of high unemployment rate probationers—and especially, high unemployment rate parolees—than of a cross-section of the predominantly successful probation or parole population.

Our main point has been simply that one strategy in routinizing evaluative research is to conduct much of it when an immediate payoff relevant to important goals can be expected. This means being alert to the existence of remediable barriers to an appreciable impact from people-changing efforts, and to promote experimentation and evaluation on measures that remove these barriers. Whenever economic sustenance or health are acutely deficient for an identifiable fraction of an agency's clientele, it is usually a safe bet that efforts to achieve other changes in these clients will be much more effective if economic and health services are combined with other change endeavors. Whenever clients have long waiting periods or difficult transportation problems in obtaining essential services, one can usually be confident that alleviation of these problems will enhance agency effectiveness, and should be evaluated by experimental innovations.

There is no simple formula for identifying the most attainable criteria that would apply at all times to all clients of all agencies, but imaginative probing of both clients and staff should facilitate such identification. A general principle illustrated in several of the foregoing examples is avoidance of adopting as an immediate objective the alteration of outcome rates that already are very high or very low. These are what statisticians call high or low "base rates," such as the high rates of relapse in alcoholism and drug addiction or the low rates of recidivism in property offenses by adult first offenders with good employment records. With such cases it may be prudent for researchers to define as their problem a more feasible goal than that of altering drastically the already very predominant outcome. This may mean simply centering

attention on a category of the population which has about a 50 percent success rate, or altering the criterion to one of partial success which is attained by about half the cases and might readily be attained by more. (On base rates and outcome prediction, see Meehl and Rosen, 1955.)

It should be pointed out immediately that there often are exceptions to the statistical rule cited above of evaluating programs with respect to measures of success that divide the population studied approximately in half. Therefore, no purely formal data on base rates alone will substitute for a thorough understanding of the clients, their socio-economic and cultural setting, and the programs they are in.

The Most Continuous Criterion

Success is too often measured as though it were an all-or-nothing matter. It is easy to assert "either you succeed or you fail," but anyone who works at people-changing knows that success is usually a matter of degree. Recidivism, for example, is measured in terms of one rearrest, reconviction, or reimprisonment, although those thus classified as recidivists differ tremendously in the immediacy, extent, and seriousness of their renewed criminal behavior. Similarly, drug addicts and alcoholics who relapse are dichotomized from those who remain abstinent, but there is great variation in the extent to which posttreatment life involves use of alcohol or drugs. Indeed, those who are non-recidivist or are abstinent are also far from uniform; they differ in the extent to which they have actually ceased their prior deviant behavior or merely avoided detection, as well as in the extent to which they have achieved other people-changing goals, such as becoming economically self-sufficient and meeting their obligations to dependents.

Any measure of the success of a people-changing effort which fails to take into account variations in the degree to which a goal has been attained, and instead, classifies all the research subjects as either successes or failures, is thereby limited in its sensitivity as an index of variations in the effectiveness of alternative programs and policies. This limitation is somewhat comparable to that which an accountant would face in advising management of its effectiveness if only permitted to report that different components of a business made more or less than 10 percent profit, rather than the exact percentages of profit. Because money is a continuous variable showing the extent of costs and incomes, accountants can make more precise analyses of the

sources of profit and loss than purely dichotomous measurement would permit. In people-changing efforts, greater sensitivity to variability in the sources of success or failure may improve the guidance of policy. This variability may be acknowledged by using continuous rather than dichotomous measures of success. This can often be done by measuring outcome in terms of other continuous variables, such as time or money.

Several alternative continuous measures of outcome that are functions of time usually are discernible if one studies the goals of any people-changing effort. For example, success in promoting abstinence in addicts or alcoholics can be measured as number of days abstinent during a posttreatment period, thus differentiating the more immediate and persistent relapsers from those who only occasionally relapse. Similarly, those who are always abstinent and those who relapse more or less frequently can be rated on the extent to which they can achieve other goals in the agency's hierarchy of objectives, such as number of days on the job in a posttreatment period, or dollar earnings (although these are functions of employment opportunities as well as of change in behavioral preferences). Groups who have similar background but are given different training can be compared appropriately by these measures if they are released in the same area and period and thus encounter the same job market.

Probably the most sensitive criterion of the effectiveness of correctional endeavors with any group of offenders is the percentage of time they are confined during a followup period. This percentage usually is calculated from the number of days, weeks, or months they are incarcerated for new infractions in a given number of years after the correctional experience to be evaluated. The average percentage of postrelease time spent in confinement by a group of released offenders reflects the frequency, duration, and interval between their subsequent incarcerations. These aspects of success or failure are not taken into account in purely dichotomous classifications, such as "Rearrested" or "Not Rearrested," and "Reconvicted" or "Not Reconvicted." By reflecting severity of as well as number of penalties, the total amount of subsequent confinement time provides a crude index of differences in the extent of societal outrage at the conduct of various groups of released offenders.

Although it is preferable to use subsequent confinement time rather than dichotomous criteria to measure a correctional program's effectiveness, the time served for new convictions is still an imperfect index because of the inconsistencies in societal re-

action to criminal law violations. Some people who continue in crime are caught while others are not. Some "beat the rap" though they are just as guilty as others who are convicted. Some who are convicted get little or no confinement while others of similar criminal record and current offense get long confinement terms. Such diversity in the penalties for crime warrants caution in conclusions on the criminality of small numbers of cases. Yet when one compares cases of similar background that have received different treatment, the inconsistencies of the criminal justice system can be presumed to be randomly distributed. When sizable groups of similar offenders from different correctional programs are compared in a followup period, differences in their *average* period of subsequent confinement presumably reflect differences in the effectiveness of the programs they were in.

To reduce the crudeness of this measure, time confined pretrial or instead of paying fines may be excluded when determining average confinement time. The wisdom of including or excluding these two items is a moot issue. In long-term followups of groups of felons who have been released to the community, however, the counting of pretrial time and confinement in lieu of fines will be of little significance in lessening the sensitivity of average confinement time as the basis for evaluating correctional efforts.

Adams (1961) reported that percentage of postrelease time spent reconfined was more sensitive than gross revocation rates in revealing the consequences of intensive counseling for "amenable" and "nonamenable" inmates in the PICO Project. This finding is shown in Table 3.1 as proportionate differences of the percentages in each row, but not in absolute differences of percentage, among the four groups compared by these two criteria. It

Table 3.1.—Comparative Performance of Four Cohorts in the Pico Project (from Adams, 1961)

Criterion of performance	Cohort			
	Counseled amenable	Non-counseled amenable	Non-counseled non-amenable	Counseled non-amenable
Percent with parole revoked or with unfavorable discharge from parole, after 24 months	29%	38%	43%	43%
Postrelease confinement time as percent of total postrelease time at 24 months	14%	22%	23%	26%

is apparent that impressive evidence of differing effectiveness was also provided by the dichotomous criterion of whether parole was revoked. Perhaps the greatest advantage of the use of confinement time over the dichotomous criterion here, however, was not its greater sensitivity as a measure of effectiveness, but its contribution to the estimation of public costs and benefits from alternative treatment programs. This suggests a different approach to selecting criteria to measure effectiveness.

The Most Support-Relevant Criterion

Much more rational choices in allocating government or other funds for people-changing would be possible if one could measure both costs and benefits of programs in a common currency. By subtracting the costs from the benefits for each type of client, one would know the net benefit—if any—from each program. This, of course, is how manufacturing firms or other businesses compare production or sales methods and choose among them: they subtract costs from income to determine profits. While such accounting cannot be accomplished with as much precision in people-changing as in manufacturing, it can be approximated, and it may often provide the type of evaluation most persuasive to those who decide on the allocation of government support.

The methods of estimating costs and benefits in monetary terms are so diverse and the problems they pose are so numerous and specialized, that it is appropriate to discuss them at some length in a separate chapter.

CHAPTER 4

THE PAYOFF: COST BENEFIT ANALYSIS

There are a variety of methods of estimating financial costs and benefits to society from people-changing endeavors. Each type of approach to a monetary evaluation offers some advantages and some disadvantages. Our discussion will begin with the simplest methods and will then examine more complex procedures. None is too difficult for a layman to comprehend. Discussion will focus on the major problems and issues encountered when assessment is done by what we have called the most support-relevant criterion—dollars and cents. It should be stressed from the outset, however, that the cost of crime to society and the benefit from reducing it, in terms of the anguish it causes, can never be fully represented by money—or by words. There are also moral principles that should not be overlooked when assessing the effectiveness of agencies that try to change persons alleged to be deviant. As well stated by John J. Galvin, Assistant to the Administrator of the Oregon Department of Human Resources and former Assistant Director of the U.S. Bureau of Prisons:

In all the scrambling toward modern business management concepts and methods, government people can too easily forget to include precautions against the erosion of legal guarantees, ethical principles and humanitarian ideals that have figured traditionally in the evolution of public policies (personal communication).

The Simplest Method: Supplementary Service Cost and Diminished Confinement Benefit

The easiest types of cost to estimate are the extra expenditures needed to add a particular supplemental service to an ongoing program. If a new kind of staff specialist, such as a psychologist or a speech therapist, is added at a correctional institution, clinic, or probation office, for example, it is much easier to estimate

what this costs than to determine the cost of operating the entire establishment.

To illustrate the process of determining cost per client, suppose that a psychiatrist is hired at \$40,000 per year, a psychiatric social worker at \$15,000, and a secretary for them both at \$8,000, with \$37,000 required for their fringe benefits, office space, and supplies, so that the total budget for a psychiatric unit is \$100,000 annually. If the unit provides services for 200 clients in the course of a year, the average cost of this service is \$500 per client per year. If the average stay of a particular type of client in the program is 18 months, the average cost of psychiatric services is \$750 for this type of client.

Similarly, if a vocational training program costs \$150,000 per year to operate and has an average enrollment of 150 students, it costs \$1,000 per year per student. If its services are offered in short course-units at all times, each course lasting a three-month quarter, the cost is \$250 per course per student.

More detailed information will be needed if one wishes to make more specific estimates, such as the cost for different combinations and frequencies of group and individual therapy, or for different types or terms of course-units. Such estimates for supplemental services are not only feasible, but are essential to sound fiscal planning and budget justification.

Benefit from a people-changing activity generally is measured as the estimated *reduction in social costs* of deviant behavior due to this activity. Thus, monetary benefit from a supplementary service is the difference between the social cost of subsequent deviant behavior by persons who have received this service and the social cost of deviant behavior by persons with similar histories who did not receive this service. Because total social costs are difficult to estimate with precision, it is often best to begin cost-benefit analysis by estimating the benefit from reducing one component of social costs, the expense of confining persons when they persist in serious deviant behavior. Reduction in subsequent confinement, usually the easiest benefit to assess with much precision, is one of the largest financial benefits.

To continue with the examples of a psychiatric unit and a vocational training program as the supplementary services to be evaluated, suppose that child molesters receiving psychiatric services subsequently are reconfined in their lifetime an average of two years less than similar child molesters who do not receive such services. Thus the cost of two years' confinement constitutes one monetary benefit from psychiatric services for

such clients. Of course, less reconfinement occurs in the average case because fewer repeat this crime, and this is the major social benefit of effective psychiatric treatment, but the confinement consequences of further crime are more readily assessed in dollars than in the anguish it causes. The assessment of additional benefits in monetary terms will be discussed later in this chapter.

In contrast, it is possible that recidivistic delinquents with a long history of gang involvement and repeated correctional confinement who receive psychiatric service at a State industrial school are later confined 1 year longer, on the average, than similar inmates of the school who do not receive this service (Guttman, 1963). If such are the facts, the benefit from psychiatry for them is negative, it is a loss, and consists of the cost of a year's confinement (plus other losses less readily assessed financially, such as the cost to society of the crimes that lead to their confinement).

Throughout this manual, reference to benefits is made only for somewhat specific types of client, for particular types of service, and with regard to a specific type of consequence. This style of presentation is motivated by the belief that the most fruitful evaluative research is not concerned with assessing a people-changing program for all persons in it, but instead, for particular types of client. This belief is supported by extensive evidence, repeatedly cited in this manual, that programs which claim to change people usually: (a) do not change subsequent deviant behavior for most recipients (e.g., recreation and counseling at prisons do not seem to reduce recidivism rates, but may have other functions, such as reducing tension from institutional life); (b) change some clients beneficially; (c) are counter-productive (negative in effect) for still others. What we seek by evaluative research, therefore, is prescriptive guidance about what works best, in what respect, for what type of person.

Estimating the cost of a year's subsequent confinement is most readily done by taking the total annual cost of a confinement facility's operations and dividing it by the average number of inmates there. Usually it is convenient to divide this figure by 12 to obtain cost of confinement per person per month. In practice, however, figures on the annual cost of operating an institution vary in completeness and may be lacking one or more of such elements as depreciation on buildings and equipment, prorated cost of land, and estimates of the supervision and auditing costs incurred by higher government offices because of the institution (e.g., the fraction of the budget of a State department of cor-

rections that is estimated as due to its services for a particular facility). The primary costs, however, are usually the salaries, food and other supplies, and maintenance and depreciation expended right at the institution. Estimates of confinement costs in the United States range from about \$3,000 to \$20,000 per year, depending on the region of the country and the services provided. The highest figures are at those facilities for juveniles that have extensive education and counseling staff, while the lowest figures are for prison work camps.

Some prison administrators object to the use of the average cost per inmate per year to estimate confinement expenses in cost-benefit analysis since so many of the expenditures at an institution do not change with small fluctuations in their population. To keep one more inmate for a year may add only a few hundred dollars to a \$5 million total annual cost of operating an institution that houses a thousand inmates. If cost-benefit analysis guides policy, however, its conclusions should, in the long run, affect the total population of a county, State, or Nation's correctional institutions. Thus, in many States the policy changes during the 1960's which emphasized expansion of community correctional services resulted in the closing of numerous prisons and youth custodial institutions, despite increases in State populations and of crime rates in the same period. Accordingly, if cost-benefit analysis is an effort to guide long-run policy, it is appropriate to use total annual confinement costs per inmate for the estimation of costs and benefits.

Since the subsequent confinement for those dealt with unsuccessfully at any people-changing agency will be at several different institutions, but it is difficult to procure cost estimates for a large number of facilities, it may be appropriate merely to utilize for all cases an estimate based on the most frequent place of confinement. Alternatively, when several institutions are extensively involved in the subsequent confinement, one might select the one believed to be about average in cost for those utilized. A more precise estimate could procure cost data from several different institutions and use a weighted average proportional to the relative frequency with which each is the place of subsequent confinement for those dealt with by the agency at which a program is being evaluated. Maximum accuracy would require use of separate cost figures for each place of confinement of each case followed up. It is usually best to begin with rough estimations in cost-benefit analysis, and gradually improve them.

Tables 4.1 and 4.2 provide some estimates for certain categories of clients of the Supplementary Costs and Diminished Subsequent Confinement Benefits of the psychiatric unit and vocational education program for which cost estimates already have been presented. While these figures are purely hypothetical, they are based on figures and benefits which are reported frequently.

The tables in this chapter have both a profit (or loss) figure (the difference between benefit and cost for each treatment), and an efficiency figure (the benefit-cost ratio). Efficiency, the benefit per dollar expended on cost, is a term applied to this ratio by Adams (in Glaser, 1974). He points out that it is a criterion independent of "effectiveness," if the latter is applied only to changes in behavior. It is also independent of profit. For example,

Table 4.1.—Supplementary Costs and Correctional Confinement Benefits of a Psychiatric Unit ("PsU") for Various Types of Clients in a Correctional Agency (Hypothetical Figures)

<i>Cost-benefit items</i>	<i>Type of Client</i>			
	<i>Child molesters</i>	<i>First-conviction rapists</i>	<i>Prior delinquency rapists</i>	<i>Recidivistic enculturated delinquents</i>
<i>Cost of PsU services:</i>				
Average duration of services	1½ years	2 years	1 year	1 year
Average cost per client (at \$500 per year)	\$750	\$1,000	\$500	\$500
.....				
<i>Social costs as subsequent confinement:</i>				
Average subsequent confinement time of clients without PsU services	5 years	9 months	3 years	5 years
Average subsequent confinement time of clients with PsU services	3 years	6 months	3 years	6 years
Benefits, as confinement time saved	2 years	3 months	0	-1 year
Monetary benefits (at \$4,000 per year confinement costs)	\$8,000	\$1,000	0	-\$4,000
.....				
Profit (benefit minus cost)	\$7,250	0	-\$500 (Loss)	-\$4,500 (Loss)
Efficiency (benefit-cost ratio)	10.7	1.0	0	-8.0

Table 4.2.—Supplementary Costs and Correctional Confinement Benefits of a Vocational Training Program ("VTP") for Various Types of Clients in a Correctional Agency (Hypothetical Figures)

Cost-benefit items	Type of Client			
	Youthful recidivist auto thieves	Recidivist unspecialized offenders	Alcoholic forgers	Professional robbers and burglars
<i>Cost of VTP:</i>				
Average no. of courses per client	3	6	4	3
Average cost per client (at \$250 per course)	\$750	\$1,500	\$1,000	\$750
<i>Social costs as subsequent confinement:</i>				
Average subsequent confinement time of clients without VTP	3 years	2 years	4 years	6 years
Average subsequent confinement time of clients with VTP	1½ years	1 year	4 years	7 years
Benefits, as confinement time saved	1½ years	1 year	0	-1 year
Monetary benefits (at \$4,000 per year confinement costs)	\$6,000	\$4,000	0	-\$4,000
Profit (benefit minus cost)	\$5,250	\$2,500	-\$1,000 (Loss)	-\$4,750 (Loss)
Efficiency (benefit-cost ratio)	8.0	2.7	0	-5.3

a program costing \$2,000 and producing \$3,000 in benefit would have a profit of \$1,000 but an efficiency of only 1.5, while a program costing \$200 and producing \$800 benefit would have a profit of only \$600, but an efficiency of 4.0.

Table 4.1 indicates a definite profit from psychiatric services for child molesters, who are usually highly recidivistic. Great benefits from psychiatric services for these predominantly older men who are deeply disturbed over loss of their sexual potency, have been claimed at a few State psychiatric centers for felons, notably that in New Jersey. On the other hand, much less benefit is ascribed in Table 4.1 to psychiatric services for first-conviction rapists, since they have a low recidivism rate anyhow. A benefit of three months reduction in subsequent confinement, which is

hypothesized for this category, exactly equals in monetary value the cost of the psychiatric services. Despite this zero profit, such services would be justifiable since there are benefits, other than less subsequent confinement, not taken into account here, especially the reduction of social costs achieved by diminishing the likelihood that these offenders will commit rape. On the other hand, the offenses for which there is a negative benefit, and hence a negative efficiency, have more subsequent confinement resulting after the supplementary service than without it, suggesting that the service actually increases crime.

For two categories of offender in Table 4.1, no benefit by diminished confinement time is indicated, and even an increase in subsequent confinement is hypothesized for the recidivist enculturated delinquents. It is anticipated that rapists who have a history of delinquency will be found to have a high rate of recidivism. This is because they share delinquent group-supported values which justify stealing what they desire from nonmembers of their groups whenever they cannot get it legitimately (Doshay, 1944). They seize sexual control from women because they have these values, not because of the psychological conflicts and misconceptions of sex more common among the usually sexually naive first conviction rapists.

Expectations that psychiatric services would have negative effects on recidivist enculturated delinquents, who often have much experience and pride in manipulating psychotherapeutically oriented personnel, are justified by the PICO Project data in Table 3.1, by a controlled experiment at California's Preston School of Industry (Guttman, 1963), and by the earlier experiment of the Grants (1959) with Navy offenders. With the great diversity in psychiatric approaches to such offenders, the effects may be highly dependent on the particular psychiatrist employed (Guttman, 1963).

Table 4.2 depicts a hypothesis that extensive vocational training would appreciably reduce the reconfinement of youthful recidivist auto thieves. This is inferred from evidence that their recidivism, though still high, is appreciably reduced by assistance that increases their employability. They are usually out-of-school and out-of-work adolescents who steal autos primarily for joy-riding and immediate transportation, rather than persons with a strong commitment to crime as a career (Hall, *et al.*, 1966).

Similarly, recidivist unspecialized offenders are usually disorganized persons without strong commitments to crime as a career, but who engage in it occasionally in crises. It is inferred

that extensive vocational training might increase their employability appreciably, and so diminish their financial crises. The alcoholic forger syndrome is a standard pattern in our State prisons, with a high recidivism rate of naive forgery in the course of drinking. It is inferred that vocational deficiencies are infrequent among them and are not highly relevant to their alcoholism problem. More speculatively, it is inferred that professional burglars and robbers are accustomed to a much higher income from crime between each arrest and conviction than that which most available vocational training for them would offer. They frequently take vocational courses in prison only as a gesture to impress the parole board favorably, and sometimes even to improve their skill in using burglar's tools. A small amount of vocational training might even be statistically a negative indicator in such cases (Glaser, 1969: 185-192).

The discussion and tables thus far simplify an introduction to cost-benefit analysis by ignoring several problems. One of these is the more difficult cost estimation necessary when the program to be evaluated is not a supplementary service, but a completely different people-changing procedure at a different location and perhaps in a different organization. There are still other questions that could be raised about the simplified discussion thus far, but it may be appropriate to address only the latter issue before proceeding to others.

Increasing Cost Information: Comparison of Alternative Programs

In the preceding section, when estimating costs of supplementary services such as a psychiatric unit or a vocational training program, it was assumed that these were the only cost items which differed for the clients who were compared. As long as the contrast involved the records of people in the same establishment—for example, a probation office, a juvenile training school, a State penitentiary, or a private halfway house—but with or without one service there, it was not necessary to determine all of the costs of operating these places; we just needed to know the costs of the extra service being evaluated.

In many legislative, judicial and administrative decisions on people-changing policies, the choice is more drastic than that of whether some persons should be given a supplementary service. There are decisions of incarceration versus supervision in the community, of confinement in a maximum security institution or placement in a minimum security camp, of parole or probation with only infrequent office contact required as against imposing

mandatory residence in a community correction center. These are the types of cost-benefit evaluation problem now to be addressed.

Determination of costs for alternative agencies or facilities is not necessarily more complex than the already discussed determination of confinement costs as a factor in estimation of benefits. First one tabulates the total annual expenditures involved in operating the agency as completely as possible, including depreciation of buildings and equipment (or rental charges), and adding the administrative costs which the agency creates for offices of higher authority which monitor it. Then one determines the average number of people who are clients of the agency per day. This can be done precisely as the total number of client-days in a year divided by 365 (or 366 in leap years). At any rate, dividing the total annual expenditures by this average number of people handled provides the average annual cost per client per year; dividing by 12 expresses this as a per-month figure.

Table 4.3 illustrates cost-benefit analysis for alternative rather than supplementary programs, still measuring benefits only by subsequent confinement. Tables 4.1 and 4.2 estimated benefits by comparing subsequent confinement of persons with and without the supplementary service that was being evaluated. In Table 4.3, however, to assess alternative programs, one can only compare the subsequent records of people in completely different programs, and use one as a standard by comparison with which the alternatives are considered beneficial or not. The traditional programs are usually the standard and innovations are evaluated by comparison with it. Table 4.3 uses prison and parole as the standard, and compares regular probation and intensive services probation with it in terms of diminished subsequent confinement benefits.

First-conviction armed robbers, to which Table 4.3 is limited, often are not highly committed to crime but have gotten into a financial crisis and seek to solve all of their financial problems with one holdup at an establishment which handles much cash. Often they use a toy or mock gun or an unloaded weapon. Frequently an extended stay in prison is more criminalizing for them than community correctional supervision would be, but the fact that a gun—even a toy gun—was displayed in their offense often results in their probation denial. As a group, their recidivism rates are below the average for all convicted felons. In Table 4.3 it is hypothesized that intensive services—such as supplementary paraprofessional staff, emergency financial assistance, and voca-

Table 4.3.—Costs and Diminished Correctional Confinement Benefits of Alternative Correctional Programs for Non-Addicted and Non-Alcoholic Armed Robbers on First Felony Conviction (Hypothetical Figures)

Cost-benefit items	Alternative Programs		
	Prison and parole	Regular probation	Intensive services probation
<i>Cost of program components for their average duration:</i>			
Imprisonment before first parole (at \$400 per month)	\$5,600
Regular parole and probation supervision (at \$50 per month)	\$1,000	\$1,500	\$1,600
Supplementary services (para-professionals, emergency residence, financial aid, etc.)	\$2,000
Total Program Cost	\$6,600	\$1,500	\$3,600
<i>Social costs as subsequent confinements:</i>			
Average subsequent confinement	2 years	1 year	½ year
Benefits as confinement time saved	1 year	1½ years
Monetary benefits (at \$4,000 per year average confinement costs)	\$4,000	\$6,000
Profit (benefit minus cost)	\$2,500	\$2,400
Efficiency (benefit-cost ratio)	2.7	1.7

tional training subsidy if needed—would somewhat diminish subsequent confinement, but would cost so much more than regular probation as not to be more profitable or efficient than regular probation if benefit is estimated by confinement cost alone. This brings us to the problem of achieving a full accounting for social benefit.

Maximizing Information on Benefits

We can define the monetary benefit from a people-changing method as the reduction in social cost of subsequent deviant be-

havior achieved by that method, as compared to the reduction achieved by another method. Thus far social cost has been measured incompletely, only in terms of subsequent confinement for deviant behavior. With each confinement the public also incurs costs of arrest and adjudication. In crimes with victims the personal injury or property loss suffered by the victims is the social cost of greatest concern to the general public.

When property offenses occur—and are the basis for most subsequent confinement of delinquents and criminals—the victim's loss is in money or goods for which a value is customarily estimated. Possibly the figures for known offenses of a person's criminal record should be multiplied by a factor representing the ratio of unsolved crimes of that type to known crimes.

For offenses against the person, such as assault, rape, or murder, financial expression of social costs is inadequate. Nevertheless, when decisions are made on how to compensate someone for such a loss, the public ultimately must assess the physical and emotional damage that these offenses do to the victims. Therefore, for cost-benefit analysis one may convert each type of crime of violence to an arbitrary dollar damage figure, based on insurance or court settlements of personal injury claims.

For drug offenses and other crimes not clearly victimizing persons other than the accused, it is most difficult to arrive at a social cost figure. Sometimes, however, one can make inferences about society's loss of the work power of the offender, or on his probable reliance on other crime for an income if the behavior which is criminal prevented him or her from being legitimately employed (cf., McGlothlin, *et al*, 1972: 14-15). For some misdemeanors, such as homosexuality in private among consenting adults, it is impossible to make any rational estimation of social cost in dollars. Perhaps legislators or bar association groups drafting criminal codes should make a financial estimate of social costs a prerequisite to declaring behavior subject to regulation by the law and a factor in determining justifiable penalties.

Estimates of the average costs for each arrest and for each judicial or parole processing can be made by a study of police, court, and parole board budgets in relation to the volume and type of transactions their units handle. The Cantor and Adams (1968) District of Columbia study arrived at figures of: \$17.67 for a juvenile arrest; \$88.41 for a juvenile hearing; \$15.99 for an adult arrest; \$11.86 to \$40.65 for adult court hearings; \$65.51 for a parole hearing. They cite 1963 figures for Los Angeles

of \$22 for a juvenile arrest, \$326 for a juvenile court hearing, \$100 for a Municipal Court hearing and \$200 for a Superior Court hearing. One suspects that the higher figures are the most accurate, and are still low for the 1970's. In any case, police, court, and parole board administrators should be able to make such estimations locally and doubtless have already done so in many jurisdictions.

Perhaps the most speculative component of any social cost estimation is the contribution that society loses if a potential member of the labor force and taxpayer is unemployed or if a person who would normally be supported by a family breadwinner becomes dependent on public welfare. This lost production is a social cost if someone is institutionalized. The useful work that inmates do in institutions is taken into account in determining the cost of confinement; this work makes such costs less than they otherwise would be. Therefore, in estimating the total social cost of continued deviant behavior, the estimated loss in production to the community resulting from incarceration of someone who could be a productive worker can be added to cost of confinement.

By keeping people in school, society presumably benefits in the long run. Consequently, school can be equated with employment in benefit analysis. Therefore, truancy in the community might be equivalent to production foregone for juveniles. It would be avoided with compulsory education in an institution that has effective incentives for schooling, and a rate of progress in school as high as that in the community.

The effect of maximizing information on benefits is shown in the hypothetical data of Table 4.4, which differs from Table 4.3 only in the comprehensiveness of its subsequent social cost estimations and consequent benefit, profit, and efficiency calculations. These yield figures several times as high as those of Table 4.3. Intensive services on probation are not as profitable as regular probation, shown in Table 4.3, where confinement is the sole basis for determining social costs, but are more profitable—though less efficient—than probation, as demonstrated in Table 4.4, where all benefits are tabulated. The gains in benefit ascribed to intensive services are primarily from more full employment, assuming that this would be the primary focus of such services for adults, even to the point of paying job placement charges of private agencies or subsidizing on-the-job training.

It is appropriate to initiate cost-benefit analysis by calculating benefits only from the social cost data that are readily estimated

Table 4.4.—Costs and Estimated Total Benefits of Alternative Correctional Programs for Non-Addicted and Non-Alcoholic Adult Armed Robbers on First Felony Conviction (Hypothetical Figures)

Cost-benefit items	Alternative Programs		
	Prison and parole	Regular probation	Intensive services probation
<i>Cost of program components for their average duration:</i>			
Imprisonment before first parole (at \$400 per month)	\$5,600
Regular parole and probation supervision (at \$50 per month)	\$1,000	\$1,500	\$1,600
Supplementary services (para-professionals, emergency residence, financial aid, etc.)	\$2,000
Total Program Cost	\$6,600	\$1,500	\$3,600
<i>Subsequent social costs in monetary terms:</i>			
Damage done by known and inferred offenses ^a ..	\$6,400	\$2,400	\$1,600
Cost of arrests ^b	400	300	200
Cost of court or parole violation hearings ^c	600	400	100
Cost of confinement ^d	8,000	4,000	2,000
Production foregone (at \$300 per month for time locked up or unemployed)	15,000	6,000	3,000
Total Social Costs	\$30,400	\$13,100	\$6,900
Benefit from alternative to prison and parole	\$17,300	\$23,500
Profit (benefit minus cost)	\$15,800	\$19,900
Efficiency (benefit-cost ratio)	11.5	6.5

^a At \$50 per misdemeanor, \$500 per non-violent felony and \$2,500 per violent felony.

A few repeaters make the averages high.

^b At \$100 each

^c At \$200 each

^d At \$4,000 per year

without much speculation (e.g., arrests and confinement), since policy arguments from such data can then be made without much challenge as to their validity. As lawyers learn, if one has two strong arguments and five weak ones, it is often best to present only the two strong ones, for argument over the weak ones will distract attention from the strong ones. One can then note separately that there are other items of benefit, while indicating clearly that their exact dimensions are less certain than those presented first.

The most valid data usually suffice to show which people-changing methods are the most profitable and efficient, even if they understate the extent of these advantages. A source of understatement or overstatement in all benefit estimation, however, is the period used for gathering followup information to assess subsequent social costs, which provides another issue to consider in this gradually more complex presentation.

Followup Periods in Cost-Benefit Analysis

To simplify discussion of cost-benefit analysis, we have thus far presented social cost data as though each of its components (e.g., subsequent confinement) were determined for the total remaining lifetime of the subjects. In practice, of course, any data collected on social costs would have to be for a finite followup period, usually of only a few years.

For estimating benefits as reduction in the cost of deviance, one must first determine the duration of the followup period—such as 3 years—for which benefits are to be assessed. The longer the period, the more accurate will be the assessment, but one must balance this against the disadvantages of waiting long for results. As suggested in Chapter 7, one usually seeks short-term evaluations first and long-term results later, but it may be possible to foreshadow long-term results from short-term trends.

One question usually neglected, but sometimes critical, is that of when the followup period should begin. If the programs compared are all in the community or all involve institutionalization for the same average duration, then the followup of clients should begin at the date of their release to the community. This assumes that the benefit to be estimated is the reduction of deviance in the community, which they all reenter at about the same time after their apprehension for deviant acts. If, however, the programs compared differ in the average duration of initial institutionalization they impose, then there are sound objections to

comparing deviance rates only from the dates of client release to the community for all programs. It would be a gross disservice to science and to the public interest to ignore, because of our biases against confining people, the one accomplishment that can be claimed for institutionalization: it prevents those confined from committing offenses in the community during the period of their incarceration. The research task is to determine whether this is worth the cost, which may or may not include more deviance in the long run after release, as well as greater financial costs in the short run during confinement.

From the foregoing point of view one can argue that followups for cost-benefit analysis of programs imposing different durations of confinement should follow clients in all programs from the time they begin to experience different types of people-changing endeavor, either in the community or in confinement. In California's Community Treatment Project, for example, the experimental subjects are paroled after about one month's confinement in a reception center, while the control group members are paroled from Youth Authority institutions under traditional procedures, which has meant an average of eight months' confinement before parole. The followup period used in comparing their infractions, however, began with parole for each group, to give them all the same period of "community exposure." (The followup periods were 15 and 24 months on parole in both the Warren 1966 and the Palmer 1971 reports, for example.)

While comparison for similar periods in the community is interesting for theoretical purposes, it is inappropriate for cost-benefit analysis if the groups compared differed in average duration or cost of prerelease confinement. It is misleading, I believe, that a net cost-benefit profit is implied in the Community Treatment Program reports when the researchers observe that: (1) community treatment as intensive as that given the experimental cases costs \$2,300 per year per client; (2) regular Youth Authority parole costs \$400; (3) the cost of confinement in Youth Authority institutions is \$5,800; and (4) reincarceration while on parole is more frequent for control than for experimental cases (Palmer, 1971: 86-87).

I would contend that to estimate societal costs and benefits from the Community Treatment Program, the followup period for *both* experimentals and controls should begin when they are committed to the supervision of Youth Authority. At the end of their approximately one-month stay in the Reception Center, the experimentals are all released to the community and begin to

incur less treatment costs but have more opportunities for new offenses than the typical control case, who is locked up for another eight months, on the average. If the experimentals had less reconfinement than the controls in the long run, however, and created less other social costs from new offenses and from idleness, there would be benefit from their lower social cost than the controls. Yet there might still be a net loss and less than unitary efficiency because of the greater cost of community treatment as compared with parole.

The foregoing possibility is illustrated in Table 4.5. It compares 15-month "community exposure" data published for the institutionalized Control Group with 24-month "community exposure" data for the Experimental Group, since with the eight months average duration of initial confinement of the controls the total period of Youth Authority control is about 24 months for both groups. By these only partly hypothetical treatment and social cost data on the most successfully treated large category, the neurotic anxious delinquents, social costs of the two programs are identical at 24 months (due to the 8-month confinement of the controls) but are greater for the controls at 60 months. This assumes a post-discharge followup is conducted, since average duration of total Youth Authority supervision is reported to be about 3 years for both groups. However, the greater cost of the community treatment program as compared with that of regular parole makes it unprofitable at 60 months despite its lesser social cost. This might not occur in actuality if the intensity of services were greater in the first year or two of community treatment than later, in the average case; the published cost data give only an average for all months and all cases. Since the Community Treatment Program began in 1962, a long-run cost-benefit analysis is overdue, but it will require different types of tabulation than have thus far been reported.

Societal and Community Cost-Benefit Analysis

Cost-benefit analysis is not only a useful method of assessing specific programs of people-changing agencies, but it is also a useful frame of reference for appraising national, State, or community policy on almost any public problem. For such matters as crime, narcotic addiction, mental illness, physical illness, and educational deficiency, for example, the government must decide how much of its funds to allocate to each method of combating the problem—detection, treatment, research, and public information. Because the objective is to give the funds available the

Table 4.5.—Cost-Benefit Analysis of Experimental (Community Treatment) and Control (Youth Authority Institution and Parole) Programs for Neurotic Anxious Male Delinquents, Using 24-Month and 60-Month Followup Periods From Date of Arrival at the Reception Center (Data from Palmer, *et al*, 1968, plus extrapolation and hypotheses)

Cost-benefit items	Control		Experimental	
	24 months	60 months	24 months	60 months
<i>Treatment costs:</i>				
Reception center (1 mo.)	\$ 800	\$ 800	\$ 800	\$ 800
Correctional institution (at \$500/month)	4,000	4,000
Regular parole supervision (at \$35/month)	385	945
Community treatment (at \$200/month)	2,600	7,000
Total Treatment Cost	\$5,185	\$5,745	\$3,400	\$7,800
<i>Subsequent social costs in monetary terms:</i>				
Damage done by known and inferred offenses ^a	\$ 500	\$2,000	\$ 400	\$1,000
Cost of arrests ^b	100	400	200	400
Cost of court or parole violation hearings ^c	100	300	100	200
Reconfinement costs ^d	500	2,000	400	2,000
Production (or education) foregone ^e	500	2,400	600	1,800
Total social costs	\$1,700	\$7,100	\$1,700	\$5,400
Benefit from community treatment program	0	\$1,700
Profit (benefit minus cost)	-\$3,400 (Loss)	-\$6,100 (Loss)
Efficiency (benefit-cost ratio)	0	0.2

^a At \$50 per misdemeanor or incorrigible act, \$300 per non-violent felony and \$2,000 per violent felony (even if adjudicated delinquency rather than felony). A few repeaters make averages high.

^b At \$100 each

^c At \$200 each

^d At \$600 per month (based on \$800 for detention and \$500 for correctional institution, with experimentals confined less time, but more often in detention).

^e At \$300 per month, not charged for total months of normal progress in schooling when not gainfully employed, either in the community or in correctional institution school programs.

greatest total impact, it is useful to estimate the marginal benefit of each additional expenditure for each method.

Presumably there is a point of diminishing returns for each method of coping with a people-changing problem. Therefore, one should strive to estimate the fiscal appropriations to each that will make the last dollar spent for any method yield about the same benefit as the last dollar spent on any of its alternatives. In practice, of course, there are many unknowns, especially on the benefit per incremental dollar spent for research and for public information. Consequently, some gambling must occur in these expenditures; the amounts spent must often be decided only by very rough guesses as to their yield per dollar.

Estimated annual costs and benefits per year of opiate addiction treatment in the United States are presented in Table 4.6. These figures are based on assumptions that current laws and law enforcement against opiate possession and sale will persist, so that most persons addicted to heroin will have to support this habit by crime. It assumes that the therapeutic community and

Table 4.6.—Estimated Annual Costs and Benefits Per Year Per Addict From Opiate Addiction Treatment Programs in the U.S., Under Current Criminalization of All Heroin Possession and Sale (Based mainly on McGlothlin, *et al.*, 1972)

<i>Cost-benefit items</i>	<i>No treatment</i>	<i>Methadone maintenance plus aid</i>	<i>Therapeutic community</i>	<i>Civil commitment</i>
No. of addicts in U.S., Dec. 1971 ..	^a 308,400	^b 40,000	8,000	18,400
Treatment Cost	\$1,500	\$3,500	^c \$3,800
<i>Social costs:</i>				
Crimes	\$10,000	\$3,500	\$ 500	\$1,755
Anti-crime measures	1,250	435	65	250
Foregone production	3,450	2,300	2,300	2,875
Total social costs	\$14,700	\$6,235	\$2,865	\$4,880
Treatment benefit	\$3,465	\$11,835	\$9,820
Profit (benefit minus cost)	\$6,965	\$8,335	\$6,020
Efficiency (benefit-cost ratio)	5.6	3.4	2.6

^a Includes 37,500 incarcerated, 40,000 on the street but temporarily abstinent, and 2,500 in temporary detoxification facilities. An additional 200 were believed to be in experimental opiate antagonist programs. Addicted persons in the street were estimated as 228,400, and the grand total of addicted persons in or out of treatment as 375,000.

^b Includes civil commitment outpatients on methadone maintenance.

^c Assumes two-thirds on aftercare at \$1,700 per annum and one third in institutions at \$8,000 per annum.

methadone maintenance programs, which must recruit their clientele as volunteers, are relatively unrestricted in recruitment and admit marginally persistent addicts in addition to those with whom it will be most successful. (The latter are those who survive a long-waiting list and other entrance ordeals.) While the figures are hypothetical, they are far from wild guesses, being based on a careful analysis of all relevant factors by McGlothlin and associates (1972). Their figures are modified only slightly here, mainly for simplification, to reduce the number of treatment alternatives considered, thereby achieving a comparative analysis in one table.

McGlothlin and associates use different "no treatment" social cost estimates for each type of treatment at each size of patient load. For example, they assume that a methadone maintenance program plus aid would recruit only advanced addicts in its first 100,000 patients, so the average "no treatment" crime costs would be \$14,000 for these addicts, although it would be only \$10,000 for the less advanced addicts recruited by therapeutic communities and civil commitment programs. We have simplified the tabulations by employing for all treatment modalities what McGlothlin and coworkers estimate as the social costs of all untreated addicts in the United States. Their estimates of these costs are yearly averages based on their inferences from evidence of the changing intensities of addiction during the lifespan of the addict, the cessation of addiction with age, and the high death rates among addicts.

Treatment cost figures are highly variable. McGlothlin and associates estimate that Methadone Maintenance Plus Aid would cost \$1,000 per addict for the first 50,000 addicts per year treated, \$750 for the next 100,000, and \$1,500 each for the next 25,000. The increase is attributed to the special inducements presumed to be needed to recruit and retain additional addicts after 150,000 are in the program. Table 4.6 takes this highest figure. They estimate \$4,000 per addict per year as the cost of operating a therapeutic community which has no profitable business or members employed at outside jobs. The cost diminishes to \$2,500 as these sources of income are developed. The pioneer therapeutic community, Synanon, became self-supporting, but most others obtain their support mainly from government agencies. For civil commitment programs they report diverse costs: for inpatients the estimates are \$7,000 to \$9,250 (depending on facility) in New York, \$12,000 in the Federal hospital at Lexington, and \$4,000 in California; outpatient estimates are \$1,750

in New York, \$4,800 in the Federal system, and \$850 in California. The figures in Table 4.6 reflect the impression that the New York estimates are the most typical, and the fact that at least 40 percent of the addicts in the United States are in New York.

Value constraints of both government officials and addicts prevent our following exclusively what Table 4.6 indicates would be the most efficient procedure—giving all addicts in the country methadone maintenance plus aid. In Washington, D.C., when all waiting lists for any treatment were eliminated, only 60 percent chose methadone maintenance. Most of the remainder of those treated requested only detoxification, which permits them to become abstinent without withdrawal symptoms. In practice detoxification is sought by addicts mainly to reduce the dosage they require to remain addicted, and thus to reduce the cost of their habit. A majority leave the detoxification centers against medical advice, before they are detoxified to the point of complete abstinence. Furthermore, it was estimated that only 20 percent of Washington's addicts were in any of the treatment programs when all were made readily available (McGlothlin, *et al*, 1972: 17).

Most public officials seem to favor complete and voluntary abstinence, which is the objective of therapeutic communities. Therefore, these establishments gain public support despite their greater cost. Actually, their average rate of retention is low, only 29 percent of addicts remaining in treatment at the end of a year after their admission, as compared to 65 percent for methadone maintenance (Joe, *et al*, 1972: 30). Taking this low retention rate into account, the social cost figures for therapeutic communities in Table 4.6 (based mainly on McGlothlin *et al*) probably are somewhat low, and thus exaggerate profit and efficiency for any cross-section of persons admitted to them. Presumably the figures are appropriate for the minority of those admitted whom these organizations are able to retain for a year or more.

Civil commitment programs are conducted in buildings that are essentially prisons—a large proportion having been originally constructed for secure custody of criminals—but are staffed with a higher proportion of psychotherapists and other treatment specialists than almost any prisons. Thus they have high treatment costs during the period of confinement but low social costs, since the addicts detained cannot commit crimes in the community to support their habit. Confinement is followed by aftercare, which

is modeled on parole and is staffed mainly by former parole officers, but with more assistance and surveillance resources in the community than is customary on parole.

Addicts in civil commitment status are under State supervision for 3 to 5 years in New York and for 7 years in California. California originally discharged about one-sixth of its addicts when they had been drug-free on aftercare for 3 years, but the initial New York experience was much less successful. In recent years, however, both States have relaxed standards for retention on aftercare and discharge. Both have also increased the availability of methadone maintenance for civilly committed addicts, especially in New York. Before these changes about one-third of those placed on aftercare in New York disappeared, as did about one-fifth in California. Most of the absconders are presumed to be readdicted, and it is known that many of those retained on aftercare use small doses of heroin intermittently. For these reasons, it is believed that the figures in Table 4.6 probably exaggerate the profit and efficiency of civil commitment when it is not combined with methadone maintenance.

The most serious omission from available data on the effectiveness of alternative treatments for opiate addiction, in this writer's opinion, is the failure to follow up addicts committed to regular correctional institutions after conviction for felonies. Parole outcome data suggest that they are more successful in achieving abstinence after traditional prison and parole than after civil commitment programs in the New York State, California, and Federal systems. This may be due to the concentration of a purely addict society in the civil commitment institutions, which perpetuates discussion of drug lore. As a matter of fact, smuggling of heroin and other drugs into civil commitment facilities is regularly reported by inmates and staff. Heroin is allegedly used there by many inmates who at the same time verbalize vehement antidrug attitudes to manipulate their therapists to place them on aftercare.

A further omission is that of knowledge about what the social costs of addiction would be if heroin maintenance were available to confirmed addicts who did not wish to enter methadone or drug-free therapeutic programs. This is the practice in Britain and many other countries, and its main advantage is that addicts do not have to steal to procure opiates for their own use. Allegedly many still steal, partly because heroin usage is not as compatible with employment as methadone. That is why the British program, which makes heroin available to addicts on demand, still tries

to induce them to change to methadone and ultimately to become abstinent (McGlothlin, *et al*, May 1972; 33-39).

Conclusion

No cost-benefit analysis can be absolutely precise and satisfactory for several reasons. Foremost, of course, is that conversion of anguish into dollars is necessarily at arbitrary rates. Secondly, one never knows of all the deviant acts that releasees commit; One only knows those for which they are apprehended and shown to be guilty. Nevertheless, benefit estimations alternative of simply leaving assessment of crime reduction up to subjective impressions, since such impressions are based on even less complete information that is less systematically tabulated and analyzed. While postrelease crime or other deviance data are incomplete, and dollar assessment rates are arbitrary, it is still reasonable to compare releasees from different programs in terms of such data. One can assume that the degree of incompleteness of data on deviance and the arbitrariness of dollar assessment rates are similar for all groups compared. This is analogous to the customary procedure and assumptions in classifying criminals by prior record (as first-, second-, or third-time offenders, for example), since such classifications also do not take into account the crimes not resulting in apprehension.

This chapter has described a variety of approaches to cost-benefit analysis, with many degrees of complexity. It is most conducive to the routinization of this type of evaluative research if one begins only with that method of estimating costs and benefits that one can do well. This should reveal in a most support-relevant manner those programs which are clearly profitable and efficient. More refined analyses can then be added gradually to provide less crucial additional details.

CHAPTER 5

COMBATING SPURIOUS CRITERIA: DEFENSES AGAINST SNIPING

A common complaint about criminological research is that it has been much more successful in discrediting myths than in establishing valid and precise knowledge of the causes and cures of crime and delinquency. Destroying myths is no mean accomplishment, however, especially when these beliefs have been the basis for ineffective expenditures of billions of dollars or for unwarranted imposition of great hardships on many persons.

While the main thrust of this manual is to set forth effective ways of conducting and routinizing evaluative research, an important aspect of routinizing this type of endeavor is to defend it against unjust and unwarranted criticism. Accordingly, it is appropriate to alert researchers and administrators in people-changing agencies to the types of illegitimate evaluations they are likely to encounter, and to counsel them on the best defenses against unfounded attacks.

The objective of evaluative research is to replace myth with reality in the guidance of policy and practice, but myths have an impressive tenacity and they arise in support of every side of an argument. The most difficult to recognize as spurious are those supporting one's own preferences. In being alert to defend oneself against unwarranted criticism, therefore, it is important to avoid making claims that are no more legitimate than those one opposes.

Individual Cases versus Statistics as Evidence

Perhaps the most commonly encountered spurious method of evaluating people-changing efforts is to draw conclusions from individual cases. Whether these are one or two dramatic cases or long lists of them, and whether they are successes cited to support a measure or failures cited in opposition, they provide no very conclusive evaluation of a people-changing method. Every

method has both success and failure cases. Evaluation can only be credible if it is expressed as a percentage, rate, correlation, or other statistical conclusion, although this type of formulation does not in itself guarantee that the evaluation is correct. Sometimes, indeed, it is not clear that a particular treatment is responsible for the outcome of the cases treated, regardless of the outcome statistics.

Individual cases *illustrate* how a method works or fails, and *suggest* explanations for its outcome, but only the statistical pattern can *demonstrate* how effective it is. Because all methods have both successes and failures, any method may evoke testimonials for or against it. One can generalize about a method's relative effectiveness only by comparing its outcome rates with those of another method with presumed traditional rates, or with an imaginary completely successful or unsuccessful method.

The more frequently evidence from separate studies of the same thing yield similar results, the more confidence is warranted in the validity of their conclusions. The possibility that some qualification is appropriate to any conclusions, however, should always be recognized in evaluating people-changing enterprises. One seeks to identify consistent outcome patterns, but some variations are inevitable in the effectiveness of any treatment method, type of client, place, or other variable, due to the many uncontrollable processes and events that can affect complex behavior.

Despite the need for statistical data and despite some normal variation in such data, every people-changing official will be confronted with individual cases cited to support or oppose each policy over which there is any disagreement. For example, whenever a prisoner commits a serious crime while on furlough to arrange housing or employment shortly before his regular parole or discharge date, the practice of granting furloughs will be denounced by local officials where the crime has been committed. In California, when this occurred in 1972, the Government pointed out that thousands of furloughs had been granted though only 1 percent of prisoners failed to return, that those on furloughs were soon to be paroled anyhow, and that without furloughs the men would be more likely to commit new offenses soon because they would be unemployed more often. In Florida during an 8-month period ending June 30, 1972, 18,313 furloughs were granted and only 46 prisoners failed to return, or one-quarter of 1 percent (*Correctional Compass*, Florida Division of Corrections, Nov. 1972). Better than 99 percent success rates have also been re-

ported in the Federal prisons, and in Michigan, Mississippi, and many other States. Despite this impressive record, any serious offense by someone on furlough—even by someone who would soon have been released by parole or discharge anyhow—places the entire furlough program under attack. While this high rate of return without crime from furlough to prison is a powerful defense against criticism that cites individual cases, a much more adequate defense would compare rates of crime after parole or discharge for prisoners not receiving prior furloughs with the postrelease crime rates of similar prisoners who had prerelease furloughs.

By 1971, in the District of Columbia, eight community halfway houses had been opened over a 2-year period. Prisoners were released by transfer to these centers before their parole or discharge and there was much uproar whenever someone was accused of committing a crime in the city while resident in one of these centers. A police lieutenant was mentioned in the newspapers as completing a research report on these centers, but the report consisted only of a list of residents of these centers who had been arrested in a preceding period. The Department of Corrections, unfortunately, attacked this spurious research only by checking this list and identifying those arrestees who had been released by the courts, implying that the arrest had been unwarranted, and that residents of these houses were subjected to undue police harassment. A more appropriate criticism could have been provided by the Department's "Research Release" of March 15, 1971, which reported that within their first 8 months of community exposure during 1969-71, 72 percent of center releasees had no further legal difficulties of any sort, compared with 56 percent of institution releasees; the 12 percent recidivism rate of new convictions from the community centers was barely over one-half the institution's rate of 23 percent; for releasees without histories of drugs, alcoholism, or physical handicaps, the recidivism rate from the centers was 6 percent, barely over one-third the institution's rate of 17 percent.

There will always be failures, sometimes dramatic, in any effort to change persons so that they cease their deviant behavior. These failures frequently will lead to attacks on those responsible for the people-changing endeavor, and to demands for alternate methods. The attacks may or may not be warranted, but their validity will be neither proved nor disproved by individual cases of failure, no matter how numerous or how glaring. Only the failure rate of the challenged method compared with the failure

rates of alternative possible methods will indicate the effectiveness of the method in question. This indication will be most adequate, of course, if rates are expressed as continuous variables, preferably as profit and efficiency in costs and benefits.

Inappropriate Statistics

There are innumerable jokes about statistics, or expressions of distrust towards them, such as the assertion, "Figures don't lie, but liars figure." Errors, whether wilful or unwitting, can be made in any observation, statistical or nonstatistical. The most sensible reaction to the possibility that statistical statements can be misleading is to be alert to the types of errors they may contain, rather than to reject all statistics.

Some of the most common errors in remarks containing statistics are not in the statistics themselves, but in the words used to present and interpret statistics. For example, a police chief asserted in a national publication that ". . . 23 percent of our homicide, 40 percent of our burglary, and 44 percent of our robbery convictees were on active probation or parole. These figures demonstrate that our own local probation concept is not working." One blatant error here, of course, is to confuse probation with parole, in a State where almost all persons convicted of the crimes cited by the chief receive a maximum sentence of life, so their release from prison is always on parole. These statistics may reflect primarily the rates of recidivism from prison, but it was only the probation system that was being attacked.

The second blatant error in assertions such as "23 percent of our homicide . . . convictees were on active probation or parole" is that this percentage is calculated from the wrong base figures for guidance of policy. The appropriate statement would not be the percentage of homicide convictees who were on probation or parole, but the percentage of parolees and probationers who commit homicide. This is only a fraction of 1 percent; indeed, much too low a base rate to be readily altered by any change in probation or parole policy.

For the prevention of homicide one would have to procure more focused statistics than those for all parolees or probationers. A first step would be to get homicide rates for parolees and probationers separately. A second step might be to obtain these rates separately for different types of probationers or parolees. Thus, homicide rates for parolees with a history of violent acts with lethal weapons might be procured, and these might be further subdivided by age and residential neighborhood. The objective

in such focusing analysis is to determine what types of offender under what circumstances have the highest rates of subsequent homicide. Even when this is done, the prospects are still small for a significant reduction of homicide or other violence by longer confinement or by special programs for those with a higher than average probability of repeating violent acts. This is the situation because thus far researchers have been able to classify in advance just a small proportion of total offenders as having the highest probability of persistence in crimes of violence; only about one-seventh of convicted violent offenders are again convicted of violent crimes after release. Furthermore, the special treatment methods thus far tried for them have not markedly altered their relatively low rate of repeating violent crimes (Wenk, *et al*, 1972). What the facts indicate, therefore, are that most of these offenders are indeed deterred greatly by present penalties, especially the vast majority who do not live by their violence, and that we should focus instead on the correlates of first arrests for criminal violence—low education, segregation, and readily available lethal weapons—if we wish to reduce violent crimes greatly.

Another illustration of misdirection with statistics is supplied by the police chief we cited (actually, he is one of the better police executives in this country). The chief writes in a newspaper article: “. . . in 1960 . . . county courts sent 34.3 percent of our convicted murderers, robbers, and burglars to State prison and . . . by 1970 that figure had dropped to 17.6. . . . Thirty-four percent is pretty low . . . but 17.6 percent is incredible.” This might be called the “mixing-apples-with-elephants” type of error. To highlight the distortion in the quoted statement, it may be appropriate to point out that by the same logic one could assert, quite accurately, that over 99 percent of humans, dogs, and insects have six legs. This is true, since there are well over a hundred insects in the world for each human or dog.

The number of persons arrested for burglary in the United States is about 23 times the number arrested on nonnegligent homicide charges, and robbery arrests are about 6½ times as frequent as these homicide arrests. Over half of those arrested for burglary are under eighteen years old, and therefore very few of them go to State prisons. By mixing burglary, for which a small percentage go to prison, with the rarer offense of murder, for which almost all those convicted go to prison, the chief conveys an image of courts turning murderers loose. Though I happen to share his critical views on plea-bargaining in our courts,

no sound ground for criticism is provided by his statistics, and they certainly are not relevant to any solution. Separate statistics on court dispositions by offense, age, and prior criminal record would be a first step towards understanding this problem, especially if they focused on the diversity of judicial practice with similar offenders.

Other possible misuses of statistics are so numerous that a separate book would be necessary to catalogue and discuss many of them. Most textbooks on statistics point out common sources of error (see, for example, Huff, 1954; Zeisel, 1968). Because of our many other concerns in this volume, the only further comment on this topic is but the admonition that statistics, like words, must be considered carefully. One must comprehend how statistics were gathered and what alternative analyses of them are feasible, if one is to determine their logical implications for public policy. A major argument for statistics, critically examined, is that they provide a check on a widespread tendency in people-changing efforts to find solace in assertions not challenged by the procurement of facts. This tendency is our next concern—reliance on verbal pontifications instead of on evidence.

Expert Opinion

In people-changing endeavors there is a tremendous cult of the expert. Public officials and the general public, when they decide that they should do something about crime, delinquency, drug abuse, alcoholism, mental illness, or other deviance, customarily propose calling in experts to tell them what to do. Frequently it is a way of ignoring a problem: "We'll hire a psychiatrist and let him take care of it." Unfortunately, there are many varieties of presumed expert, there is great diversity in their expertise, and there is no simple way to determine their qualifications.

Policy discussion on people-changing frequently becomes a "battle of the experts," with faith in some spokesmen and not in others determining the winner, rather than the evidence for what they say. In discussions of policy to reduce delinquency and crime, one regrettably finds officials in different components of the criminal justice system—the police, the courts, and correctional agencies—and sometimes independent consulting psychiatrists all pitted against each other in talking to the general public and legislative officials. To these are increasingly added ex-convicts, and for drug-related offenses, ex-addicts. Though they all proclaim the same objectives of protecting the public and rehabili-

tating the offenders, each offers a somewhat different policy preference based on different assumptions, experiences, values, and theories.

Persons who have job interests or an ideological commitment to a particular type of treatment are especially resistant to evidence that does not support the endeavors they favor. Such evidence they dismiss as "inconclusive," but they welcome any similarly inconclusive statistics that support their position. Of course, all knowledge of treatment effectiveness is imperfect and can be both criticized and improved. What we have to rely upon other than expert opinion, preferred theories, or personal impressions is only a preponderance of the evidence from evaluative research. When such research repeatedly yields data that show consistent patterns, our confidence in the conclusiveness of the evidence grows.

If research evidence is incompatible with our expectations, it is natural for us to look with an especially critical eye for defects in the research procedures. Unless we can clearly show that defects in research procedures probably account for the deviation of findings from expectations, however, contrary evidence should make us lose some of our prior confidence in the theories, impressions, or presumed expertise on which our expectations were based. Of course, if we can show that there are serious and relevant defects in the research, it is appropriate to specify what would be a more rigorous research design. Only from the continual interaction of evidence and inference does scientific opinion progress.

Conclusion

It follows from the foregoing that the only defense against spurious statistics, questionable criteria of effectiveness, and reliance on untested claims to expertise is the institutionalization of evaluative research as a continuous component or concomitant of all people-changing endeavors. This institutionalization—which is here called "routinization"—is the concern of this entire manual. That the cult of the expert and reliance on individual case data frequently misguide people-changing efforts will be further illustrated, especially as an introduction to our next problem: For evaluation, what rates should be compared with what other rates?

CHAPTER 6

EXPERIMENTS, PSEUDOEXPERIMENTS AND QUASI-EXPERIMENTS: WHO SHOULD BE COMPARED?

“Expert” Opinion and Illustrative Cases versus Controlled Experiments: An Example

One of the most dramatic examples of the cult of the alleged expert in people-changing, and of the determination of practice by isolated illustrative cases instead of by relevant statistics, has been the movement for megavitamin therapy. Since the late 1950's, large doses of various vitamins, especially the B-complex nicotinic acid derivatives such as nicotinamide, have been recommended as wonder cures for schizophrenia. Some enthusiasts have even claimed great prospects for the megavitamin treatment of alcoholism, drug addiction, crime, and a variety of other types of deviant behavior, which proponents of the treatment ascribe to an imbalance in neural chemistry.

With schizophrenia especially, numerous illustrative cases of impressive success with megavitamin therapy have been reported. So persuasive were the claims from such cases that the American Schizophrenia Association and other organizations raised hundreds of thousands of dollars to promote this approach, and it has been widely publicized in newspapers and magazines. Although most psychiatric opinion initially opposed it, by the early 1970's nicotinamide was frequently prescribed for schizophrenia, in addition to other therapies, especially if the patient's family requested it.

Although many cases of success and many of failure in treating schizophrenia with nicotinamide could be cited, there were several reasons why neither of these types of illustration constituted strong evidence for the effectiveness or the ineffectiveness of this treatment. In the first place, the symptoms of schizophrenia are often vague and variable. Experts frequently disagree in diagnosing it, partly because many persons show

fairly clear schizophrenic symptoms at times but lack them at other times. Thus much apparent recovery could be only a temporary remission, which may occur without any medication, or it could reflect the subjects, having been erroneously diagnosed as having had this ailment. Secondly, the nicotinamide was generally provided in conjunction with other types of therapy. Even telling the patient and his family that the vitamins should help recovery might conceivably be a form of therapy because of the power of suggestion. Therefore, even when the patient seemed to improve after nicotinamide treatment, one could not be certain that the vitamin intake was responsible for the change, rather than other forms of therapy.

The achievements of organic medicine, especially the discovery of antibiotics for curing or preventing infectious diseases, are in a very large measure a consequence of controlled experiments to determine the effectiveness of alleged therapeutic agents. In psychiatry, clinical psychology, and other disciplines concerned with changing deviant behavior, the use of controlled experiments to test therapeutic methods is much rarer. Therefore, despite the tremendous investment in research on mental health in the United States and the extensive controversy over nicotinamide treatment of schizophrenia, our National Institute of Mental Health's *Schizophrenia Bulletin* had to refer to an Irish study to find a controlled experimental test of this treatment.

The account by McGrath and associates (1972) of this Irish research provides such a clear illustration of the ideal method of evaluating people-changing procedures that it may be well to quote at length from this account before discussing experimental methods further.

As nicotinamide appears to have few side effects and does not interfere with other medication, a very simple experimental design was adopted. During a specified calendar year, every admitted patient who was diagnosed schizophrenic by the senior psychiatrist-in-charge at each of the four participating hospitals was included in the study and allotted a code number. Each patient was given, three times daily, two tablets which contained either 500 mg nicotinamide (niacinamide) or an inert substance. The allotment of these active and inert tablets was made on the basis of sets of randomized numbers. Since the identical-appearing tablets were supplied to the hospitals in containers bearing only the patient's code number, neither hospital staff members nor patients knew who was receiving nicotinamide and who was receiving placebo.

In addition to nicotinamide or placebo, each patient received the "normal treatment" for schizophrenia as prac-

Table 6.1.—Final Assessment of Patients Treated for One Full Year With Nicotinamide for Schizophrenia (From McGrath, *et al*, 1972:76)

Final assessment	Patients receiving nicotinamide		Patients receiving placebo		All patients	
	No.	Percent	No.	Percent	No.	Percent
Recovered or much improved	58	65	68	72	126	68
Improved or not improved	31	35	27	28	58	32
Totals	89	100	95	100	184	100

ticed in the hospital which had admitted him. . . . Full patient-progress notes were kept, and an itemized summary sheet was completed shortly after each patient's admission and again at the completion of the trial. If a patient was discharged from the hospital, treatment continued on an outpatient basis. . . .

As is true in any treatment program for schizophrenic patients, there was an inevitable "dropout" of an appreciable proportion of cases. This loss of patients was due to many factors, including failure to cooperate and moving to a different geographical area. There was no significant difference, however, in the dropout rate for patients on nicotinamide and for patients on placebo. . . .

The final assessment of patients who took the tablets for a full year and who were assessed at the end of that period is given in Table [6.1]. The difference in favor of the placebo group was not significant ($X^2 = .87$, $p > .10$).

In summary, this double-blind collaborative study failed to demonstrate any therapeutic effect of the addition of nicotinamide (3 g per day for 1 year) in the treatment of a consecutive series of 265 schizophrenic patients.

The conclusions of this experiment may, of course, be modified eventually on the basis of new experiments, which perhaps will indicate a specific type of client whom some megavitamin dosage may benefit, but it is quite clear that the knowledge provided by one good controlled experiment is much more conclusive than the suggestions provided by thousands of illustrative cases or hundreds of alleged expert opinions.

Blinding, Sampling, Randomizing, and Matching

A number of the features of the Irish experiment described above merit special attention. The first is simply that the researchers did not follow any of the practices customary when people-changing agencies decide to try out a new form of treatment. They did not provide megavitamin therapy to every schizophrenic patient, to every patient who volunteered for it,

or to every patient for whom staff recommended this form of treatment. To have followed any of these practices customary with new modes of treatment would have greatly impeded finding out whether this new approach had any effect on the course of schizophrenia. Instead, the Irish researchers chose only to supply the vitamin treatment to half of the newly received patients whose condition was diagnosed as schizophrenia, and only for 1 calendar year.

By supplying the new treatment only to newly received patients, the researchers presumably minimized the possible masking of vitamin effects by the effects of hospitalization on patients. Through supplying the vitamins only to half of the newly received schizophrenic patients, they provided themselves with a control group of the other half, who did not receive the massive doses of nicotinamide, and could thus be compared with the experimental group that received this vitamin. By terminating acceptance of new cases to the experimental and control groups at the end of 1 year and waiting an additional year for followup data on the last cases they permitted assessment of consequences of this vitamin therapy after a uniform followup period of 1 year for all cases.

By "blinding" in evaluation research we refer to not letting participants in a people-changing program know that they are involved in a special treatment or evaluation, or at least, not letting them know the exact nature of their involvement, such as whether they are in the treatment or the control group, or that there are two such groups. These types of blinding are not always possible and may not even be desirable in many evaluative research endeavors, but in the Irish experiment they could all be accomplished because the vitamin and the placebo tablets looked alike and tasted alike, and each patient's package was identified only by a number. It was a "double-blind" experiment, in that neither the treatment staff nor the patients knew what the numbers meant; they did not know which patients were receiving placebos instead of vitamins. The researchers checked the records to determine who was in each group only after they had collected from treatment staff and classified all of the 1-year followup evaluations of patient progress.

When single-blinding is employed rather than a double-blind procedure, the research subjects do not know who is in the experimental and who is in the control group, but the treatment staff know this. If such knowledge had prevailed among staff in the Irish experiment, so that they knew which patient was get-

ting the vitamin and which was receiving the placebo, they might inadvertently have communicated this knowledge to the patients. This communication might have affected the patients' recovery independently of any possible organic effects of the vitamin. Also, even if the patients did not learn whether they were in the experimental or control group, the staff's knowledge of these groupings might have affected how patients were treated or how their progress was assessed. Whenever subjects or staff are aware that they are in an experimental or a control group, the possibility that this knowledge is a major influence in subsequent changes or diagnoses of behavior should be considered. (For examples of such effects, see Rosenthal, 1966.)

The Irish experiment involved sampling only in the sense of sampling 1 year out of all possible periods of time, and in the sense that the participating Irish hospitals may be considered a sample of all mental hospitals. More often we wish to sample in a more restrictive sense, that of limiting the research to a fraction of the total number of cases that might be available for study. For example, if the Irish experimenters had a hospital system which received 2,000 schizophrenics per year and their research budget required them to limit the vitamin study to only 100 treatment and 100 control cases, they would have to select as their research sample only one-tenth of the schizophrenic patients whom their hospitals received during 1 year. There are several alternative procedures that could be followed in selecting such a 10 percent sample, each procedure having some advantages and some disadvantages when compared to the others.

The simplest method of selecting a 10 percent sample might be to take every tenth case that is received and make it one of the research sample, which would be further divided into experimental and control groups. Where each potential subject receives a registry number from the people-changing agency, has a Social Security number, or is simply numbered serially by the researcher, one may select a 10 percent research sample by selecting all clients whose number ends with a certain digit—for example, "3" or "7" or any other number. For a 20 percent sample one may use two numbers—for example 3 and 8, or any two numbers, and for a 30 percent sample use three different arbitrarily selected last digits. For a 5 percent sample one could take all cases with both a particular last digit and an odd number as the next-to-the-last digit. To obtain a random 2 percent, only cases that combine a particular number as the last digit with two other specific numbers as their next-to-the-last digits could

be selected. By the appropriate combination of such procedures one may select any fraction desired. When there are punched card or electronic tape administrative records on all potential research subjects, it usually is a simple task to use data-processing equipment both to select the sample by a last-digits procedure and to print a list of the subjects selected.

Several objections are likely to be raised to the last-digits method of selecting fractional samples. One frequently expressed but almost never realistic objection is that there might be some regular sequence in assignment of numbers resulting in most cases with a particular last digit being different from cases that have other last digits in their registration, Social Security, or other serially assigned administrative number.

A more important objection to the last-digits method of selecting a sample is that the method of selection could be discovered by treatment subjects or staff, thus ending the blindness, and in some situations encouraging manipulation in the designation or recording of numbers in order to get some people into or out of the research sample. To prevent this one could employ a table of random numbers, available in most statistics textbooks or in separate books of statistical tables. After preparing a list of all *potential* research subjects numbered serially, one could simply start at an arbitrarily selected point in the table of random numbers, using only random numbers no larger than the number of cases in the list of potential subjects, and selecting sample subjects by the successive random numbers in the table until the size of sample desired is selected. This list of selected cases would have to be kept secret to preserve blindness. This was, essentially, the method of differentiating the experimental from the control subjects in the Irish experiment.

The Irish example illustrates an additional practical consideration in sampling. Their research sample consisted of all newly received schizophrenic patients, and all were given what the patients and staff thought were vitamin pills. If less than all of the persons of a standard category at a people-changing agency are selected to receive what appears to be a special treatment, not only is blindness eliminated, but both clients and staff may make invidious comparisons between those in the research sample and those not in it. Also, it may be administratively inconvenient to select certain clients for the treatment or the clients may resist the selection process, and therefore, the research instructions on who is to receive the treatment and who is not to receive it may often not be followed. These problems often

justify selecting as a research sample *every* person of the desired type (e.g., schizophrenic) received at certain facilities in a specified period. Facilities and a time period must then be selected sufficient to yield the size of sample desired.

Selection of all cases of a facility or a time period as a research sample creates risks that the time period or the facilities selected will have atypical clientele, thereby limiting the validity of generalizations from the sample to cases elsewhere or at other times. One can check for evidence of such atypicality in age and other measurable traits using appropriate statistical techniques (especially analysis of variance). One can also try to make selections that minimize possibilities of atypicality, such as not selecting highly specialized facilities or holiday periods. Conducting the same kinds of evaluative research in diverse periods and facilities permits statistical analysis of the separate and interacting effects of the multiple factors affecting outcome. This is not impeded if samples are not typical in all proportions to the total population to which generalizations will be made, provided the research sample is large and diverse. Furthermore, of course, if repetition of evaluative studies of a type of people-changing program at different times or settings yields similar conclusions, confidence in the conclusions is increased whether or not the totality of the samples studied is exactly proportional in all attributes to the totality of potential subjects for the program. This has occurred, for example, in assessing the effectiveness of programmed education and token economies in achieving some limited types of goals with many categories of client.

In the Irish experiment, as indicated, randomization was employed to separate experimental from control cases, using a table of random numbers. One could also separate the cases by a mechanical random act for each case—such as tossing a coin, throwing dice, or drawing cards from a shuffled deck—designating in advance which outcomes from this act would indicate the selection, such as all “heads” go into the experimental group and all “tails” into the control group. Actually, use of a table of random numbers is less work than a mechanical process of randomization and is less prone to error (e.g., from an unbalanced or unspun coin, or from a poorly shuffled deck).

The advantage of randomization is that it makes chance alone determine the group into which a subject is placed, so that all traits have the same probability of being in each separate group. If experimental and control groups are selected by judgment that they are similar, there is a prospect that some unintended but

unnoticed difference will distinguish the groups and may account for differences in their subsequent behavior regardless of the treatment which is to be evaluated.

There are two types of limitation on achieving equivalent groups by randomization, small size of the groups selected and constraints on random assignment. When experimental or control groups are small in size, chance alone may make those in one group different from those in another, even with randomization procedures that would make differences negligible for large numbers of cases. There is no specific group size that guarantees perfect matching through randomization; all that one can say from the mathematical laws of probability is that when all groups are randomly selected, the larger the size of the smallest group to be compared with other groups, the lower is the probability that there is any difference between the groups in their proportions of any characteristics, measurable or unmeasurable.

Whenever there is any doubt that the groups will be similar through chance alone, one precaution is to control for measurable differences. If this is done in advance it is known as *stratified* random sampling. Stratification is done by first dividing all the cases from which random selection is to be made by some attribute that is thought to affect outcome of treatment; for example, one could divide them into age or prior criminality groups. This produces, before random selection, a set of separate "strata"—groups selected to assure their relative uniformity in some attribute. Thus the strata may consist of all clients under 21 years of age, all those 21 through 35, and all those 36 and over; alternatively, the strata may consist of all probationers who are first offenders and all who have a prior criminal record, or stratification may be by several variables, such as first offenders under 21, recidivists under 21, and all probationers 22 and over. Random selection of a research sample or random division of a research sample into experimental and control groups is then done separately on each stratum. The stratification before randomization guarantees that all of the randomly selected groups will be identical in their proportions of all the categories into which strata are separated. The random selection after stratification maximizes the probability that the groups have identical proportions of all other variables, measured and unmeasured or unmeasurable.

An alternative method of maximizing equivalence on measurable dimensions deemed important is *edited* random sampling. In this procedure one first randomly divides all cases by a random-

ization method, then checks to see that the groups are similar on measurable attributes believed to affect outcome. If one group has a higher proportion in a category of some attribute, one exchanges cases between groups by purely random selection from the overrepresented component in each group to the underrepresented component in other groups.

A few hypothetical examples may be useful to illustrate edited random sampling. If, despite purely random selection of a research sample of 160 cases from a total prison population of 3,000, one discovers that chance variation has produced the unusual result of a research sample with only 60 blacks from a prison population that is half black, one may randomly select for removal 20 whites from the 100 whites in the research sample and replace them by 20 blacks randomly selected from the prison population. If it is discovered that a randomly selected experimental group has a median age of 22 and the control group has a median age of 24, one may divide each of these two groups into age strata (for example, under 18, 18 through 21, 22 through 25, 26 and over) to determine in which strata one group has a higher proportion than the other group. In this example, the control group's high median age may be due to its having 30 percent in the 26-or-over stratum and only 20 percent in the 18-through-21 stratum, while the experimental group has the reverse proportions—30 percent in the 18-through-21 stratum and 20 percent in the 26-or-over stratum—with both groups having about 25 percent in the under-18 and the 22-through-25 strata. By randomly selecting one-sixth of the 26-or-over cases in the control group for transfer to the experimental group and one-sixth of the 18-through-21 cases in the experimental group for transfer to the control group, one will produce two groups with the same proportions—25 percent—in each age stratum and the same median age. It is, of course, crucial to evaluation that this editing be done before treatment services are provided, and that the random selection and editing be done from the records, without the subjects being aware of the process.

In some research the potential subjects are scattered over a wide geographic area. For example, researchers may wish to contact a sample of all clients or ex-clients of correctional or mental health agencies within a metropolitan area, or they may wish to contact a sample of all probationers in a large State or in the Nation. In such research it is customary to conduct pure, stratified or edited random sampling of geographic units rather than of people, making the random selections from lists of coun-

ties, precincts, or other geographic units. After this *area* or *cluster* sampling is completed, all subjects or samples of the subjects within the selected geographic areas are used in the research. This process saves travel costs and preserves much randomness in selection of subjects. (Further technical details on sampling are available in social science research texts, notably Kish, 1967.)

Stratification or editing with randomization is only feasible if a treatment program to be evaluated is to begin with large groups selected at once from cases already available for assignment to the program. Thus, if admission to a stipend-paying vocational training program is to be offered to half of the 200 applicants on a waiting list for it among unemployed juvenile probationers, or if release with methadone maintenance is to be offered to one-third of the 300 qualifying applicants for it in an institution for civilly committed addicts, these methods of selecting an experimental group to receive the program may be optimal. More typically, as in the Irish experiment, the service is to be offered to a specified fraction of all newly received or newly released clientele of certain agencies, with new cases entering at an irregular rate that averages only a few per week. In such circumstances pure randomization is all that is practical, and one can only check after the experiment is completed to see if the experimental and control groups were similar in measurable attributes, and if so, whether these attributes are associated with the outcome of treatment.

More often than not, political or administrative constraints make an experimental design with any type of randomization absolutely impossible. Thus, if there is a waiting list for a program, it may be deemed ethical or politic to assign to the program only those who have been longest on the waiting list. With new programs it may be most efficient or expedient simply to establish the program first in one arbitrarily chosen location, and to provide it for all those there who qualify for it. Also, presumed experts often persuade officials to institute a new program for everyone in their system at once, or for everyone at a particular location, or for everyone admitted during a specified period.

When evaluation of a program is requested after the methods of assigning cases to it described in the preceding paragraph have already been employed, and are unalterable, one cannot randomly select an experimental group with the program to be evaluated and a control group without it. Instead, the experimental group is replaced by a *treatment* group of those who have already been arbitrarily selected for the program, and the control group is re-

placed by a *comparison* group of persons from a prior time or other location who match those in the treatment group as closely as possible but were not in the type of program that is to be evaluated. This replacement of randomization by some form of presumed matching is the essence of the difference between an experiment and what Campbell and Stanley (1966:34) call a "quasi-experiment."

Matching may be done by a variety of methods, all of which are less satisfactory than randomization under ideal circumstances, but may nevertheless inspire much confidence when several different methods of matching treatment and comparison groups yield similar assessments of a program. One method of matching is simply to compare the posttreatment record of the clients of a people-changing agency before the program to be evaluated was initiated, with the posttreatment records of later clients who were in the program. A followup period of the same duration should be used for everyone in each group. One may also compare the posttreatment behavior of clients in a program in one people-changing agency with that of clients in similar agencies elsewhere which do not have programs of the type that is to be evaluated. One may go further in matching, by comparing only individuals of similar attributes with these groups, such as persons in similar age ranges, with similar offense histories, previously resident in the same neighborhoods, and so forth.

The major deficiency of the above-described methods of matching obviously is that the treatment and comparison groups may differ in attributes or in treatment received in ways other than presence or absence of the program being evaluated. These other differences, such as different economic conditions when each group was treated or differences of the neighborhoods in which they were located, rather than the program being evaluated, may account for differences in outcome between the groups with and without the program.

Sometimes matching is done on a one-to-one basis. Every individual in the treatment group is matched by deliberately selecting from persons eligible for the control group one person who is closest to the individual of the treatment group on some variables selected as the basis for matching, such as age, offense, or ethnic descent. This process has the risk of making the two groups compared resemble each other closely, but making them both different from typical populations elsewhere, thus limiting the extent to which generalizations from the matched groups will apply to populations elsewhere. Also, in matching very closely on some

variables one may mismatch on other variables. Randomization has the advantage, for large groups, of making it highly probable that all groups randomly differentiated are identical in their proportions of every feature that is frequent in any of them.

There are many further ramifications, problems, and solutions in blinding, sampling, randomizing, and matching that have not been covered in this discussion. Some of them will be illuminated by further analysis of alternative research designs.

Alternative Evaluative Research Designs and Their Pitfalls

As indicated, the double-blind controlled experiment illustrated by the Irish research represents an ideal procedure for evaluating people-changing practices, but for a variety of reasons this method of research often cannot be employed in pure form to measure people-changing effectiveness. Therefore, a series of alternative evaluative procedures will be considered here, beginning with the controlled experiment as a standard. Reasons for deviation from this standard will be indicated, and alternative designs or analysis will be suggested for the circumstances where such deviation occurs.

A. *The Classic Controlled Experiment*

This design was illustrated in its simplest form by our Irish example. It is represented with the following symbols (adapted from Campbell and Stanley, 1966:25):

R	Experimentals:	X	O ₁
R	Controls:		O ₂

The first line represents the experimental group receiving the treatment "X" to be tested (in the Irish case, nicotinamide). The second line represents the control group receiving no treatment (in the Irish example, those who received the placebo). "R" indicates that all cases are first divided randomly into the experimental and control groups, to maximize the probability that the cases in each group have about the same proportion of every personal attribute. This is to minimize the probability that the selection of better risk cases for one group than for another accounts for differences in their outcome. "O₁" and "O₂" stand for the posttreatment observations of the experimental and control groups, respectively (these yield data such as the statistics presented in Table 6.1).

Evaluational research in people-changing agencies could follow

the classic design above much more often than usually is realized, but often such a design is not sought by researchers, is not permitted by authorities, or is not feasible even if requested and authorized. Some political and administrative reasons for this resistance or difficulty in controlled experimentation (many of which also occurred in the history of organic medical science) have already been indicated. Therefore, alternative research designs that may be substituted for the classic type of inquiry should be considered.

B. The Prescreened Controlled Experiment

One source of resistance to controlled experimentation is simply that the treatment to be tested, if more "lenient" than traditional practice, appears to endanger the public or to conflict with government goals other than changing those adjudged deviant (for example, with the goal of deterring people who contemplate committing deviant acts).

This type of resistance was evident when the California Youth Authority's Community Treatment Program proposed immediate parole of committed youth offenders for intensive assistance in the community, youth who would otherwise have been confined for an average of 8 months. It was feared that the public often would object to such release, that victims of crimes by these youths and law enforcement personnel who had just arrested them and assisted in their prosecution might be angered to encounter them on the street soon afterward. This prospect of a negative public reaction was alleviated by having the Youth Authority screen all offenders committed to their custody soon after they arrived at the Reception Center serving the area where the research was to be conducted. In this screening some youths, notably sex offenders and others who had committed notorious crimes, were declared ineligible for immediate parole because of anticipated community resistance to their release at that time. In practice this generally meant that about a quarter of the males and a tenth of the females were eliminated from the pool of eligibles from whom the experimental and control groups were randomly selected (Warren, 1966).

The Prescreened Controlled Experiment design may be represented by the following symbols:

S ₁	Ineligibles:	O ₁			(O ₃)
S ₂	Eligibles:	O ₂ :	R	Experimentals:	X O ₄
			R	Controls:	O ₅

In the above, S_1 represents the cases screened out as Ineligible for the experiment and S_2 represents the Eligibles, for whom the experiment is permitted. O_1 and O_2 represent initial observations of the eligible and ineligible groups to see what attributes differentiate them. This is extremely important, since generalizations from the experiment will only be applicable to cases resembling those in S_2 , the eligibles, who are randomized into an experimental and a control group as in the classic controlled experiment design. O_3 is optional, a later observation of the screened out Ineligible cases to see how their outcome compares with those of the experimental and control cases. The criterion observations of the latter groups are indicated by O_4 and O_5 .

There are numerous reasons for prescreening, and diverse ways of doing it. Perhaps the most common screening method is to provide a type of treatment only for those who apply for it, or only for those who express both a desire for it and are willing and able to pay for it. Restriction of people-changing efforts to such cases may be made for economic reasons, on ethical grounds, because of legal requirements, or because the treatment itself—such as psychotherapy—is believed by its proponents to be effective only for those who volunteer or only for those who pay for it. Conclusions from studying such selected persons may not apply to other persons.

A different type of screening occurs when a treatment is deemed dangerous for the client, rather than for society. This is familiar to the public in heart transplant and other therapies of organic medicine. In efforts to change behavior it is frequent in the prescription of drugs to combat addiction. When the drugs, such as methadone, are themselves addictive, or when there is a risk of organic damage or discomfort from side effects, these therapies are restricted not only to those who volunteer for them, but also to those who clearly have been heavily addicted for a long period. Frequently these drugs are provided only to adults who have been unsuccessfully treated by less dangerous therapeutic programs. (For information about a more complex prescreening design combining purposeful and random allocation, see Wilkins, 1969: 152-155).

Whatever the reason and procedure for prescreening before a controlled experiment is conducted, it is crucial that differences between the cases screened out and those judged eligible for the experimental and control groups be identified as well as possible. Minimally, the proportion of prospective clients deemed ineligible, or the proportion not volunteering and the methods by which

volunteers were recruited, should be indicated. These types of information are necessary to qualify generalizations from the findings of the prescreened controlled experiment, to caution that results apply only to persons resembling those in the pool of eligibles from which the experimental and control cases were selected.

Usually a new treatment for a serious problem, if well publicized as highly promising, will be requested for or by many more persons than can initially be provided with it. Under such circumstances officials customarily propose some type of screening, such as supplying the treatment services only to those who have been most difficult to change by other methods, or to those for whom the program seems to involve the fewest risks. Whenever any screening is necessary because the attractiveness of a program creates a demand that it be supplied to more than the number for whom it can be made available, the quickest and most convincing argument for increasing the availability of the program, if it is effective, is to screen out cases by randomly separating potential clients into an experimental and a control group. Should prescreening rather than a classic design be pursued, it is appropriate to select for the pool of eligibles only the types of clients for whom the program is expected to yield the greatest cost-benefit profit or efficiency.

C. Contaminated Randomization and Pseudoeperiments

Even when classic or prescreened controlled experiments are initiated, a number of impediments not yet discussed prevent completion as planned. This occurs because most people-changing research does not consist of evaluating one pill versus another, where one is a placebo, as in the Irish experiment cited; therefore, the identity of experimental and control group members usually cannot be hidden. The treatment variables being investigated may include separately, or in various combinations (with interaction among them also studied) such highly visible people-changing measures as: vocational training, counseling, education, medication, segregation, special types of staffing, changes in staff-client relationships, variations in client responsibility, monetary motivation, job placement, confinement, or dispersion. Ideally the experimental and control group should be different in treatment received only on that item or combination of items on which effectiveness is to be measured, but all of these treatment methods are visible to the experimental group members who receive them, and frequently persons in the control group or their

friends or relatives become aware of the services denied them.

Because of the awareness of treatment methods by the subjects of people-changing efforts, the administrators, the friends or relatives of subjects, the victims of subjects, and the general public, there is a tremendous diversity of interference with controlled experiments. Most frequently, in my experience, staff somewhere in an agency or the clientele or their friends or relatives regard the experimental services as highly desirable, so they try to provide them for members of the control group whom they regard as especially deserving, or in whom they have some personal interest. To expand the availability of the experimental services, they may also arrange to transfer out of the experimental group clients whom they think will succeed without the special services, or who have a personal interest in being in another program, perhaps in a different geographical location. If the persons transferred out of the research-designed program assignments are randomly selected, a few transfers will not seriously harm evaluation; if the interference is not randomly distributed and if it affects a large proportion of the research subjects, total comparisons of experimental and control group data may be meaningless.

Another frequent but indirect source of contamination in randomized assignment of services to experimental and control groups is failure to predict accurately the flow of eligible cases to a research project. For example, in 1969 a controlled experiment with methadone maintenance for civilly committed heroin addicts in New York required that a physician screen each addict as nonpsychotic, nonassaultive, and an opiate user for at least 4 years. When a psychiatrist accustomed to middle-class clients was hired for this screening task, he viewed nearly all cases as not fitting this criterion. In the early 1970's, after the Federal Bureau of Prisons initiated a carefully planned experimental program for juvenile and youthful offenders at its Robert F. Kennedy Youth Center in West Virginia, the Federal court judges drastically reduced the imposition of confinement sentences in Federal institutions for young violators of Federal laws. When such developments sharply curtailed intake at the experimental research centers, there were economic pressures to use the staff and facilities there for other subjects of diverse types. Both the experimental and the control group cases, and nonrandomly selected additional subjects, were then hastily assigned to the programs first intended only for scientifically selected experi-

mental subjects. This process has occurred repeatedly in many correctional, mental health, and addiction treatment programs.

The dropouts from a voluntary treatment program often are inappropriately viewed as a contamination of the randomization process. Of course, dropouts in controlled experiments distress sponsors, administrators, and researchers, but it may be a realistic feature of a particular treatment method that it results in many dropouts with certain types of clientele. This may be an important fact to know, but a fact for which dimensions and consequences are precisely learned only by research, preferably with controlled experiments. In Los Angeles, for example, an inventory of fatherless boys was made in several black ghetto and Chicano barrio schools. The households with such boys were randomly divided, with "Big Brothers" offered to boys in the experimental half. In a number of cases the mother or the boy, or both, declined the relationship with a "Big Brother," so that the youths maintaining such a relationship were no longer comparable to all boys in the control group. Nevertheless, the dropouts from a voluntary experimental group are only a contamination of randomization if they are ignored. (See discussion in Chapter 10 on the bias from ignoring dropouts of inhouse trials.)

A program to change people should be evaluated on the basis of all those whom it undertook to change. Therefore, those who do not remain in the program must be considered with those who remain; together they comprise the totality that was to be changed (Empey and Erickson, 1972:74-77). In the Big Brother project, for example, if no contamination other than that described occurs, a comparison of subsequent arrests, school performances, and other criteria for the original experimental and control groups will indicate the overall effectiveness of offering Big Brother services to fatherless boys in these neighborhoods. If those families or boys in the experimental group who declined Big Brother services are appreciably different in certain attributes (e.g., age, or number and attributes of siblings) from those who accepted these services, and if these attributes are identifiable in the control group, then comparisons of outcome can be made separately for these categories of youth or family that have the highest rates of acceptance of Big Brother services. This type of analysis might suggest an appropriate prescreening policy for the program that would perhaps increase its cost-benefit profit by concentrating its services where they will be most effective.

The contamination of randomization in a classic controlled experiment can be represented as:

R Experimentals:	T	X	T	O ₁
R Controls:	T			O ₂

Here the T represents transfers out of the experimental and control groups, respectively, after randomization has occurred, and either before or after the experimental action. A transfer may be formal, where some nonrandomly selected members of either group are officially assigned to programs other than the ones to which they were randomly assigned for research, or it may be an informal and unofficial de facto transfer. The latter occurs whenever the services to be evaluated are either denied to nonrandomly selected members of the experimental group or are provided to members of the control group. It is the unofficial type of transfer that is most difficult for a researcher to prevent, or even to be aware of. If it is extensive and the researcher is unaware of it, of course, the investigation is clearly a pseudo-experiment, but this is often known, if at all, only after a large investment in research has been made.

In a prescreened controlled experiment the contamination that concerns us here is only that which occurs after the eligible cases are randomly divided into an experimental and a control group, although transfers also occur from the eligible and ineligible groups. This can be represented as follows:

S ₁ Ineligibles:	O ₁ ,	T ₁					(O ₃)
S ₂ Eligibles:	O ₂	T ₂	R Experimental:	T ₃	X	T ₅	O ₄
			R Control:	T ₄			O ₅

Here T₁ and T₂ may only change the definition of eligibility, but these changes must be described as well as possible so that the final description of the eligible cases, those that are randomly divided into an experimental and a control group, indicates the type of client to whom the results of the experiment appear to be applicable. Indeed, any reduction in the proportion of potential research cases declared ineligible increases confidence in the general importance of the experiment, so transfers from the ineligible to the eligible groups before randomization occurs are often especially desirable. Even when transfers from ineligible to eli-

gible status are made on the basis of criteria that are not clear, as when they reflect the unfathomable hunches of screening officials, this should not greatly disconcert researchers if it greatly reduces the proportion who are ineligible and it is done before randomization occurs.

Transfers of cases frequently consist of shifting individuals from the ineligible group into the group receiving experimental services. This may not impede the prescreened controlled experiment if the researchers are aware of it and simply do not tabulate data from these individuals with the data from the experimental group which is compared with results from the control group. It is primarily transfers made from the randomly separated experimental and control groups, either formally or informally, that seriously impair experiments.

As previously stressed, random assignment is the most certain way of maximizing the probability that an experimental and a control group are similar in all experiences and attributes other than the people-changing effort that is to be evaluated by providing it only for the experimental group. While contamination of the randomization may damage the experiment, the destruction may not be complete. There are several ways of salvaging order out of chaos in random assignment.

If the attributes of the cases transferred out of the experimental or the control group are known, and an appreciable correlation of these attributes with outcome rates has been established, one can assess the direction of the bias resulting from the transfers. For example, if the experimental group received or the control group lost more high than low failure risk cases, one can assert that this contamination biased the research against showing favorable results for the experimental program. If a followup of the two groups nevertheless reveals that the experimental group is more successful than the control group, one can conclude that the experimental measures are clearly beneficial, for the contamination's effect is to understate the actual benefits. This occurred, for example, in New Haven's Residential Youth Centers, described by Goldenring (1971:407-416), where the control group averaged better work attendance, more earnings, and fewer arrests than the experimental group of out-of-school and out-of-work youth, but the reverse was true 6 months after the experimental treatment began.

A second and simpler salvage is possible if the lost cases are few, or if no known attributes of cases transferred out of the original experimental or control groups have a marked correla-

tion with outcome. One may then estimate their potential impact on the experiment's findings by appropriate assumptions. The most conservative assumption is that all the cases lost from the control group would have been great successes had they remained in it and that all of the cases lost from the experimental group would have been failures had they remained in it. A less conservative assumption is that their actual outcome would have been the average for cases with their attributes; the assumption made will not be crucial if the number of cases lost is small. In any choice, one can calculate what the difference in outcome rates for the total experimental and control groups would have been with the assumption chosen. If the experimental program would have been clearly beneficial even if the most conservative assumption were valid, then the contamination of randomization has not been sufficient to destroy the value of the planned experiment.

The final recourse in coping with contamination of randomization in a controlled experiment, to be employed if all other ways fail, is to regard the project as clearly a pseudoexperiment, but nevertheless as yielding potentially valuable information on the experimentally treated cases. This implies completely abandoning the idea of regarding the experimental and control groups as equivalent in all respects other than the treatment which was to be evaluated, but comparing the criterion information for all or part of the experimental group with the same type of information on other groups that can reasonably be inferred to have been similar to the experimental group before the treatment. This brings us to the last recourse of comparison for evaluative purposes, which is often the first and only recourse if no controlled experiment was ever intended. It is also an important type of supplemental analysis for data from a controlled experimental group.

D. The Nonequivalent Comparison Group Quasi-Experiment

No knowledge on the effectiveness of a people-changing effort is acquired only by learning the subsequent rates of behavior of those subjected to the effort. Instead, effectiveness is assessed by comparing these rates with some standard, preferably the rates that evidence suggests would have characterized the group studied had they not been the subjects of the people-changing endeavor. The most rigorous method of procuring such a standard is the random division of prospective cases into an experimental and a control group. When such a controlled experimental design is not even attempted, or is seriously contaminated after it is

attempted, all or part of the treated cases for whom criterion information is available may be employed to evaluate the treatment by comparing this information with what is inferred to be the most appropriate standard available.

As has already been indicated, much evaluative research in people-changing that does not meet the randomization standards expected in a controlled experiment has appropriately been labeled by Campbell and Stanley (1966:34) as "quasi-experiment." If such research does not employ a randomly selected control group, it uses some other type of *comparison group* for the appraisal of its outcome data on the *treatment group*. The comparison group may consist of the research subjects themselves, at an earlier time, or it may be another group believed to be similar to the treatment group in all respects except the treatment experience that is to be evaluated. This other group could simply be the subjects of another study, perhaps in another geographic area, or it could consist of clientele of the agency in which the evaluative research is conducted, but from a period before the treatment being evaluated was introduced. What all such comparisons attempt to do is to simulate a controlled experiment by seeking the closest thing to a control group that is available, with which to compare the treatment group.

A comparison group that was not randomly selected with the treatment group may not be the latter's equivalent in all particulars other than the treatment being evaluated, but a randomly selected control group may also not be fully equivalent to the experimental group. As previously indicated, some nonequivalence occurs by chance alone even with random selection, especially in small-sample studies. This can be guarded against on variables of known importance by stratified or edited random sampling. Secondly, as already discussed, both official and unofficial contamination may develop to make the experimental and the control groups no longer equivalent through random selection, or no longer sharply different in having had or not had the treatment that is being evaluated. Thirdly, somewhat diverse but related unplanned consequences of research operations, commonly labelled "placebo effects" and "the Hawthorne effect," may create important differences between the experimental and control groups that usually are not anticipated in research designs.

The term "placebo effect" is derived from the frequent medical observation in some ailments, that patients given completely inert medication will recover more rapidly or, at least, will suffer less than patients who receive no medication at all. In many institu-

tional situations, and sometimes in the community as well, the subjects in a controlled experiment and the staff know which administrative unit is the experimental and which is the control group. This knowledge may constitute a challenge, motivating one or another or both groups to perform better than they otherwise would.

The term "Hawthorne effect" derives from a classic study at the Hawthorne plant of the Western Electric Company in Chicago that began with an assessment of the effect of lighting on productivity. To their surprise, the researchers found that neither increases nor reductions of illumination affected production; instead, all groups, including the control groups that had no change in lighting, produced more when they knew that their production was being measured by the researchers (Roethlisberger and Dickson, 1969:14-18). In many people-changing agencies the attention given an experimental group, and the anticipation that it will perform better, may result in its performing better regardless of the actual effect of the experimental services or conditions provided for it.

Because of these possible consequences of research procedures or settings, confidence in the conclusions from initial studies is greatly enhanced when experiments are repeated in a variety of contexts and their findings are tested with diverse comparison groups. Secondly, results will be considered more certain if, when possible and ethically justifiable, subjects and staff are not informed that they are in an experimental study or, minimally, do not know whether they are in the experimental or the control group.

A third implication of placebo and Hawthorne effects is that, whenever possible, information on the attributes and outcomes of eligibles and ineligibles, experimentals and controls, or research and comparison groups, should be obtained by what Webb and coworkers (1966) call "unobtrusive measures." This means that pre- and posttreatment observations should be made by extracting data from administrative records routinely maintained for purposes other than evaluative research, rather than by personal contact with the research subjects through interviews, by direct observation or by interviewing employers, neighbors, or others in direct contact with the subjects. The latter point, of avoiding any contribution to the public labeling of the research subjects, is especially important for both research validity and ethical reasons.

There are numerous other potential sources of systematic error

(or "bias") which should be minimized in selecting research and comparison groups for evaluation of people-changing efforts. One type might be called "history effects" (this, and the other "effects" discussed in the remainder of this chapter, are given labels derived from Campbell and Stanley, 1966:5). These effects are the consequences of events independent of the procedural variable being studied, events that may uniquely affect outcome data on research or comparison subjects during a particular period. For example, Adams and coworkers (1968) found that releasees from District of Columbia correctional facilities had recidivism rates between January and July 1968 considerably higher than those in any of four previous 6-month periods. This was ascribed to widespread social disorder among the predominantly black residents of our Nation's capital following the assassination of Martin Luther King that spring, and may possibly also have been due to increased police and court initiative in arresting and convicting during this period. Clearly, if recidivism rates of a research and a comparison group were found to contrast, but one group was released during this high recidivism rate period and one was not, the results could create a quite erroneous assessment of the treatment.

History effects on evaluative research frequently result from fluctuations in economic conditions, and from changes in policies or personnel that determine the criteria for defining clients as successes or failures. Obviously, such events can greatly affect both comparisons of groups from two different periods, and comparisons of the behavior of a single group of subjects before and after a treatment experience. One remedy is to avoid, if possible, comparisons from periods clearly different in relevant respects. Another is to make multiple comparisons, from different periods. Consistency in findings of *relative* effectiveness of different treatment methods in various periods, is sought; no method need be constant in its effectiveness. Instead, it should be constant in being more effective than another in each period studied, to inspire confidence that our conclusions on relative effectiveness will be valid in the future. On the other hand, extending studies to different periods may reveal an interaction between history and treatment effectiveness. For example, vocational counseling or training may alter outcome rates in times of full employment but not in times of great unemployment, or vice versa.

A source of error in comparisons that is particularly distorting if conclusions are made by comparing the behavior of subjects

before and after a treatment experience, or by comparing subjects of different age levels, can be called "maturation effects." This refers to changes in the subjects that occur over their lifetimes, regardless of whether they are involved in people-changing programs. Thus it has repeatedly been shown that recidivism rates of criminals decrease with age (Sellin, 1958; Glaser and O'Leary, 1966), and that people "mature out" of opiate addiction (Winick, 1962; Snow, 1973). From this it follows that research subjects generally should diminish their rates of these types of deviant behavior whether or not they are subjected to special efforts to diminish these rates. Instead of studying posttreatment rates of the research subjects in comparison with their pretreatment rates, it is therefore preferable to study the change in rates for a treated and an untreated group of the same age over a span of years encompassing the treatment.

Another possible impediment to measuring the separate impact of a people-changing program is what can be called a "testing effect." This may occur if measurements are made before treatment, as when a personality test is used before and after a counseling or psychotherapy program to measure the program's impact, or is used to stratify or edit random sampling. The testing itself can tell the subjects the concerns of the tester, and may assist and motivate them to score more favorably when they are retested. This was previously discussed, and examples were cited, in Chapter 3, when discussing the most objective criterion.

Testing effects can be eliminated, of course, by relying only on unobtrusive data for pretreatment information. The impact of these effects when obtrusive measures are employed can be assessed by using what Campbell and Stanley (1966:24-25) designate as the "Solomon Four Group Design." This can be represented as:

R	Pretested Experimentals:	O_1	X	O_3
R	Pretested Controls:	O_2		O_4
R	Untested Experimentals:		X	O_5
R	Untested Controls:			O_6

Here random selection is employed to form four equivalent groups, only two of which are pretested, as indicated by O_1 and O_2 . If the difference between O_5 and O_6 (the observations of outcome for the untested experimental and control groups) is greater than the difference between O_3 and O_1 (the observations of out-

come for the pretested experimental and control groups), then the diminution of difference can be ascribed to the pretesting.

When pretesting occurs, a type of "matching" between the research cases and cases to form a comparison group frequently is undertaken which can create distortion by what are known in statistics as "regression effects." This phenomenon was pointed out long ago by the anthropometrist Sir Francis Galton, who noted that children of exceptionally tall parents were likely to be shorter than their parents and children of exceptionally short parents were likely to be taller than their parents, since all extreme variations tend to regress toward the average of the group to which they belong. Therefore, if one matches low-scoring individuals receiving a special instructional program with low-scoring individuals not receiving it, one will find that both groups tend to increase in score on retest simply because the low scorers will include a disproportionate number of people who were performing at below their average level in pretesting.

Distortion from regression effect will especially impair evaluation if a low-scoring group is used as its own comparison group, by evaluating instruction in terms of the difference between its own pre- and postinstruction scores. Of course, the extent of distortion from this kind of regression effect is less when the reliability of the tests used is high. Similarly, if one selects for a therapy program only those who score high on a test for defects of personality, part of any decrease in score on retesting will be due to regression effects, part will be due to testing effects, and presumably part will be due to the therapy. Regression effects can also occur when samples for research or comparison, or both, groups are selected by extreme scores on unobtrusive measures that are correlated with the behavior that the people-changing effort is trying to affect, since the unobtrusive measures (such as earnings, grades, number of arrests, days unemployed, etc.) also reflect some fluctuation due to chance, on which extreme variations tend to regress toward a mean. This type of regression effect is often subtle and difficult to prevent completely. Indeed, regression effects as a source of error can occur with any matching. Their magnitude can sometimes be estimated by taking pre-treatment and posttreatment measurements in different score ranges and noting the change in the extent of score shifts at different ranges (cf., Campbell, 1969:419-425).

E. Pre-Post Comparisons

Whenever clients' records are available for a considerable time before and after a treatment that is to be evaluated, it will

probably be interesting and useful to compare their pretreatment with their posttreatment behavior or experience. For example, number of arrests, time confined, or months employed during 5 years following release from confinement can be compared with number of arrests, time confined, or months employed during 5 years before their confinement. This type of analysis is often called a "pre-post comparison."

If possible, one should also gather before-and-after observations for the same periods of persons in a control or comparison group. The objective in before-and-after observations of a treated group is to determine whether their behavior changed, following the treatment, from what it had been before. The objective in making the analogous observations on untreated persons, similar to those treated, is to infer whether the treated persons would have changed to the same extent even without their treatment (whether this change is due to history, regression, or maturation effects).

Pre-post comparison is illustrated in simple form by the finding that 46 percent of the clients of a St. Louis alcoholic detoxification center were arrested in the 3 months before their first admission to the center, while only 13 percent of these clients were arrested in the 3 months following their discharge from the center (Law Enforcement Assistance Administration, 1970:xiv). Since their median age was 48 years, it seems unlikely that this change was due to their aging in this brief period. Nevertheless, it certainly would be desirable to make such a pre-post comparison also for similar alcoholics receiving the alternative to a stay in the detoxification center, a commitment to the county jail. And it would be most desirable to make the pre-post comparison for a longer period than 3 months.

Another type of pre-post study is represented by Frances Gearing's comparison of the percentage of time that addicts were employed or in school before and after their entrance to a methadone maintenance program. These comparisons were made separately for every pre-post period which is a multiple of 3 months, from the short term of only 3 months to a maximum of 42 months. The percentage of time employed or in school was greater following than before the treatment for each of these durations. This difference reached a peak in the 33- and 36-month comparisons, in both of which she found nearly three times as large a percentage of time employed after entrance on methadone maintenance as in the same period before.

One defect of the comparisons above was that data on employ-

ment before entrance on the program were often based mainly on the addict's recall (only the involuntary unemployment due to incarceration could be verified) while the employment data for the period in the maintenance program were derived from current reports, frequently verified by the supervision staff. Therefore, the posttreatment employment data may be more complete than the pretreatment data. Secondly, there were larger samples on the short-term pre-post comparison than on the long-term comparisons, since the long-termers were a more select group who had lasted that long on the program. Also, the long-termers were included in the short-term pre-post comparisons (based on data from when they were in the program for a shorter term). Thirdly, of course, it would be interesting to compare this pre-post change record with that of addicts in other programs, including penal incarceration, civil commitment and therapeutic communities (Gearing, 1969).

In the Provo Project, pre-post comparisons were conducted of two pairs of groups. All members of one pair were boys deemed by the county judge to merit probation by his usual criteria, but were randomly divided before being informed that they were to receive probation, with half receiving regular probation conditions and half being required to participate in an experimental treatment program. This program required daily work when not in school, gave probationer groups much responsibility for positive and negative sanctions on their members, and promoted a distinctive style of staff-client relationship. The second pair consisted of: (a) Boys deemed by the judge to merit commitment to the State industrial school, by his usual criteria, but placed instead on probation in the experimental program described above. They were informed that this placement was in lieu of incarceration. (These persons were at first randomly divided, with half actually being committed to the State school, but after a brief period such division was terminated for newly adjudged boys, and all recorded by the judge as appropriate for incarceration were placed in the special experimental program.) (b) Boys sent to the State industrial school from the rest of the State (plus the few from the research county who had been randomly selected for incarceration) were then matched and selected for purposes of comparison with the experimental group (Emery and Erickson, 1972: Chapter 2).

For each of the groups described in the preceding paragraph, the number of arrests experienced during 1 year before the arrest that led to their entering the project sample were com-

pared with the number of arrests during 1 year after their release to begin their sentence (for those receiving probation or probation with the special program conditions) or after their confinement (for those receiving incarceration). Comparisons were also made for 2-, 3-, and 4-year pre-post periods.

In every one of the above-described comparisons, for each of the four groups and for each of the four durations, there were fewer arrests in the period after release than in the same length of time before the arrest that brought them to the project. For the two randomly differentiated probation groups, the reduction of arrests after the project was greater in the experimental than in the control group for each duration of followup, but only in the 1-year before-and-after comparisons was this difference between the experimental and control groups considered statistically significant (at the .05 probability level). By contrast, for the comparison of incarceration cases, the post-release reduction in arrests was greater for those sent to the State industrial school than for those placed in the special probation program in the 1-year pre-post comparison period, but this pattern was reversed in the 2-, 3-, and 4-year pre-post comparisons. Indeed, for these incarceration cases, the longer the time of the before-and-after comparison after 1 year, the greater was the reduction of arrests for the experimental community-treated cases as compared with those sent to the State industrial school. They were only significantly greater (at the .05 probability level) at 4 years, which was the longest duration of pre-post comparison undertaken in this project. Incidentally, all of these contrasts were greater when arrests for moderate and serious offenses were considered alone than they were when mild offense arrests were also considered (Empey and Erickson, 1972:Chapter 10).

More complex multivariate analyses of these pre-post data can be made statistically, to try to determine what percentage of the posttreatment reduction in deviant conduct is due to maturational effects or other variables and what percentage is due to the intervention or is inexplicable. Such analyses often use dummy variable multiple regression techniques, but the assumptions necessary with these techniques may make the conclusions highly speculative (Empey and Erickson, 1972: Chapter 11). At any rate, the foregoing examples of pre-post comparisons should confirm the opening assertion of this section that such comparisons will probably yield interesting and useful evaluative data.

Conclusion

This chapter has stressed that evaluation of people-changing methods requires *comparison*, in conjunction with a followup period. The concern here has been with procedures and problems in answering the question: Who should be compared to whom? Techniques of blinding, sampling, randomization, and matching were described as well as several types of research design, and many pitfalls in using them were indicated. Numerous additional designs and pitfalls are discussed in works on research methodology (e.g., Campbell and Stanley, 1966), but this chapter has focused on those which the author has encountered most often during more than two decades of involvement in evaluative research in correctional and other people-changing agencies.

Emphasis on pitfalls may discourage administrators from conducting evaluative research, since in their dealing with politicians or with financial contributors they wish to inspire confidence by expressing great certainty as to their effectiveness. As problems such as crime, alcoholism, drug addiction, and mental illness persist, and as the educational level of the public rises, however, there is increasing demand for scientific evidence on program effectiveness, and there is growing recognition that no panaceas exist for these problems (except for those problems that one is willing to solve by ceasing to define them as problems). For most people-changing objectives, as Campbell and Stanley (1966:3) assert:

. . . we must increase our time perspective . . . even though we recognize experimentation as the basic language of proof, as the only decision court for disagreement between rival theories, we should not expect that "crucial experiments" which pit opposing theories will be likely to have clear-cut outcomes.

Adams (1974) points out that not just in corrections, but also in research and development in industry, less than 5 percent of the studies undertaken demonstrate a means of increasing effectiveness. As was stressed in our preceding chapter, research is also valuable when it indicates that measures advocated as effective are ineffective. This was illustrated by our Irish example, at the beginning of this chapter. The utility of both positive and negative findings will be enhanced if their pursuit is institutionalized and their findings are made cumulative. Further requirements for achieving these goals will be the concern of our next chapters.

CHAPTER 7

CROSSING THE TIME BARRIER: OBTAINING OUTCOME DATA

It has been repeatedly stressed here that the evaluation of a people-changing effort requires information on the behavior of the clients for some period of time after the effort to change them occurred. As indicated in the preceding chapter, the key principle of evaluation is to follow up those persons whom an agency tried to change to see if they do change, and to follow up a control or comparison group to see whether they also change even if nothing is done to them or if an alternative kind of treatment is given them.

Unfortunately, the time perspective essential for evaluation is infrequent in the thinking of administrators. They are currently required to think only of budget justifications, which do not reflect effectiveness. They report such matters as the number of clients their agency served in the past year, how many it is handling now, and how many they expect it to have in the future. Frequently these are the only valid statistical data they submit to justify their budget, rather than statistics on their effectiveness.

A Cohort Approach

To consider the effectiveness of a people-changing agency scientifically, one should conceive of its clients as a set of *cohorts*. A cohort is well defined by Ryder (1965) as "an aggregate of individuals . . . who experienced the same event within the same time interval." The aggregates that concern us consist of those persons in specific kinds of people-changing programs in particular periods, since the events we wish to evaluate are the varieties of treatments they received. With a cohort perspective one asks such questions as: What happened to the clients of 5 years ago? What proportion are now in prison, and what proportion are in mental hospitals? How many are "doing well"

in the community? What proportion of the past 5 years did each spend in circumstances other than those they now are in? What types of persons were helped most by what programs? How does the record of those released 5 years ago compare with the record of those released 2 years ago, or with the record of those who got out 10 years ago?

Established businesses customarily evaluate their activities over the years in terms of income, expense, and profit trends, both for the business as a whole and for its component departments and products or services. Thus, sales of an item are described as 5 percent up from last year, costs as up also, and perhaps profits as down. The administrators of people-changing agencies also think of statistical trends regarding the number of clients they handle, the number of personnel they employ, and other items, but they are not accustomed to thinking of trends in profit or other expressions of effectiveness.

Assessing the effectiveness of a people-changing agency, and especially assessing trends in effectiveness, requires thinking in terms of cohorts. Clients of 10 years ago, for example, may be assessed as of 5 years afterwards, and the findings may be compared with the current assessment of similar clients of 5 years ago. Each such group thus compared (e.g., those of 10 years ago and those of 5 years ago), constitutes a cohort, an aggregate of persons who experienced the services of an agency during a given interval of time (e.g., during calendar 1963 or calendar 1968), evaluated as to their behavior through a particular followup period (e.g., 5 years).

Comparisons of outcome may be made separately for the annual cohorts from each past year, or cohorts can cover other periods of time (e.g., persons released during a 6-month interval, or a 2-year interval). Different followup periods may also be used, such as 2 years or 4 years, but in any cohort for which one presents evaluative statistics, each case should have the same duration of followup. Variations of followup period were illustrated in Chapter 4.

Ultimately, with the institutionalization of evaluative research in people-changing agencies, officials should be able to present 2-year or 5-year or other specific period outcome rates for each treatment measure. These rates will ideally be expressed as cost-benefit profits. Officials should be able to present such rates for various types of clients, in diverse types of treatment programs, and for different periods. How the outcome data are ob-

tained may be the most crucial problem, however, and that problem is our concern here.

Procurement of Followup Information

One of the major deterrents to evaluation research in people-changing agencies has been apprehension that it would be impossible to procure adequate followup information. The difficulty such procurement entails varies greatly with the kind of cohort that is followed up, the duration of the followup period, and the outcome criterion that is employed. It is a certainty that any such data collected will be imperfect. It is also almost certain, however, that any systematically collected and tabulated outcome data will permit more adequate evaluation of a program than the usual assessment from unsystematically collected and incompletely tabulated subjective impressions. Furthermore, the difficulty of getting useful followup information tends to be greatly exaggerated.

Ease of followup varies directly, of course, with the extent to which the subjects studied remain under the control of the agency engaged in the followup effort. It may be useful to consider the problems of procuring followup information under two contrasting degrees of control, predominant or complete control on the one hand, and on the other, when all or most subjects are released without any restriction or control.

A. The Simplest Circumstances: All Subjects are Under Supervision of the Organization Conducting the Evaluation

When all members of the cohorts to be followed up are under the control of the evaluating organization, or of an organization cooperating in the evaluation throughout the followup period, outcome information should be most readily available. In practice, agency files sometimes make this information unavailable, but presumably the agency has somewhere basic data on every client under its control.

Postrelease information adequate for much evaluation is collected automatically in agency files, for administrative rather than research purposes, whenever all members of a cohort are released under supervision in the community, on probation, parole, or aftercare. In most cases, the information is only adequate if the term of supervision imposed is as long or longer than the followup period desired for evaluation purposes. The supervision agency receives, minimally, information on whether its releasees (probationers, parolees, and others) are still in contact

with the supervision staff, have been arrested, have been confined following arrest, or have disappeared from supervision. If arrested or reconfined, the reasons for this action also are learned by the supervision agency. Maximally, it may receive much more information, such as data on postrelease work record, earnings, residence, family relationships, associates, apparent attitudes, and so forth.

Such parole performance information would be adequate, for example, for a cost-benefit analysis of community correctional centers or work release by comparing the subsequent performances of cohorts of parolees placed in these programs before release with the performance of cohorts of similar prisoners paroled directly from regular institutional custody. Aftercare records would suffice for cost-benefit comparison of release to methadone therapy with release to regular aftercare supervision.

Surprisingly, the postrelease information needed for such evaluations is often difficult to collate, even when it is all contained within the files of a single organization. Frequently a releasee returned under a new sentence or other commitment order receives the next registry number even if the newly received person was committed in the same institution previously. While the file opened under the new number will contain information on the subject's previous commitment, the file under the number of his previous commitment is closed, much of it may be destroyed or buried in dead record archives, and a cross-index for tracing cases from their old file numbers to their new ones is not maintained. I was told in one of the largest county correctional agencies that it would take at least a day of clerical labor per case to trace in the agency's files what happened to juveniles it placed on probation in an experimental program 5 years earlier. This is an agency, incidentally, with extremely elaborate computerized records, but its information system is designed to chart the volume of transactions and their conformity to arbitrary standards of workload, rather than to evaluate their effectiveness in recidivism reduction or cost-benefit.

Ideally, for evaluative followup purposes, a correctional or other treatment organization should employ only one number for each client on all of his or her commitments. This is conveniently a number useful for tracing the individual through other agencies, such as the Social Security number, or the Federal Bureau of Investigation number for adult felons. Alternatively, if the number is changed with each new commitment, index files relating all numbers and names to each other may be

maintained, to indicate where subsequent information is located on individuals returned under new registration numbers.

It should be stressed again that any body of followup information on a cohort of cases will be imperfect. There will be some pure errors in records, possibly arising from use of aliases or multiple Social Security numbers by some individuals. Also, even if records are complete for new arrests, convictions, mental commitments, or other information on official reactions to the subject's deviant behavior, these official reactions are imperfect indices of the rates of deviant behavior. Measuring an individual's deviance by the official record of his or her deviance can certainly be misleading, since many people commit offenses for which they are not caught, and addicts or psychotics sometimes have relapses of their deviant behavior which officials do not learn about.

Despite these imperfections, available records of followup information on supervised persons provide invaluable data for comparing statistically the outcomes of large cohorts of people who are followed up. It can usually be assumed that errors and incompleteness in case records are fairly randomly distributed, so that approximately the same proportion of error is present in the figures for all groups compared. If the official record of renewed deviance is found to be 30 percent for Cohort A and 60 percent for Cohort B, for example, the actual rate may be 50 percent for Cohort A and 80 percent for B. As these presumed actual rates suggest, the ranking of the two cohorts by outcome is likely to be the same by official and actual rates. In any event, we can never know actual rates; we can only infer their relative magnitude for different cohorts from the most complete and accurate data available.

It is appropriate to try to eliminate any possibility that less complete followup information is received for one cohort than for another with which its outcome is compared. Such variable completeness creates a large systematic bias. When there is no evidence of such a bias, however, we can assume that significant and consistent differences in outcome rates found from the most complete data we can procure indicate actual differences in relative outcome rates.

B. The More Complex Usual Circumstance: Some or All Subjects are Not Under Supervision of the Organization Conducting the Evaluation.

The major problems in procuring outcome data arise when some or all of the cases to be followed up are no longer super-

vised by the agency that seeks information about them. What if a State office wants to follow up a cohort of persons committed to its juvenile correctional institutions 5 years ago and a cohort of similar juveniles placed on probation 5 years ago, to provide more conclusive evidence than current rhetoric reveals about the consequences of plea bargaining and consequent grossly disparate decisions in our juvenile courts? What if a State department of corrections only half of whose prisoners are released on parole, and with diverse durations of parole supervision, wishes to determine how similar parolees and dischargees released 5 years ago have compared in criminality during the 5 years following their release dates, to determine if the public is better protected by parole than by discharge without parole? (Problems of systematic bias in such a study will be discussed later.) What if a State or Federal research agency wishes to follow up cohorts of narcotic addicts convicted of crimes but receiving civil commitment to addiction treatment centers with a cohort of similar addict criminals in the regular correctional system, to determine the wisdom of concentrating addicts instead of distributing them among other offenders? How would these State or Federal agencies procure the relevant followup data?

No procedure will provide absolutely complete information for the above comparisons, but at least some of a large variety of methods of collecting adequate data for evaluation are always possible and practical. It is appropriate to list all potentially possible methods below, starting with the most ideal but least frequently available method, and proceeding to the less ideal. One should also note that some of these methods are always feasible, and that use of multiple sources of followup data produces more complete and valid knowledge than use of a single source.

- (1) Request Current Fingerprint Arrest Reports ("rap sheets") from the Federal Bureau of Investigation.

Every adult in the United States, when formally arrested for a felony, and many when arrested only for a misdemeanor, is supposed to be fingerprinted. Usually they are fingerprinted, and sometimes juveniles are also. The fingerprints are sent to the Federal Bureau of Investigation with information on the reason for the arrest, as well as some identifying information such as name, date of birth, sex, race, and known or reported prior criminal record. In an increasingly large number of States the fingerprint reports also are sent to State bureaus of identification,

and sometimes to county, metropolitan, or other regional fingerprint record collection agencies.

An additional set of fingerprints is usually taken whenever a person is jailed or imprisoned, and frequently when he or she is committed to a State hospital or a public facility for addiction treatment, or when placed on probation or parole. The report accompanying the fingerprints usually indicates the reason for the confinement or release, and the sentence. Any change of status of a fingerprinted individual, such as discharge from sentence, and especially issuance of a warrant for the arrest of the individual as wanted (for example, escapees from institutions and absconders from supervision), is also reported to the fingerprint collection agencies.

Fingerprints are also similarly distributed whenever a person enters or leaves service in the U.S. Armed Forces or in a variety of types of employment requiring security clearance.

The purpose of these procedures is to supply police, courts, and other agencies with information about the possible criminal record of an individual and on his military service or other important career characteristics. The record thus produced is useful in determining if an individual is wanted by authorities somewhere, and to assess how trustworthy or dangerous he or she is likely to be. On request, every police or court agency and a number of other establishments authorized to make security checks can secure a summary of all fingerprint reports ever sent to the FBI on an individual. Usually this request for a summary of the prior record accompanies the submission of a fingerprint report on a current arrest or confinement. The provision of these summaries, known as "rap sheets," has been made tremendously more rapid in recent years through computerization of record storage, retrieval, and transmission. Indeed, it is now possible for many police agencies to procure criminal record information on an individual from the FBI in a matter of minutes, or even in less than a minute.

Obviously, the easiest way to procure the criminal record of a cohort of persons dealt with by a people-changing agency in a past year would be to request from the FBI the current rap sheets of everyone in the cohort. Each individual for whom a fingerprint record is initiated receives an FBI registry number, so the records can be quickly collected by the FBI if this number is provided; with somewhat more effort, the records also can be collected from identifying information other than the FBI number. The FBI has the obvious advantage over local or State crim-

inal record files of having information from every State, and even some foreign criminal record information, so its files would be more complete than others on individuals who incurred their criminal record in several different States or countries.

For about 5 years after the end of World War II the FBI routinely sent to prisons any subsequent fingerprint reports on persons whom the prisons had released, but this service was then discontinued for economic reasons for persons no longer on parole, perhaps because the information was seldom utilized. Since 1950 only a few followup studies have been provided with current FBI rap sheets; most requests for current sheets for any appreciable cohort of individuals have been denied on economy grounds. Even when the requesting agency has offered to pay for the service, and even when the requesting agency was the U.S. Bureau of Prisons, it has often been impossible to procure rap sheets for former prisoners except by requesting them a few cases at a time, without indicating that it was for an evaluative study. Finally, after a decade of announcing that it would be available "next year," in January 1973 a computer terminal in the Bureau of Prisons was made operational for unrestricted procurement of current rap sheets on past Federal prisoners, but only for those arrested after January 1, 1970, or released after February 1, 1973.

As I stated nearly 20 years ago, "the number of cases and the completeness of followup possible with FBI records would permit intricate statistical controls not practical in most research with other data" (Glaser, 1955). As I said again just over 15 years ago: "Use of FBI fingerprint records would permit us to know more about the types of former offenders who are not again convicted of major crimes, and concerning the relationship of such law-abiding behavior to the correctional treatment which they received. Nonrecidivists are the great unknown of criminology" (Glaser, 1957:683). Finally, nearly 10 years ago, I was more optimistic that this FBI function was about to expand, and I asserted:

The FBI's information on the criminal record of felons after their release from probation, prison, or parole, and their experience and resources for handling these records, make them the agency best equipped to ascertain the long-run felony recidivism rates which follow alternative judicial and correctional action for particular types of offender.

... maximum benefits will come not from omnibus evaluation of an overall program or a broad policy for all criminals, but from specific evidence as to which practices reduce re-

cidivism most for which types of offender (Glaser, 1964: 34).

After the above publication, the "Careers in Crime" section was added to the FBI's annual *Uniform Crime Reports*. This section provides postrelease arrest data for atypical samples of Federal offenders. It was begun with a group of subjects who were already recidivists and older than most offenders. Rather than being oriented to providing guidelines for sentencing and correctional decisions for specific types of offenders, this tabulation appears to have been initiated solely to denounce collectively a variety of practices that the FBI then called "leniency," including release from prison by parole rather than by unconditional discharge. The director of the FBI's Uniform Crime Reporting Section advised, however, in a letter of comment on a draft I prepared:

"Careers in Crime" was never intended to measure the success or failure of correctional programs since all persons entered into the system, until a legitimate followup period could be established, were, of course, failures (Glaser, 1967: 112).

After repeated disappointment, one hesitates to be optimistic again, but the FBI's gradual reorganization during the 1970s and its provision of some computer access to its Federal offender files for research by the Bureau of Prisons, suggest that guidance of judicial and correctional policy from the accumulated information in FBI fingerprint files will increase. Guidance could be much more extensive, routinized, and influential, if an office were established in the FBI specifically to assist State and local agencies in policy evaluation research through procuring follow-up statistics on cohorts of their past releasees. Until this is developed and even afterwards, much outcome information can be collected from more local resources.

(2) Request Current Fingerprint Arrest Reports ("rap sheets")
From State, County, or Regional Criminal Identification
Centers

Several of our larger States have for decades had their own central criminal identification agencies to serve local police, court, and correctional agencies seeking information on the criminal records of persons in their custody. These centers have procured fingerprint reports overlapping those of the FBI, and in some cases additional information from the agencies within their State.

During the late 1960's and early 1970's, with increased Fed-

eral and State funds available for combating crime and with the development in this period of improved electronic data processing, there has been a rapid increase in the number and quality of such State centers, as well as county and other regional centers in some areas. This effort has been spearheaded by a number of special demonstration projects funded by the Law Enforcement Assistance Administration of the U.S. Department of Justice, notably "Operation Search." Perhaps the most ambitious single State center of this type has been NYSIIS, the New York State Identification and Intelligence System.

While all of these criminal record centers have been oriented primarily to improving the speed and completeness with which criminal record information on individual cases is provided, especially to law enforcement agencies, the utility of their records for evaluating judicial and correctional policy has not been overlooked. It has been given second priority because there has been little demand for it, and because the criminal record centers were sponsored by police and prosecution offices, but when local correctional researchers have requested followup data on cohorts of past offenders from State agencies, they frequently have received excellent service.

The New York State Division for Youth's (1970) study of offenders referred to its treatment centers was unique in that State's history in having recidivism arrest data, thanks to NYSIIS. While they have not yet followed up comparison groups or attempted cost-benefit analysis, the groundwork for correctional evaluation has been created there. Similar projects based on followup information from NYSIIS have been undertaken in the Nassau County Probation Office and in other agencies.

Florida, Alabama, and several other States in the Southeast have initiated followup research utilizing State criminal record information centers. California has long had such service, and its Department of Corrections and Department of the Youth Authority probably have conducted more correctional evaluation research than all other States combined.

As indicated earlier, State criminal record files may not be as complete as FBI files with respect to the criminal record that a State's releasees accumulate in other States. This only parallels the fact that the most complete official record possible is actually incomplete, because it lacks data on crimes for which an individual was never caught. Nevertheless, if there is no reason to believe that compared cohorts will differ in the proportion of their crime that is recorded only out of State, outcome

rates based only on State data should yield valid evidence of the relative criminality of the different cohorts. Indeed, most offenders do not wander far in their criminal careers, and even the criminal records of a single county or large city can provide valid indicators of the relative recidivism rates of different groups of offenders. Nevertheless, many agencies already obtain, for other purposes, much information on the recidivism of their releasees in other jurisdictions.

(3) Data From Inquiries Made by Agencies Holding One's Past Releasees

Whenever a correctional institution receives a rap sheet indicating previous incarceration of one of its newly admitted inmates, it is customary for the institution to send inquiries to the places where previous incarceration occurred. This inquiry is not universal and does not have a standard form, but it is widely done for it usually requests information on the subject's behavior while previously confined. Its main purpose is to obtain warning of any inmate's having been a danger to the security and order of the institution during prior incarceration, so that appropriate precaution can be taken during the current confinement. Sometimes the forms will make additional inquiries, for example, asking about work and educational performance. Usually these inquiries are answered routinely, then either discarded or filed and not used further.

The presence of such inquiries in the files of past releasees of a correctional institution, combined with the institution's own record of those releasees whom it has received again, provide the data needed to tabulate a reconfinement rate for past releasees. Frequently the inquiries do not indicate why the individual has been reconfined, nor what the new sentence is, but this information could be requested. Robert B. Levinson, Coordinator of Mental Health Services for the U.S. Bureau of Prisons, has suggested (in a 1972 conversation with me) that the American Correctional Association sponsor use of a standard form for the exchange of information among correctional institutions whenever one prison receives an inmate who was confined previously in another facility. This would provide information useful for both operations and research.

(4) Other Types of Followup Information

There are a variety of methods of tracing cohorts of people to discover their experiences after they received or failed to receive particular types of treatment. The tracing of thousands

of delinquents was done repeatedly, for over four decades, by the Gluecks (1930, 1940, 1950, 1968). Lee N. Robins and her associates (1966) traced 524 persons seen by a child guidance clinic 30 years earlier and 100 control subjects. In another study (Robins and Hill, 1966; Robins and Murphy, 1967), they traced 296 black children from entrance into first grade in St. Louis schools until completion of the upper limit of juvenile court age. More recently Wolfgang and associates (1972) traced during the late 1960's the delinquency records of 9,945 Philadelphia schoolboys born in 1945 who lived in Philadelphia between their 10th and 18th birthdays. Many other examples could be cited.

All of these studies relied heavily on local school, police, and juvenile court records, but used a variety of other methods to find out what happened to the people in the cohorts they were tracing. In addition to schools, police, and courts, social service exchange registers, Selective Service records, medical and mental health records, credit information files, and other compilations, as well as telephone directories and city registers, can more or less readily be used to: (a) locate people for interview; (b) find out what information about them has become a matter of public record; (c) find nonpublic information about them that can be made available to responsible researchers if there are guarantees that data on individuals will be kept confidential and only statistics on groups released. The availability and utility of records depends on the authority and responsibility of the agency collecting the followup data, the kind of data sought, and the legal and ethical restrictions that are therefore applicable.

None of these sources of information will be complete, but my comments on other types of imperfect data are applicable. If the deficiencies of data appear to be randomly distributed among the cohorts one is comparing, one can with confidence assume that the relative outcome rates the data reveal will be comparable to actual rates.

Systematic Bias in Followup Data

A major source of systematic bias usually exists whenever a cohort that is intensively supervised in the community is compared with a cohort that is less intensively supervised. This could occur, for example, if the postrelease deviant behavior of parolees is compared with the postrelease deviant behavior of discharges, or if the postrelease conduct of parolees or probationers intensively supervised in small caseloads is compared with the conduct of similar persons under regulation supervision in stand-

ard caseloads. Intensity of supervision determines how much officials know about releasees. Those who are discharged without supervision or are discharged with little supervision, therefore, have their postrelease deviance less fully reported than those released with more supervision. Evaluation using as its criterion the rate of reported deviant conduct is thus biased against a favorable assessment of intensive supervision.

While the possibility of this type of bias has been mentioned in reports and comments on experiments with diverse intensities of supervision, little has been done to try to take the effects of such bias into account in assessing outcome. In general, *the best correction for differences between cohorts in completeness of criterion information is to compare them only as to outcome data on which information is equally complete for all cohorts compared.* It may therefore be appropriate to compare rates of deviance for parolees and dischargees, or for other cohorts differing in intensity of supervision by only recording offenses for which they are apprehended by the police, ignoring arrest or re-confinement resulting just from being taken into custody for technical violations by a parole, probation, or aftercare officer.

It may be argued, however, that the procedure indicated above sometimes creates a counterbias instead of eliminating bias, for under this procedure the offenses discovered by the supervision staff and resulting in their taking a client into custody instead of calling the police would not be counted. In contrast, similar offenses by those unsupervised or less supervised would lead to arrest by the police and be counted. A first step to correct both bias against intensive supervision and possible counterbias is simply to find out what the customary role of supervision staff is in the arrest of persons whom they supervise.

In many supervision agencies staff never make arrests, but request that this be done for them by the police, both for violations of supervision regulations and for new offenses. If researchers find out how often supervisors request arrests for new offenses, they will be able to estimate the extent to which bias results from either counting or not counting these arrests. It is my impression that most supervision staff rarely request arrests for new offenses by their clients; they usually learn of such offenses only when they are notified that the police have arrested their clients. The supervisors only request an arrest by the police when their clients have not reported and cannot be located, so that they are presumed to have absconded. If this is the usual practice, then data on police-initiated arrests for new offenses

will suffice to provide an unbiased comparison of the rates of new criminality of cohorts differing in the intensity of supervision they receive.

If supervision staff does take persons into custody for new offenses, as may occur with intensively supervised juveniles, it will be difficult to make an unbiased comparison of the post-release criminality of their clients with that of similar persons who are not under supervision in the community. The bias will then be against the intensively supervised cohort if all reconfinement is counted, and it will be against an unsupervised or less supervised cohort if only arrests by the police are counted. All that one can do to offset such a bias is to estimate its magnitude and try to take it into account in assessing findings. If one cohort in fact has much less crime than another, this may be evident despite bias of the data against revealing it. Thus, if data collection is biased to reveal a larger proportion of the crime committed by Cohort A than of the crime committed by Cohort B, but the data show a lower crime rate by A than by B, then one can have great confidence that A indeed has a lower rate than B.

Still other types of systematic bias may occur in the assessment of outcome rates for two cohorts with different intensities of supervision. The most intensely supervised group could conceivably have a record of less crime than the less supervised group simply because those who are closely supervised are reconfinemented for violations of rules (e.g., not working, drinking excessively, being with disapproved companions, being out late at night) before they have a chance to commit crimes, or before they are arrested for whatever crimes they may have committed. Some commentators have claimed that a high confinement rate of supervised releasees for rule violation and a low rate of new crimes by them implies that close supervision and rigorous rule enforcement are crime prevention methods (cf. Ohlin, 1951:43-44). Others object that reconfinement for rule violation reveals the arbitrary power of supervision officials to impose penalties for petty deviance, not even criminal, often greater than the penalties which courts would impose for felonies, and that such reconfinement creates in the imprisoned rule violator a sense of injustice conducive to recidivism (Irwin, 1970:170-173; American Friends Service Committee, 1971:90-91). Reconfinement merely for noncriminal rule violation seems to be decreasing markedly, however, due to court decisions requiring a quasi-judicial due process for proving violations before reconfinement

is permitted. As a result, in many jurisdictions most confinement for parole or probation violation only follows a judicial conviction for offenses committed while under supervision.

Whether intensity of supervision actually reduces or increases crime, and the social costs as well as the possible benefits it produces, can be assessed well only by long-run followup of cohorts of similar individuals who received different intensities of supervision. The effects of indeterminacy of sentence and of styles of parole or probation supervision on crime rates would certainly be much more conclusively demonstrated than they are now if they were evaluated with followup periods of 15 or 20 years or longer, so that all subjects studied had extensive opportunities for freedom if law-abiding. Whether such a long followup can be made without waiting that long will depend in large part on the accessibility of good records for relevant samples of persons released 15, 20, or more years ago, and whether they can be followed up currently, preferably in FBI records. Any routinized evaluative research on current cases could be tremendously augmented, and well-tested conclusions could be procured much more rapidly if extensive retrospective research were done on the postrelease outcome of cohorts released decades ago.

Most often the cohorts differentiated for comparison in evaluative research are those that received distinctively different types of treatment—psychotherapeutic, educational, employment assistance, etc.—rather than different degrees of supervision. Each cohort different in treatment, however, may also receive a different average intensity of supervision during the followup period, due to differences in the proportions receiving parole, for example, or in the duration of their probation supervision. Such possible sources of systematic bias in outcome data should be investigated in all evaluative research and appropriate adjustments should be made to control such errors or to estimate their impact. Thus one could restrict assessment of outcome to similar members of the cohorts compared who had similar intensities of supervision, or one could limit criterion data employed to those which were unaffected by intensity of supervision.

Duration of Followup

One obvious deterrent to evaluative research is the length of time needed for a followup of treated and comparison or control cases. The pressure for evaluation tends to be greatest for procedures that have been given support because they are new

and controversial; great promises usually have been made for them, but since they also have articulate critics, there is a demand for quick evaluations. Yet any procedure to change the behavior of people can be tested only by allowing enough time to elapse to permit adequate observation of whether or not the subjects do behave differently. How long a period suffices for adequate observation? Can we make it shorter?

There are no single answers to the above questions, because the answers depend upon the people-changing problem, the criteria of effectiveness, and the treatments to be compared. As already indicated in Chapter 4, if the criterion of treatment achievement is cost-benefit profit and efficiency, and if one group is institutionalized much longer than another, a long followup usually is necessary. This requirement arises since there are clear short-run benefits of complete protection of society by incarcerating persons who victimize others, even if the cost of incarceration also is high. Prolonged followup, however, may reveal that these short-run benefits are offset by greater post-release cost in the long run, from released inmates committing more victimizing deviance than is committed by similar persons treated only in the community.

Obviously, for any given evaluative study, the longer the followup period, the more confidence is warranted that additional time would not alter the conclusions. But what clues are there of diminishing returns from additional waiting? Conversely, when is it best to be patient and not trust initial conclusions? Several factors may be most important to consider in trying to answer these questions.

One key determinant of appropriate length of followup is the probability of change in the lives of the clients at the time of the people-changing effort. A long followup is appropriate to test treatments for clients of types that have generally had a very low probability of changing. These include, for example, alcoholics, opiate addicts, or persons who have supported themselves exclusively by crime most of their adult lives. Treatment which occasionally has impressive short-run achievements with them such as aversive conditioning and social pressures in an institution or residence that are not readily sustained after their release, seem at most only to defer relapse, not to diminish its prospect appreciably. Similarly, persons who have a very high probability of changing probably will not have their high probability of success greatly enhanced in the long run, even by measures that appear to help them. Thus the Provo probationers in

a special treatment program (described in the preceding chapter's discussion of pre-post comparisons) had a greater decline in arrests during their first year on probation than the control group under ordinary probation, but in longer-run followups (of 2, 3, and 4 years' duration) both of these probationer groups had about the same greatly diminished arrest rate (Empey and Erickson, 1972:210).

The prospects that a short-run followup will be sufficient seems greatest for clients who have not been highly committed to the behavior a program is trying to alter, and have also not been highly committed to alternative types of behavior. Most persons arrested for crimes of a predatory type (burglary and theft, particularly auto theft) are teenagers or youths in their early twenties. Even when they have records of recidivism, especially when they have not become highly specialized in their offenses, most have also had some work experience and orientation to legitimate occupations, but little success at it. A "change in luck" and in social support, at crime or at work, may alter their subsequent careers appreciably. As indicated in Chapter 3, it is with groups and criteria for which the base rates of success are closest to 50 percent that people-changing efforts seem most likely to be able to alter outcome rates appreciably, although this numerical rule does not hold true in every situation. When clients have little experience with success at alternatives to deviance and little commitment to deviance, even a short period of what is for the subjects an unusual success at legitimate pursuits in an available postrelease occupation and social setting may greatly increase long-run prospects of less deviance.

One clue that a short-run followup is significant in these cases is a progressive improvement in the success rate of the treated group as compared with the control or comparison group, with each added interval of followup period. This was evident in the Provo Project, where the incarcerated group had increasingly more arrests per case than the community-treated group with each additional followup year (Empey and Erickson, 1972: 184). In the Adams (1961) comparison of treated and untreated amenable in the Pico Project, a progressively increasing advantage of the treated over the untreated group was especially evident when reconfinement time was used as the criterion, rather than arrest rates. Because reconfinement time reflects immediacy and seriousness of relapse in addition to its occurrence, it seems to be an especially quick and sensitive index of shift from commitment to crime as a livelihood. Unfortunately, there has

not been nearly enough repetition of the type of investigation of criteria sensitivity that Adams pioneered.

Another pioneering study of a type that should be repeated was Ohlin's (1951) 20-year test of Illinois parole-prediction tables. In comparing experience for four 5-year periods, from 1925 to 1945, for 27 variables, he found 12 that were much more stable in their prognostic significance from one period to the next than the remaining 15. He also found that for a 12-item prediction table there was a fairly constant ratio between the violation rate from a 1-year followup and the violation rate from a 3-year followup. From these "parole period ratios," as he called them, he could predict the violation rates after 3 years, for groups of a thousand or more parolees, from data on their violation rate at 1 year, making an average error of only two percentage points (Ohlin, 1951:Appendixes D, E, F and G).

It was on the basis of the discovery of relatively constant ratios between outcome from short and long followup periods that Ohlin (1954) proposed a system for "the routinization of correctional change" which suggested the title of this manual. This system would require continual checks on the failure rates of alternative programs for different types of offenders, and possible alteration of policy when only a 6-month or a 1-year followup indicates that a particular type of client is having appreciably more difficulty in one kind of program than in another. Such continuous assessment of change by routinization of evaluation would be much easier than it is now, however, if our data sources were improved.

Conclusion

Prerequisite to evaluation of people-changing efforts is follow-up information on the behavior of clients after the attempt to change them, followup of an appropriate control or comparison group, and a cohort perspective toward all past clients. Procurement of followup data presumably is simplest while the subjects on whom information is desired are under supervision of the organization conducting the evaluation, but collection of such data is frequently impeded by multiple administrative numbers and files for the same person, and by deficient or discarded administrative records. When some or all of the subjects are no longer under supervision of the organization seeking followup data, information must be procured from Federal, regional, State, or local record compiling agencies or from other tracing resources. These usually are oriented to procuring data on one individual

at a time and often resist or give low priority to requests for data on statistical cohorts of appreciable size. Such obstacles are gradually being overcome, and alternative sources of information often are available when obstacles are encountered with one potential source. One such alternative source may be the releasing organization's own postrelease correspondence on its clientele.

All followup information on deviant behavior, and much on conforming behavior (e.g., employment), is likely to be incomplete. Use of such data for evaluation of people-changing efforts is nevertheless justified if we can assume that the degree of incompleteness is about the same for all groups compared. Such an assumption is questionable, however, whenever we can infer that systematic bias makes followup data more complete on one group than on another with which it is compared, a circumstance especially probable if one group is more intensively supervised than another during the followup period. One remedy for this problem is to limit the followup data utilized to those which are most likely to be equally complete for all groups compared, such as data collected by a nonsupervising agency, and long-term followup information.

While the longer the duration of a followup period the more adequate evaluation will be, a major impediment to evaluation research support is the desire of evaluation sponsors for immediate assessments. A long followup is especially crucial to valid conclusions on efforts to change behavior patterns that usually are highly persistent, such as an addiction or professional crime. One clue that initial followup results will not be reversed in time is a progressive increase in the effectiveness revealed by each additional increment of followup duration. A possible source of confident followup abbreviation is discovery of a relatively fixed ratio between outcome rates from short and long followups.

While followup information is essential to evaluation, information on the treatments to be evaluated and on the clients to whom the treatments were applied also is essential. Enhancing the quality and efficiency of such data collection is the concern of our next chapter.

CHAPTER 8

AUTOMATING INPUT DATA: THE INTEGRATION OF RESEARCH AND OPERATIONS RECORDS

In order to obtain pretreatment and treatment data for evaluative research, reliance has frequently been placed entirely on the files maintained by people-changing agencies for administrative or casework purposes. This certainly is cheaper and more convenient than procuring separate data for evaluative research. Indeed, for retrospective inquiries on events and conditions before the research began, existing records may provide the only information available. Nevertheless, currently available records usually are grossly deficient for both operations and research. The analysis of these deficiencies will suggest that records can be improved most if they are designed to serve both operations and research purposes simultaneously.

Deficiencies of Operations Records for Research Purposes

Despite the economy involved in using agency administrative and casework records for research purposes, scientific investigators of the correlates of favorable and unfavorable outcome from treatment have long complained about the inadequacy of the data for compiling statistics and tabulating statistical relationships. Prominent among the deficiencies noted are the following:

A. Operations records vary greatly in completeness. Some administrative or casework staff jot down detailed information on all items, but some make few or no entries on many items, even when standardized forms are used. Yet the compilation of statistics on an item requires an entry on that item for all cases. One cannot determine the relationship between vocational training in prison and subsequent criminality, for example, unless records indicate fairly completely which prisoners received vocational training and

which did not. In addition, they should indicate the type and amount of training for each case, and the grades or credits achieved in it.

B. Operations records often vary in their terminology for describing the same item, in the aspects of the item which they emphasize, or in the dimensions they employ to indicate the item's magnitude or quality. Such variation, of course, impedes tabulation of statistics on an item. One cannot compile statistics on the vocational training received by a cohort of past prison releasees, for example, if one individual's file only mentions the vocation that he or she studied, another individual's file only indicates that the subject was industrious at vocational training, and a third individual's file only comments on the high or low degree of skill demonstrated in this training. This diversity is further confounded when some staff use one set of categories, such as "adequate" and "inadequate," and others use alternative terminology, such as the estimated percentile rank, in describing the degree of industry or skill manifested. Records that consist of narrative accounts or comments usually contain all possible mixtures of terminology, as well as much variability in the thoroughness with which they describe their topics.

C. Operations records are often bulky and inefficient when used for the retrieval of information. Administrative or case records frequently consist of long narrative reports with a large number of diverse documents overlapping in their information and jumbled in a thick file. Compilation of statistics on hundreds or thousands of cases from such files therefore requires a tedious and error-prone search that is extremely costly and inefficient.

D. Operations records simply were not designed for research purposes, and therefore many neglect to record the kinds of information researchers desire.

It is partly because of these deficiencies of operations records for research purposes that so much evaluative research in people-changing agencies is done on a project basis, instead of being routinized like the accounting system in a business. Evaluative research projects are launched repeatedly, each designed to collect its own information on the cases that it is to follow up, instead of relying on data in the administrative or casework files.

Despite the shortcomings of operations records for research

purposes, the alternative source of data for evaluative research—information collected specifically and only for research needs—also has serious deficiencies. These defects vary, depending on the method of collection employed.

Deficiencies of Special Records for Research Purposes

Let us first consider the common procedure of devising forms which operations personnel are requested to fill out for each of their clients and send to the research office. The serious faults common in data collected by this method stem from problems of communication and reinforcement.

If forms are to be completed away from the office that uses them, they should be instituted only after much pretesting by those who are to make the entries and trial tabulation and analysis of the pretest data at the collecting office. This is to minimize the possibility that items on the forms will be interpreted differently in the field than at the office. Even if the forms are pretested carefully, however, there are likely to be frequent shifts in the way that details are construed in the field, especially when changes occur in personnel or procedure at the location where the forms are completed. If data collection is to be valid and reliable, the research offices of most people-changing organizations must undertake frequent monitoring and training of the treatment and clerical personnel in the field who fill out research forms. But even when these precautions are taken, information on research forms is often carelessly and indifferently reported simply because there is little reward for doing it well or penalty for doing it poorly.

When a research office which receives a form does not have direct authority over the persons who fill it out, and when those who fill it out are not likely to see the form again or be reprimanded if their entries are inaccurate, one cannot expect good data collection. The task of completing such forms gets low priority and falls behind schedule when other work is pressing, and is then done hastily and poorly. Filling out forms which requires judgment by professionals is often delegated to clerks, who are given instructions by the professional on rather mechanical methods for deciding what entries to make. In time these emergency shortcuts become routine procedures, as in the following case.

During the 1950's the U.S. Bureau of Prisons developed "research forms," each as a separate mark-sense IBM card, on which institution caseworkers were to make a summary of the

diagnosis and recommendation on each prisoner at the conclusion standard set of categories, many of them judgmental, on criminal, personality, and socioeconomic factors in each case. It was suggested that I use these cards in the evaluative research I had been recruited to undertake in the Federal correctional system. I discovered that many of the cards were incomplete and that they were not received regularly in Washington. I inquired at a Federal penitentiary as to how these forms were prepared there. The caseworkers did not recognize them, but someone thought they were filled out by the clerks in the records office. At the records office they said that these were sent to Washington by one of their clerks, "Mr. Smith," who was then on vacation. When I visited Mr. Smith after his return, he had a large stack of files on his desk from which he was hastily filling out the forms by scanning several pages of narrative reports prepared by caseworkers, psychiatrists, and others on each prisoner who had been received in the past month. There was only a crude relationship between the judgments indicated in these files and those which Mr. Smith entered on the cards; indeed, he had developed perfunctory entries, on which he could decide largely by guess after only a quick glance at each file.

This illustration parallels data collection circumstances in many other agencies. Even where professional persons, such as probation officers or psychologists, provide information for research, unless they routinely receive feedback which has consequences for them, their entries on the data forms will often be made with little care or consistency. Usually this carelessness in completing research forms (after their novelty wears off) contrasts sharply with the concern shown in preparing forms consulted subsequently by others in the same agency (e.g., court, clinic, or prison) as a basis for important decisions. One thing is certain: any researchers who employ data sent them by personnel who do not use again the reports they send to the research office would do well to make frequent field observations of how the reports are actually filled out.

An alternative to using either operational records or forms filled out by operations personnel for researchers is to send research personnel into the field to gather information by interview or observation, or consultation of local records. This has the obvious advantage of giving the research office more direct control of its data collection. Often this control is essential if new types of tests or interviews are used, if a detached and objective perspective is desired, or if the expertise needed for the data

collection is not available in the field. On the other hand, it is expensive. Not only will research employees sent into the field have to incur considerable transportation and hotel expenses, but they often must work inefficiently because the subjects from whom they seek information can be seen only in their leisure time or at the convenience of their custodians. There are other disadvantages in addition to expense.

Because research personnel in the field are outsiders, they may not be sensitive to all the aspects and implications of what they see or are told. Therefore, they may frequently be misled or make erroneous inferences. Information collected by outside researchers may thus be different from that which would be obtained by those continually on the scene as part of their everyday life. Sometimes this problem is solved by research offices employing personnel on a permanent basis at field data-gathering locations, but this may also involve expenses or drawbacks compared with what will be proposed later as a more optimum solution, to obtain simultaneously information for operational decisions and information for evaluative research.

Deficiencies of Records for Operation Purposes

Shortcomings of operations records for research purposes already have been indicated, but it is appropriate to point out here that these records also are frequently an impediment to operations. This will justify our ultimate recommendations in this chapter that routine records be revised to serve both research and operations more adequately.

The deficiencies of operations records as a source of statistics will not concern us here. Incompleteness of entries on some items of the forms used, or variations in terminology from one report to the next, may not be a serious problem in most cases to those who consult records for operational decisions such as deciding the best assignment or recommending a change of status for a client. What is important to them is that the records contain the information most relevant for the specific decision they must make in a particular case. What is relevant depends upon the client involved; the same information is not needed for everyone, and it can be in diverse terminologies as long as it is clear. It is also important, of course, that preparation of operational reports be efficient, that information desired be readily found in the reports, and above all, that the information be valid. Deficiencies of operations records in meeting these various needs will be dis-

cussed separately for narrative reports and for performance ratings.

A. Narrative Reports

Operational records are especially likely to impair the effectiveness of people-changing agencies if the utility of the records is not systematically tested, if they are inefficiently prepared or used, and if their preparation is a substitute for rather than an adjunct to people-changing endeavors. It will be contended here that all these faults are especially characteristic of that favorite product of casework, the narrative report.

Staff presumed to have special treatment functions and expertise—such as caseworkers, counselors, probation officers, parole officers, psychologists, and psychiatrists in all the many types of people-changing agencies—are called upon not only to deal with their clients, but to provide guidance in agency decisions. Their reports are relied upon to justify either their own decisions or those judges, superintendents, boards, commissions, or committees. This is particularly crucial in decisions on when and under what circumstances the persons reported on should have their freedom reduced or increased. These reports are also used as guidance for a variety of decisions regarding the particular type of service that should be given a client, and for recommendations—often prepared years later—that affect the opportunities of clients in subsequent pursuits. It is from these reports in the files that letters of reference are prepared when requested, for example, by employers or schools to which former clients of a people-changing agency may apply.

These reports thus have many functions that are important and manifest, which warrant their receiving much staff attention. The reports also have latent functions, however, which probably account for the fact that their preparation tends to receive an even larger proportion of the time and energy of treatment staff than usually is assumed. For example, Federal probation officers calculate workloads by counting the preparation of one presentence investigation report as equivalent to the supervision of four probationers or parolees for a month, but time-activity research indicates that the average officer gives more time to the completion of one presentence study than to the supervision of a dozen probationers or parolees for a month (Glaser, 1969: 299-303). My colleague Robert Carter advises he also found a 12-to-1 ratio in a Washington State study. A time study in the California Youth Authority indicated that when a

parole agent's caseload is reduced, the parolees are still supervised much as they are in a larger caseload, but that the extra time is devoted mainly to paperwork (Johnson, 1952).

How does one explain this unexpected, high priority for report preparation, at the expense of activity with or on behalf of the client? Several factors seem to be operating. In the first place, the report preparation produces a tangible document in which the author can feel some pride or can perceive remediable deficiencies, while people-changing endeavors with or on behalf of the client, such as counseling and investigating, are often of uncertain value. In most cases, if benefits or defects of treatment services are inferred, they are perceived weeks or months after they are provided, but the merit or deficiency of a narrative report is sensed as it is written, and can be reassessed immediately by rereading.

Secondly, the quality of a report's preparation is quickly evident to the superiors of the person who prepares it, while the quality of relationships with clients and of services to clients are not so immediately apparent, especially when clients are not in an institution. Therefore, superiors are much more likely to praise or criticize a staff member because of the reports he or she submits, and to promote or demote on this basis, than to take these actions because of the much less visible impact of the staff member on clients. This occurs especially in the typical probation, parole, and aftercare agency, where staff see clients infrequently.

Thirdly, the liberal arts education and middleclass background of many treatment staff, in contrast to the predominantly low education, lower socioeconomic class background, and minority ethnicity of most of their clients, often affects work preferences of staff. These personnel often are uncomfortable and inept in communicating with clients, but gain a sense of achievement from the creative writing involved in preparing narrative reports to be read by persons of higher rank than themselves. It is small wonder, in such circumstances, that extra time from reduced caseloads tends to be used by treatment staff more for polishing reports than for trying to improve their services to clients. Diagnosis permits pontification; it is easier than treatment and tends, therefore, to replace treatment.

It would follow from the three points above that, while narrative reports have the manifest function of providing information that may guide decisions on the treatment of clients,

these reports have the latent function of reducing tensions in treatment staff. There is also a latent function of the reports in addition to those indicated by the three points above; narrative reports may relieve officials of feelings of guilt or anxiety about the merits of their decisions, and they may protect officials from criticism for these decisions.

It has frequently been observed that humans tend to defer action until they can indicate satisfactorily, to themselves and to others, a verbal justification for a decision to act in a particular way. Many narrative reports in people-changing agencies are expected to conclude with a recommendation for some action with respect to the client, such as granting or denying probation, parole, transfer, penalty, or a particular assignment. Other reports are specifically intended to explain, and thereby to justify, an action that has already been taken, such as issuing a warrant for arrest of a parole violator.

In much writing that concludes with a recommendation, the recommendation is anticipated in preparing the entire report. Once a decision is made on what the recommendation will be, the rest of the report tends to be written to support the conclusion, in spite of efforts to show all pros and cons. When the report must explain or justify an action already taken, however, it usually is especially slanted, for the author has a stake in showing that the action was proper. Thus petty infractions that previously would be overlooked, and alleged misconduct—perhaps only the impressions of questionable observers, which usually would be given no significance—are added to the accounts of major infractions to support a diagnosis and decision that a particular client is dangerous or sick, and therefore that an arrest warrant was justified. Any board or official acting on the basis of a slanted report, whether favorable or unfavorable to the client, is likely to feel much more confident than he would if his information were less conclusive in its implications. Should events later prove that his actions were inappropriate, he can point to the report as evidence that his decisions were reasonable under the circumstances.

An additional defect of narrative reports is that it is extremely difficult for officials to test their validity. A very imperfect test occurs if reports are received from several independent sources, each making recommendations autonomously. In such circumstances, consistency creates confidence in validity, while inconsistencies foster doubt and suggest items meriting further inquiry.

The question of validity applies both to the information which the reports present and to the inference that this information justifies a particular recommendation. The only immediate basis for assessing validity, especially if only one report is received, is confidence in the author of the report and an impression that the account is plausible and reasonable. Therefore, a narrative report's validity and often the competence of its author, tend to be assessed more by the verbal style and skill of its rhetoric than by the objective merits of its contents.

The validity of narrative reports is especially difficult to check scientifically. The deficiencies of operations records for research purposes, which were described in the preceding section of this chapter, are more acute with narrative than with most other types of operations documents. Narrative statements are especially diverse in completeness and in the terminology they employ for describing a particular item. They are also most difficult to use when seeking specific items of information they are presumed to contain. The latter defect is an impediment for operations use as well as for research; if one desires a particular fact, such as the intelligence test score of a client, the personality assessment he received from the psychiatrist, or the names of the client's criminal associates, it is much easier to find these in standardized forms that have a space for these items than to dig through narrative accounts in search of them.

The remedy for these problems with narrative reports, of course, is to have precoded reports in standard categories which staff can simply check to indicate the information they wish to report. Before considering the merits and faults of such reports, and how the merits might be maximized, it may be well to consider the special problems of reporting and utilizing staff observations of the behavior of clientele.

B. Performance Ratings

The growing recognition that people-changing can be done, if at all, not just by treatment staff but by any employee in contact with the client, has increased demands that a large variety of personnel provide higher officials with ratings on agency clients whom they observe in the course of their work. For example, supervisors, house mothers, teachers, nurses, coaches, and guards—in a variety of total institutions, residences, workshops, and schools—now submit periodic reports on the clients assigned to their care.

It is presumed that these "line personnel" derive a special expertise in assessment from seeing their clients for several hours at a time on many different days of the week or month, and when the client is relaxed and with peers rather than across the desk in an interview cubicle. Competence of line staff in rating clients may also come from their similarity to the clients in class or cultural background, or from their expertise in the activity at which they observe the clients, such as vocational training or athletics. These ratings by line staff may be submitted directly to decision-making officials, or may be given to the caseworkers, who summarize and interpret them in their narrative reports for higher officials.

One frequent source of defect in performance ratings by line staff is deficiency of feedback to them, but another type of defect may reflect the source of feedback. If the entries on rating forms have no consequences for the person who fills out the forms, as when he or she mails them off and never hears about them, the forms are likely to be completed indifferently and carelessly. Conversely, if these ratings are regularly discussed with others, there will be more concern shown in preparing them. When clients see the reports made on them, however, staff may be reluctant to rate anyone unfavorably, or their ratings may reflect efforts to motivate or manipulate particular clients or groups of clients more than efforts to assess them accurately.

The effectiveness of performance ratings as devices for motivation may, of course, justify disclosure of the ratings to clients even at the risk of some impairment of the validity of the ratings. In any case, if ratings of clients are discussed by the raters with those who subsequently use the rating forms, such as caseworkers, idiosyncracies in the entries on the forms may be taken into account for operations decisions, if not for research.

If performance ratings are shown to clients, the ratings should be especially objective, for they must then refer only to behavior explicitly encouraged. For example, in rating performance on vocational training, a secret rating form might ask, "Does subject seem to like his trade?" or "Does subject tell others that he (or she) wants to pursue this trade after release?" But if such a rating sheet were shown to the subject it would simply encourage verbal behavior to manipulate the performance rating. A sheet shown to the client should preferably refer only to nonverbal behavior that is explicitly encouraged, such as "Does subject begin work immediately on arrival and continue until

told to stop?" or "Does subject clean the tools and equipment before turning them in?" In most situations it probably is not wise to assume that any performance rating sheets can be kept secret. Therefore, items should usually be phrased to refer only to behavior overtly encouraged, rather than to behavior that can be construed favorably only if the subject does not know it will raise performance ratings.

The language in performance ratings poses the problems of narrative reports if it calls for answers other than "yes" or "no" or other alternatives that can be checked off on a printed form. If the form consists of open-ended questions or instructions, such as "Describe ways in which subject shows positive or negative motivation," to be answered by any comment deemed appropriate, the ratings will be diverse in coverage, completeness, and terminology. If the reports are structured, as when staff are asked to check one of a set of scaled adjectives (such as "excellent," "superior," "fair," "inferior") to indicate a client's skill or diligence, different raters will have different interpretations of these adjectives. Some, for example, may call three-fourths "superior," and still others describe most as "fair." One solution is to ask them to classify everyone by a particular ranking of fractions, such as "top third," "middle third" and "lowest third" of their group, or of all persons rated. In any case, it will be useful for interpretation of these reports if the distribution of ratings made by each staff member is tabulated, so that the rater and everyone else know what fraction of the clients have received each of the alternative possible ratings.

In designing instructions for performance ratings, it is well to bear in mind that the most objective measures frequently are the best. In work assignments of a standardized nature, this may simply consist of the amount of goods produced or tasks performed in a given period, compared to averages or percentiles for similar clients at that assignment. In training or education assignments, it may consist of scores on standardized tests. For most situations, however, no such measures are readily available. In these typical circumstances a procedure used by Catholic University psychologist Antanas Suziedelis (1963) for Federal prisons provides an excellent model for the collaboration of researchers with operations personnel in the design of rating forms (Glaser, 1964: 245-250). This model employs standard procedures in psychological measurement applicable to many situations where rating forms can be useful.

The first step in this suggested procedure for scientific development of a rating form is to recruit aid from persons familiar with the situations where ratings are desired and expert in the tasks pursued there. Request that they submit short statements describing clearly desirable or clearly undesirable aspects of a client's behavior in such situations. The language in these statements should, if possible, refer to objectively observable behavior rather than to inferred situations. Thus, in vocational training activities, either in classes or on the job, instead of statements to the effect that the subject seems interested in the work he is doing, it would be preferable to have statements such as: "The subject begins work immediately on arrival and continues until it is time to stop," and "The subject tries to avoid repeating mistakes after they are pointed out to him." Actually, about half the statements should ask for agreement or disagreement with statements describing the person to be rated unfavorably, and about half should seek responses on statements describing this person favorably. Such a mixture will correct for the problem of "response set," in this case, the tendency of some raters to agree disproportionately with unfavorable statements, and of others to agree disproportionately with favorable statements, without considering the statements carefully. In addition, care should be taken to select statements applicable to all situations where the rating form is to be used.

To maximize the variety of statements that will be suggested, it is preferrable to solicit them from persons with diverse characteristics, from different locations, and with different kinds of relevant background. Small groups of these persons might be brought together in "brainstorming" sessions to collaborate in developing statements. Or contests can be held among individuals or among groups, to motivate the formulation of as many distinct and relevant statements as seems possible.

After a large number, perhaps several hundred, of these statements are collected, they should be screened to eliminate duplication. If several statements refer to the same thing, those with the clearest phrasing should be retained, or the best phrases in each should be combined in a new statement to replace the others. "Double-barreled" statements, such as those that refer to two or more distinct qualities (laziness and lack of skill), should be avoided in a single sentence. These should be replaced by separate statements, each dealing with an indicator of one quality in the performance of the person rated.

The result of the screening process described above will be an "edited list." This list should be distributed to a group of persons who would be receiving the performance ratings and making decisions on them. Such recipients, for example, might consist of superintendents and assistant superintendents, or members of release-granting boards or commissions.

The persons receiving the edited list should be asked to rate each statement on it by some sort of numerical score or adjectival check list indicating the extent to which a subject described by the statement should be regarded favorably or unfavorably. Thus each statement could be rated by checking whether it describes behavior that is "very favorable," "unfavorable," "neither unfavorable nor favorable," "favorable," or "very favorable." Alternatively, but less preferably, these adjectives could be represented by numerical scores, for example "-2," "-1," "0," "1," and "2." The ratings received by each statement in this process should be tabulated, and those rated most inconsistently, or those most often receiving ratings in the neutral category as neither clearly favorable nor unfavorable, should be deleted from the edited list. The resulting set of statements might be called a "pretest list."

The pretest list now should be printed as a Trial Rating Form. This form should have spaces to indicate the location (e.g., institution or office, and shop, class, or other unit within it) where the rating is done, the client to whom the rating applies, and the name of the rater. Each statement from the pretest list should be followed, on this form, by a "yes" or "no," or possibly by a more complex scale, such as "often," "occasionally," and "never." Instructions should indicate that the rater should circle the term after each statement that indicates the extent to which the statement describes the person being rated.

The Trial Rating Form should be sent for field trial to a considerable number of the settings for which the rating form is intended. Officials at these locations should be asked to try out the rating forms on their clients, selecting for this task staff who will be expected to use the rating form when it is final, and applying the forms to all clients who would normally be the persons the staff members would rate. Wherever possible, if two or more staff members are all considered sufficiently familiar with a given client to provide a rating for him, each staff member should be asked to rate the same subject on a separate Trial Rating Form, without consulting the other. To assure that these

directions are followed carefully, it is best to have someone from the research office supervise these field trials.

When these Trial Rating Forms have been applied in the field trials, the completed forms should be analyzed by the researchers from a number of different standpoints. To check on the reliability of items, the two or more forms for a single client that were prepared independently by different staff members should be compared with each other. This will indicate the statements on the pretest list on which independent raters most often agree when assessing the same client, and the statements on which they most often disagree. The latter, of course, are unreliable components of a performance rating. If a scale of answers is employed for each statement, rather than just "yes" and "no"—for example, if the possible entries are "often," "occasionally" and "never"—disagreements involving adjacent categories should be tabulated separately from disagreements involving more contrasting ratings. Thus, ratings of "often" by one person and "occasionally" by another, on a given subject, indicate that they both rate the subject more similarly than when one indicates "often" and another "never." On the basis of this reliability analysis, statements on which disagreement is greatest can be deleted from the form.

After reliability has been investigated, it will be well to check on selectivity. An item is considered unselective if almost every client receives the same rating on it. It may be appropriate to check on selectivity separately for each staff member who submits ratings on many persons. If some statements are always rated in the same way by raters who apply it to many subjects, the statement may be unselective even if there is diversity in the total ratings received on it; some raters may always apply the statement in one fashion (e.g., scoring everyone "yes" on it) while other raters always apply the statement in another fashion (e.g., scoring everyone "no" on it).

Sometimes, after this series of screenings, the statements left to be used for a performance rating instrument will be relatively few, but they will be those which appear to be most clearly favorable or unfavorable, reliable, and selective. A final Performance Rating Form might then consist of only one sheet. Sometimes such a sheet is employed without attempting to summarize it by a single score or grade (e.g., "outstanding," "superior," etc.). More often, however, there is a desire to consolidate the ratings on several statements into one overall assessment, or possibly into

a few assessments on different aspects of performance, each measured by a number of separate statements.

The simplest way to summarize a performance rating based on a number of separate statements is to assign a numerical score to each statement and add the numbers up for a total score. Thus a rating based on "yes" or "no" entries next to 10 statements could yield ratings of from zero to 10, indicating the number of items on which the subject was given a favorable assessment. A favorable assessment might be a "yes" on one statement, such as "Finishes his tasks before leaving them," but a "no" on another statement, such as "Gives excuses instead of taking the blame for his mistakes."

If the final form consists of numerous statements, it might be useful to reduce them to several scores for different aspects of the performance that seem to vary somewhat independently of each other. There are a number of statistical techniques for measuring the extent to which clusters of items measure the same thing, and do not measure what the items in another cluster are measuring. The most common method is factor analysis, but this has many variations, and a specialist in psychometrics should be involved in applying such procedures. In the Suziedelis study cited above, a factor analysis revealed that ratings of work performance by Federal prison inmates reflected four general factors, which were labeled:

- (1) "Good and Hard Work"
- (2) "Expressed Interest and Satisfaction in Work"
- (3) "Leadership"
- (4) "Dependency on Supervisor and Conformity"

Each of these factors was measured by a separate set of statements on which "yes" or "no" answers were circled, and the final product of the rating consisted of four scores, one for each of the above mentioned aspects of work performance. Some items measuring these factors, however, were on easily manipulated behavior, such as "Inmate asks about the salary scale of his work in the community," for the second factor. It was assumed that the form would be kept secret, a questionable assumption in a prison. If such items were eliminated, not only would the form be shorter, but the factor analysis might not have grouped the items into the same four factors.

The ultimate utility of a performance rating is its relevance to the decisions officials must make, and this is where these ratings

tie into evaluation research. If performance is to be considered a predictor of the outcome of a people-changing effort, then it is appropriate to investigate the correlation between performance ratings and subsequent behavior. It is possible that, for many programs, clients of a particular type who perform well will not persist in deviant behavior, whereas similar clients who are not in the program, or who perform poorly in it, will persist in the deviant conduct such as crime that the agency is trying to reduce. Only followup studies will reveal this.

Investigation by higher officials of the correlation of performance ratings with followup data will test the validity of the previously described judgments on whether performance rating statements in an edited list are clearly favorable or unfavorable, rather than not clearly related to the concerns of the officials regarding a client on whom they must make decisions. If some ratings are found to be irrelevant to subsequent behavior, regardless of what decisions are made about a client, there may be a desire to cease collecting these particular ratings.

More complex methodological procedures could be discussed, such as using multiple correlation methods to derive a set of diverse weights for different items when combining them into a single score, but these more technical refinements are beyond the scope of this manual's concerns. What should have been indicated by the foregoing discussion of performance ratings is the possibility of gradually improving operations records through using them as research data. This brings us to the theme of this chapter, the desirability of integrating research and operations records.

Automating Input Data for Evaluative Research

Research can guide the improvement of every type of operational report or form routinely prepared in people-changing agencies, not just the performance rating forms described in the preceding pages. Appropriate research can help to make every document of maximum utility for administrative and treatment objectives. If these potential improvements are achieved, agency records will also become more useful for research purposes. Indeed, if the record improvements for operations objectives proposed here are attained, research data superior to those now available will be gathered automatically during routine report preparation for operations needs. It is in this sense that one can describe

these ideal operations records as automating the collection of input data for evaluation research (Glaser, 1965b).

The improvements to be suggested need not be made all at once. They can be piecemeal renovations of reporting and recording procedures, each introduced only after research has demonstrated its advantages for operations objectives over the more traditional reports of records that the new routines would replace. Several distinct steps are involved in making such improvements, and these will be discussed one step at a time.

A. Content Analysis

The first step in improving a particular report is to identify the distinct types of information it contains. This problem is most critical with narrative reports, or with any sections of a report form that provide space for comments on some broad topic (e.g., "Institution Adjustment," "Family Relations," or "Post-Release Plans"). It is with such reports that modifications can achieve the greatest enhancement of utility for operations and efficiency in preparation, so that replacement of narrative reports should have first priority.

An appreciable sample of the reports should be analyzed in terms of the kinds of information they contain, and the proportion of all reports that include each kind of information. This content analysis is achieved by classifying the entries in the reports as logically as possible, preferably using simple terminologies commonly found in the reports themselves. This will produce statistics on all the topics and subtopics found in the reports and on all the kinds of information recorded on each.

To illustrate, a content analysis of the Admission Summary narrative reports at a penitentiary for adult male offenders might find that 99 percent deal with the topic "Employment History," even if a small percentage simply report "Never Employed" as their total information on this topic. Most reports, however, would include information on the subtopics "Last Job," "Longest Job," "Last Period of Unemployment," and "Longest Period of Unemployment." The kinds of information recorded under each of these subtopics would usually include year and duration. In addition, most accounts of jobs might note the type of employer, type of work subject did, and last rate of pay, while some accounts would indicate salary or status advancement on the job, and many would disclose reported reason for termination. The unemployment period information might often

include data on the sources of support while unemployed, and the standard of living maintained. A small minority of the narrative reports would provide information on additional subtopics in employment history.

The above type of analysis should be undertaken for every topic covered, such as criminal record, family upbringing, marital history, and military service. On Admission Summary narratives the task of content analysis may be so large, and the subsequent steps toward revision so time-consuming that agencies may wish to begin their efforts at report improvement by starting on shorter narrative reports, such as those for preparole investigations or the periodic "Progress Reports" prepared on inmates or probationers.

In the course of preparing these content analyses of a specific type of agency report, it might prove interesting to tabulate the frequency with which particular kinds of information are repeated within individual reports. These content analysis tabulations should cover all parts of any report, including both narrative sections and categorized short-entry sections, when both are part of the same report. Such a tabulation will aid in pinpointing any excessive duplication.

B. Preliminary Assessment of Contents

Even if nothing more than a content analysis of a large sample of an organization's standard reports were completed, the people who use these reports probably would find it interesting to learn what the tabulations reveal. Not only should the findings be presented to them, but their comments should be solicited in a systematic fashion, with respect to the following issues:

(1) Which topics, subtopics, and kinds of information tabulated in the content analysis of the reports are most useful to them?

(2) Which are least useful?

Frequently the respondents or the researchers, or both, will raise a third issue:

(3) What useful additional information should also be included in these reports?

By discussing these three issues with persons who use the reports, one gives them a stake in the redesign of the reports. By involving such persons, especially higher officials, in an assessment of the implications of the content analysis, one gives them a sense of collaboration in the researcher's effort to improve the reports. Such collaboration raises prospects that the re-

searcher's ultimate recommendations for revising the reports will be supported. Also, by addressing the three issues listed in the preceding paragraph, staff may be prompted to think about their reports objectively, as devices for accomplishing certain tasks rather than as creative writing that is an end in itself.

A number of procedures can be employed to procure more rigorous data than mere impressions on the utility of the various parts of a report. One method of gathering more objective information is to attach a log form to all of the files containing the type of report with which one is concerned, and to have staff note on these logs: (1) each time that they use the file; (2) for what purpose they use it; (3) what part of the contents of the file—even what sections of specific forms and reports—prove useful to them for this purpose. Alternatively, the log might be attached not to the file, but to the specific report in which one is interested, but this might then discourage normal use of the report, especially if the information sought is available but less adequately presented elsewhere in the file. (For an account of such logging, see Miles, 1965.)

Still another procedure for determining what information is used in decisions, devised by Leslie T. Wilkins, has been widely used for diverse types of personnel and decision in people-changing agencies (Wilkins, 1965: Appendix IV; Carter, 1967). He and his associates or proteges employ a decision simulation game which is designed by first conducting a content analysis of the information in the reports. Each topic found frequently is printed on one side or part of a card, and actual information on that topic from one report is printed on the other side or on a lower part of the card. The cards are then arranged in a file, in a rack, or on a table, so that people can only see the topic of each card—for example, "offense," "prior criminal record," "employment history," "marital status," and so forth—but not the entries about that topic for a particular individual covered by the report.

In applying the Wilkins simulation game, a group of decision-makers in an agency using the type of report to be studied are asked to imagine that they must make a particular kind of decision about a client on whom this report would be available. For example, the decision-makers could be probation officers who must decide what to recommend in presentence cases, they could be judges deciding what sentence to impose, they could be parole board members deciding whether to grant or deny parole, or

they could be institution officials deciding on the assignment to be given a newly admitted inmate after the admission classification study has been completed.

Data are procured from this game by showing the decision-makers the cards with only the topic of each card visible, then asking them on which topic they would want information first if they were to make their particular kind of decision on an individual. They are given the card for the topic they designate, and after they read its contents they are asked if they would be ready to make their decision from that information. If not, they are asked to pick the card for the topic they would wish to know about next, and after they read it, they are again asked if they have reached a decision. This sequence is continued, recording all their answers and the cards they select until they indicate that they have reached a decision. They are then asked what their decision is, and if they are certain about it. If they are not certain, they are asked on which additional topic they would desire more information. This is continued, with recording, asking them after each card whether they would still make the same decision and whether they are now certain about it. Sometimes, when they indicate certainty, they are asked to read the remaining cards anyhow, and to indicate which, if any, of the additional information cards cause them to change their mind about the decision they made, or alters their degree of certainty about it.

To assure confidence in the conclusions from these Wilkins simulation games, several sets of information cards should be prepared, each on a different client, thereby providing a variety of types of cases on which to make decisions. The clients thus depicted should be representative of the diversity of persons dealt with in the agency where decisions of the type studied are made. Of course, the correct names or other identifying information in the records on which a set of cards is based should not be used on the cards, for one should not jeopardize the confidentiality of file information. Each decision-maker asked to play this game should play it several times, if possible, using a different set of information cards each time, in order to reveal the factors influencing his decisions on a variety of types of client.

The data compiled from this game usually reveal that a majority of decision-makers make most decisions with confidence after they see only a small fraction of the information available in a report. Information cards after the first few

seldom cause them to change their minds. The cards selected before decisions are made with certainty indicate the most influential topics of information in the reports. The topics that are neglected, and that, when shown to the players, do not result in their changing their decisions, can presumably be deleted from the reports without impairing their utility to decision-makers.

Also extremely important to the assessment of the utility of information in reports are data on the correlates of past treatment decisions and on the correlates of treatment outcome. Any attribute or circumstance of clients that has been highly correlated with the decisions a people-changing agency may make, such as granting or denying parole, and with the subsequent behavior of clients, certainly should be included in the information available for making decisions. Because evaluative research identifies these correlates and consequences of decisions, those engaged in such research and those making crucial case decisions should have an interest in collaborating to improve the utility of their information.

C. Precoded Forms: Some General Observations

The content analysis, and all methods for assessment of content discussed above, complement each other. From the content analysis one sees what information is contained most frequently in the reports and what is included only infrequently and haphazardly. One may also become aware of much duplicated information cluttering the files. From discussions with staff and higher officials using the reports, one learns what they are most interested in having these documents contain. If the actual use of the reports can be logged, or if simulation games based on the topics in the reports can be developed and played, one can gain additional evidence on what information in the reports is actually used in the agency's operations. If there has been relevant followup research, one may also learn what types of information provide the most valid guidance for the agency's case decisions.

In undertaking to improve a given type of report in a people-changing agency, one should always be able to procure, minimally, a content analysis of past entries in these reports, and staff assessment of the contents. From this, with or without logs of use, simulation game data, or knowledge of outcome correlates, one should be able to replace narrative reports by precoded forms.

These forms would contain spaces for entering information on all of the topics and subtopics which the content analysis and its assessments indicate are useful enough to procure for every client to whom they are applicable. For example, in an agency for adult offenders one might find space on a precoded Admission Summary for Employment History, with spaces under it to enter "None," but if this is not entered, to record information on Longest Job and Last Job, according to our discussion in the Content Analysis section of this chapter, the Admission Summary might also have space for some less frequent items such as Military Service, including branch of service, dates, rank, and type of discharge, if the content analysis and assessments concluded that information on these matters would be useful.

A precoded form thus contains topical headings and phrases under each topic, so that staff members, instead of preparing a narrative report, simply check the appropriate items or write short entries to provide the requested information in the indicated spaces. Such forms are precoded in the sense that the topics they cover and the language they use is similar from one case to the next. The forms may also be precoded in the sense that they have marginal spaces or numbering of the items to permit efficient transfer of their information to computer cards, tapes, or discs for statistical tabulation.

It is not suggested here that precoded forms should completely replace all the narrative content in the casework reports of people-changing agencies. What is suggested is that standard types of information in standard types of reports should be reported merely by checking the appropriate items on a precoded form. Anything not standard would still have to be reported by narrative phrases or sentences, or by short paragraphs.

The advantage of precoding is that it eliminates most of the deficiencies ascribed in the early sections of this chapter to both records now used for research purposes and reports now used for operations purposes. For both research and operations, precoding contributes to efficiency because it results in standard types of information being recorded on all cases in a standard manner and in a standard location. For research it adds greatly to the validity and efficiency of information collection because the entries are made in the records by operations people in the course of their routine work, and on forms that they and their colleagues and superiors will have to refer to again. This reuse

of the forms in operations motivates those completing these forms to make the entries accurately, and increases the likelihood that should errors be made someone will find them and correct them.

For operations purposes, precoding has the advantage of getting field treatment staff out of the office and in greater contact with their clientele, or performing other functions more useful to the organization than literary production. Precoding may reduce the number of clerks or secretaries needed. Precoding may also make reports more valid. It is somewhat harder to slant the information in a precoded report to justify a particular recommendation or action than it is to slant a narrative report so that it emphasizes and dramatizes what one wishes will be inferred, and omits or distracts attention from information supporting an alternative conclusion. Precoding can thus reduce the extent to which reports serve as rationalizations and excuses for decisions, and may result in decision-makers, being more deliberate and responsible in their actions.

Finally, precoding helps to routinize evaluative research. This routinization comes mainly from the fact that precoding makes collection of many of the data needed in evaluative research automatic as a part of operations requirements. Precoding of the background and diagnostic information collected for operations purposes when a client is first received by an agency, and precoding of the logs and performance ratings prepared for operations purposes during the course of treatment programs, provides evaluative research with most of the input data it requires. If precoding is also done for progress or violation reports on long-term community surveillance (as in extended probation, parole, and aftercare), then these reports may provide evaluative researchers with all of their needed outcome data.

Precoding also helps to routinize evaluative research in another way. When input or output data for program evaluation are derived from operations records, the reports from such evaluations will employ the language and the categories with which operations staff are familiar. Terminology may then be largely identical in evaluation and operational reports, and such reports may often be combined. Such a development should certainly foster more communication between researchers and operations officials, and more use of research findings.

D. Initial Versions of Precoded Forms: An Example

When conversion from narrative to precoded forms is initiated, considerable resistance can be expected. Persons who

devote themselves primarily to preparing narrative accounts and assessments of clientele and who take pride in these documents will object to the notion that their creations can be replaced by what they are likely to label derogatorily as "a bureaucratic form." They will contend that the unique features of every case, and the dynamics of personality development and response to treatment, cannot be captured in a few categories.

Persons using narrative reports as their source of information on which to base decisions, in contrast to persons who prepare the reports, are likely to have mixed reactions to suggestions that the reports be precoded. On the one hand, they will admit that when they wish to look up some detail about a client, such as employment history or military record, it is often hard to find it in the narrative report, and they are not sure about the amount of information that will be there when they do find it. On the other hand, they will say that it is good to have some sort of narrative summing up of the principal justifications for a recommendation, rather than having to infer them from a sheet of atomized separate bits of information.

It should be remembered that the persons who prepare and use the narrative reports will usually be the ones asked to prepare and use new precoded forms. Therefore, it is important for acceptance and efficient utilization of the new forms that these people be involved in planning the changeover, that the change be preceded by much trial and revision of draft forms to minimize objections to any of their details, and that compromises be made on any aspects of proposed forms to which there is strong resistance. Since precoding can be expected to reduce the work of those who prepare reports, a proposal to develop precoding will be accepted most readily when there is a pileup of paperwork and delay in getting reports completed in time for required decisions.

Such circumstances existed and all of the consultation, trial, and compromise recommended above were stressed in a successful conversion from largely narrative to highly precoded forms that I initiated in 1969 as Associate Commissioner for Research of the New York State Narcotic Addiction Control Commission. The conversion was completed and successfully institutionalized by my successors after I left this agency in January 1970. The forms, and the kinds of experience associated with developing them and getting them accepted, provide illustrations that may be useful to persons in other agencies who wish to revise their

reporting procedures to improve both operations and research.

The first reports we sought to improve were the narrative accounts of the experience and performance of addicts in the Commission's confinement institutions. These reports, prepared by caseworkers at the institutions, were used by members of the Commission primarily in deciding whether to release an inmate to aftercare in the community, or to transfer an inmate from one institution to another. The reports were also used by superintendents and other officials of the institutions in deciding on the transfer of an inmate from one program to another within the institution, in reacting to the misconduct record of an inmate, in considering inmate requests for furloughs, and in many other decisions.

The reports described above were to be prepared every 3 months or whenever the Commission was requested to consider release or transfer of an inmate, whichever came sooner. One reason for our effort to improve these reports was a lag in their preparation; they were often not available until much more than 3 months had elapsed, and the Commission's deliberations on a client frequently had to be delayed because reports were not ready. Also, there was considerable dissatisfaction with the content of many reports. This was ascribed largely to the fact that the agency had grown rapidly and many of its staff were inexperienced.

Form revision efforts began with a content analysis of these reports. This was followed by the research staff's discussing the content analysis findings with institution caseworkers and supervisors. Precoded report forms were then drafted and applied to a few inmates at Commission institutions by a team of persons from both the institution and the research staffs. This teamwork bridged some social and cultural gaps between these two staff groups, thus helping to legitimize collaboration and interest in this project at the institution.

One of the emphases I gave to the initial drafting of these precoded forms was always to provide caseworkers with an opportunity to note on the form any information they thought important, but not covered in the precoded categories. Most sets of categories ended with "Other," to be checked when the caseworker thought the client did not fit any of the more specific categories. There also was a space to specify what was being referred to as "Other." Also, almost every page of the form had a space at the bottom for "Comments." This space was to be used whenever some subtopic not covered in the form was deemed

worth mentioning in a particular case, or whenever the caseworker who checked a category on the form felt it necessary to qualify such an entry, for example, to indicate why there could be some question about the appropriateness of this category for the particular client.

In our staff's trial of their draft forms, they frequently made reliability checks to see if two or more of them agreed in the entries they made independently about one client. They prepared instructions for using the forms in hopes that this would foster reliability and validity, although from my experience I stressed that those who routinely use forms will not long check details in the instructions, so that the desired interpretation of the form should be evident without special directives.

When our staff were fairly well satisfied with their revised drafts of a precoded form, we arranged to have several caseworkers apply the forms to inmates in their regular caseloads at the institutions. When their entries in the "Other" categories or in the spaces for comments were frequent, our staff revised the precoded categories to include entries for the most common special circumstances cited, or to eliminate ambiguities which resulted in a preference for the "Other" category or a need to add comments.

When the revised forms were then shown to still higher officials, especially to one fairly new to the Commission, some suggestions were made calling for additional information in the form that we and many caseworkers whom we consulted thought could not be procured with any validity. These were mainly items about what were called "psychosexual factors," such as "sublimation of sexual energies." Indeed, the psychiatric orientation of the senior officials and of many of the casework and clinical staff resulted in the reports including many items that we researchers, mainly sociologists with a background in criminological research, thought would be unreliable, invalid, or irrelevant. We included these items on the precoded form anyhow, first because the forms had to convey what the preparers and users of the reports thought would be important, and second because we anticipated that our subsequent research would demonstrate which entries were relevant to outcome and which were not.

The precoded "Intramural Case Summary" form resulting from these efforts was submitted to the Commission for approval around the time of my departure in January 1970, but was not adopted until March of that year. It is reproduced as Appendix A

to this volume to illustrate the compromises between narrative and precoded format that often are necessary in the initial changeover from one to the other. This form was reported to be a great timesaver in both preparation and use when compared with the narrative reports which it replaced in most cases, and it also provided more of the information officials considered important. Its sections for "Comments on Above," at the bottom of each sheet, were seldom used, but they provided a safety valve whenever anyone preparing a report was dissatisfied with the categories the form provided.

Changes in the Commission's policies in 1970-72 made this form quickly obsolete, primarily because the average time of initial confinement for its civilly committed addicts dropped from 12 to about 5 months. Most of the 47,000 clients in the programs it operates, funds or certifies are now in methadone therapy programs in the community. This form is now an optional summary record form in the institutions, its entries are not tabulated, and it is used somewhat differently from one institution to the next. Nevertheless, the effort to create it started a movement that swept through the organization, for reasons I shall attempt to elucidate.

E. Revised Versions of Precoded Forms: Some Examples

Once the movement to replace narrative reports by precoded forms had been in progress for a few months and was well known to many officials in the organization (even before the adoption of the long form in Appendix A was finally negotiated), the Division of Research was repeatedly consulted by officials in various other specialized units of the Commission for assistance in the replacement of their narrative report procedures. For example, the Commission initially required from its aftercare staff a thorough investigation and a narrative report on an inmate's proposed postrelease living arrangements and employment before the inmate could be considered for release on aftercare. When hundreds of its civilly committed addicts suddenly became eligible for such consideration during 1969, the aftercare staff had much difficulty keeping up with the demand for these reports, and this was held partly to blame for the Commission confining its addicts longer than had initially been anticipated. Accordingly, the head of the aftercare service, Meyer H. Diskind, recruited staff of the Division of Research to work with aftercare officers in developing a largely precoded investigation report, with only a short narrative section as its conclusion. The officers

were delighted with what they called "the new checkoff forms." It helped them greatly in catching up in their work and improved the clarity and efficiency with which their information was communicated and utilized. Similarly favorable reactions were reported in other parts of the organization where we assisted in the replacement of narrative documents by largely precoded forms.

As indicated, the first precoded forms adopted by an agency in which staff are accustomed to narrative reports are likely to be compromises. They will often include many items that fascinate the more influential members of the clinically oriented treatment staff, who are impressed with the insights they presume that their narrative reports transmit.

There may well be reason to doubt in advance the reliability and validity of these items for most cases. Indeed, if extensive field trials are undertaken, with pretest forms filled out on one client independently by several caseworkers, the unreliability of some items may be demonstrated before a form is adopted so that concensus can be reached on eliminating such items. For two major reasons, however, it is wisest to compromise, if possible, when some influential officials or staff have a strong attachment to certain items, and to include on initial forms enough items to please every major faction.

One reason for compromising somewhat on the inclusion of items on a precoded form is that strong initial dissatisfaction with a form will impair its careful use. It is preferable to include all topics fervently desired by influential persons preparing or using a form than to risk the prospect that they will immediately encourage replacement of the standard form with their own variations independently in many cases, thus severely impairing the use of the forms for research. Secondly, precoding of forms soon generates considerable pressure by staff to revise or delete sections they cannot fill out with confidence. Revisions difficult to make before the precoded form is adopted become easily accepted after the form is used for a while, for reasons which can readily be elaborated.

Once a narrative report has been replaced by a precoded form there is a great reduction in the sense of literary creativity that staff feel in report writing. They may still be involved in some writing for narrative sections of the largely precoded report forms, but this is much more limited in length and content than the traditional narrative reports. When every caseworker must classify every case by the same set of subjective impressions,

because the favorite subjective categories of a few of them have been included in a standardized precoded form, the reliability and relevance of the more speculative items are likely to be widely questioned, even by the original proponents of the items. Once staff do not have paragraphs and pages to regard as their clinical achievements, they develop more interest in other activities of their job. There will then be pressure by staff to make the reporting system even more efficient.

The task of researchers in facilitating the revision of a precoded form is much simpler than their work in the conversion from narrative to precoded forms, but they are logically the same in function. Content analysis is the first step, but with precoding this is simply a matter of listing all categories and tabulating the frequency with which each is used. If forms have been designed for electronic data processing, this listing and tabulation can be done by a computer, and may already be available in routine statistical compilations by the agency. Inspection of these tabulations will reveal what categories are rarely used, and on what items there is extremely frequent use of an "Other" category.

Examination of a sample of forms from the files will indicate the frequency with which space for comments is employed. A separate content analysis and tabulation of topics covered in these comments may suggest need for revising the precoded items. The computer can also print a list of all cases on which "Other" is checked under any topic, and the forms from a sample of these cases can be examined to determine what is specified under "Other." Finally, if short narrative sections remain in the largely precoded form, it may be of interest to subject them to content analysis, as was done with the more predominantly narrative report that preceded the precoded form.

Discussion of these content analysis findings with the staff who prepare the precoded forms and with the officials who use the forms should again be fruitful. Indeed, staff at many levels are likely to have already voiced strong opinions that certain sections of the form are useless, and they may have instituted deletions or other revisions even before a content analysis is provided. If the researchers have been alert they will already have undertaken some logging of the use various parts of the forms receive, they will have conducted new reliability tests, and they will have expanded knowledge of the correlates of entries on the form.

By 1972 the forms in Appendix B were employed by the

Commission, several of them having gone through revisions twice after the narrative forms were largely replaced by precoded forms in 1970. The first page of Appendix B is a form for monitoring employment on aftercare, which presents one of the most salient indexes of rehabilitation as far as societal costs are concerned, and is thus a boon to cost-benefit analysis. Fingerprint arrest reports on persons under Commission supervision are supplied automatically, as received, by NYSIIS (described in Chapter 7). The second and third sheets of Appendix B are the two sides of an admission summary report, entitled "Report of Admission to Treatment and/or Service Program." The fourth and fifth pages of Appendix B are both sides of the "Aftercare Chronological Summary." This provides in precoded form what is believed to be the most relevant, valid, and reliable information for aftercare decisions. It can also be invaluable as a data source for evaluative research to determine, from the followup information it provides, what Commission programs are most effective for various types of addict.

The examples presented in Appendixes A and B were selected because I was familiar with their evolution and could describe the principles and processes involved in this development. Doubtless there are better forms, and further application of the principles presented here could improve these forms. For the forms in Appendix B the computer prints the entries on the first few lines and staff enter the remainder, which can then all be stored by the computer and be retrieved in almost any variety of desired tabulation. This reflects another trend, even more developed in other people-changing agencies, which merits the concluding observations for this chapter.

F. Computerized Record Systems

The integration of operations and research records is accelerating in people-changing agencies because of use of computers for information storage and retrieval. It has frequently been possible to receive grants or appropriations to purchase computer services, but they are often used without appreciation of their potential for program evaluation and decision guidance. This often reflects the direction of the computerized record system by persons more experienced at business and industrial applications of computers than in designing and administering evaluation research for treatment agencies.

One of the problems to avoid in computerized information systems is an excess of unimportant input information. Just

because the computer can process and analyze an immense amount of data, some agencies have propagated a large number of excessively detailed forms on which staff record a tremendous amount of information on their clients, and especially on their own activities. This type of data collection is appropriate only for special studies; when it becomes routinized it creates a paperwork overload that is seriously burdensome and results in entries being made carelessly, indifferently, or even with deliberate misrepresentation.

When information forms are introduced for use with a computer, they should be designed to replace and improve upon the traditional forms in the agency, rather than to supplement them. The logical steps needed for this are the same as those already described for developing or revising precoded forms. One should begin with content analysis of current forms and assessment of their use, then proceed with new forms only in consultation with staff and with extensive preliminary trials. Unfortunately, this has not always been done. Indeed, at this writing there are, to my knowledge, no computerized information systems in people-changing agencies that have fully integrated operations record-keeping with program evaluation, but several are developing in that direction.

One such development, still in progress, is the RAPS-2 system of the Federal Bureau of Prisons. Initially instituted on a trial basis in only three of the Bureau's approximately 30 facilities, this program links computer terminals at separate prisons with the file room serving the Bureau and the U.S. Board of Parole in Washington, D.C. The terminals at the institutions have a typewriter keyboard and a television screen; there is no paper in the typewriter, but the operator sees the typing printed on the screen as though on a paper form. The terminal is programmed so that the entries will be for whatever numbered square of a precoded report form the operator designates. Actually, the terminal is also punching holes in a teletype tape, and the operator can run this through a machine to print on forms for use at his facility. The tape can also be run through a data phone connected to the Washington terminal to print a copy on a form there, or to store it in the computer for retrieval whenever it is needed. Ultimately it will be linked to the FBI's computerized records for followup data, and thus permit its use in evaluation research.

Appendix C consists of two of the forms in this RAPS-2 system of the Bureau. The first page, entitled "Program Analysis Sheet," contains the essentials of an Admission Classification

Summary. The first two lines contain routine administrative identification and scheduling information. In the columns below that "Rating" classifies the inmate into three categories according to his presumed prospects for change; "Age," "Prior Commitment," "Sentence," and "Custody" have self-evident information; the three levels under the heading "Priority" refer to frequency of program review, as 3 months, 6 months, or a year.

The 10 lines numbered 12 through 21, in the lower half of this Program Analysis Sheet in Appendix C, are each for a different aspect of the Bureau's potential efforts to change an inmate. These aspects are indicated in the column under the heading "Correctional Factors." They range from improving his "Environment" by changing his "Economic Status" or his "Family Conditions," to improving his "Character Traits" with respect to "Self-Control," "Interpersonal Relations," "Standards and Values," or "Aspirations." In the column headed "Need Level," as shown, entries are made to indicate whether the inmate's need is "High," "Low," or "None," for change in each of the Correctional Factors. Entries are made here whether or not anything can be done to fill the need, if fulfillment of such need is believed to be relevant to the individual's rehabilitation. Thus, if the inmate is poor the entry in the Need Level column of the Economic Status line may be "2" for "High"; if he is quite deliberate and careful in his behavior, the entry in this column on the Self-Control line may be "0" for "None."

Whenever the entry in the Need Level column is other than None, the remaining columns are supposed to indicate the "Activities Planned" to meet the inmate's indicated need. For each planned activity there may be an entry in the adjacent column, on the "Constraints Preventing Plan." Thus, "Time Too Short" or "Lack Program" may be entered next to an entry in the Activities Planned column for "On-the-Job Training." The constraint of insufficient motivation, that the inmate simply does not want the activity prescribed for him, presumably comes under "Other."

The second page of Appendix C contains the form headed "Inmate Activity Report," which is essentially a summary of progress made toward achieving the plan covered in the form already described. Lines 1 through 7 of the Inmate Activity Report identify the inmate and indicate what program this sheet covers; a separate copy of the form is used for every program that an inmate enters or completes. Lines 8 through 9 contain data gathered when he enters a program, and entries are made

on the additional lines if he completes the program or if he withdraws from it.

With these two forms retrievable from the computer, any officials at the institution or in Washington can be informed immediately of the progress of an inmate with respect to the correctional needs projected for him, and the additional activities which the inmate should be encouraged to pursue. The computer is programmed to print a narrative summary of this information for the Parole Board Examiner before he interviews an inmate. Below this it can print the Examiner's summary of interview impressions, for presentation to the Parole Board.

From the standpoint of evaluation, the potential of these forms is also tremendous. In the first place, they can, of course, provide data on the activities completed by inmates before their release. They can relate these activities to the presumed correctional needs of the inmates, to activities planned for them, to the constraints on these plans, and to the age, prior commitments, and sentences of the inmates. Finally, when this is linked to the FBI computers, they can relate this information to the subsequent criminal records of the inmates.

The RAPS-2 sheets in Appendix C consist mainly of objective items, but are subjective in their "Rating" of the inmate with respect to prospect for change, and in their estimations of his "Need Level" under each of the ten "Correctional Factors." These subjective estimations are those which correctional administrators must make, however, if they are to decide rationally on the rehabilitative programs they should provide for an inmate. By making these judgments explicit and in the record, this record system creates a pressure for assignment of inmates to activities on the basis of their presumed correctional needs rather than on the basis of administrative convenience, as is often customary. By linking this record with data on activities actually provided, and with data on subsequent criminality, a foundation has been created for assessing both how accurately subjective judgments are followed in practice, and how relevant this is to subsequent criminality. The forms provide an invaluable input for evaluative research, an input acquired automatically and routinely in the course of maintaining records for operations requirements.

Conclusion

Because operations records in people-changing agencies so often are incomplete, unstandardized, and disordered, they are

deficient as sources of research information about clients and about the treatment that clients receive. Most special data-gathering for research, however, also has serious shortcomings: if it is done by operations personnel for researchers, difficulties tend to develop in trying to maintain adequate control and reinforcement of data-gatherers to assure their doing careful work; if the data are gathered by research employees at operations sites, the task becomes expensive and it is prone to error from the unfamiliarity of the research personnel with the persons and settings where data are collected.

Also, operations records often impede the attainment of operations objectives. For reasons that this chapter elucidates, these shortcomings of operations records are most notable in narrative reports, which frequently are more a liability than an asset in endeavors to increase the effectiveness of people-changing agencies. Also grossly inadequate, in many instances, are the records for collecting line staff's assessments of the clients with whom they work.

This chapter and the manual's appendixes provide extensive descriptions and illustrations of procedures for replacing narrative reports by precoded forms, and for improving performance ratings by line staff. These procedures can enhance the quality of information available for both operations and research needs. Efficient use of computers for records storage and retrieval also requires the suggested types of improvement in recording information. Therefore, the growing use of electronic data processing may greatly hasten the integration of operations and research records, expanding the usefulness of both.

CHAPTER 9

COMBINING INPUT DATA: FROM SCORE SYSTEMS TO WORD SYSTEMS

While the preceding chapter discussed the switch from word systems to number systems in record-keeping, this chapter, for a different problem, discusses movement in the opposite direction, from numbers to words for the summation of statistical information.

In people-changing agencies, and in much of this book, we often speak of persons and programs as though they each had only one or two dimensions. Thus one speaks of burglars or of unaddicted burglars, and of their being in vocational training programs or in vocational training with monetary incentives. The burglars may also be young or old, regularly or never employed previously, first offenders or recidivists, married or single, and have numerous other characteristics. Similarly, vocational training can be described as trade studied, skill attained, and many other attributes.

There are several reasons why our generalizations use so few dimensions, accenting some, neglecting others. In the first place, our language and our minds seem able to conceptualize readily no more than three or four dimensions at a time—some people have trouble after two! Secondly, if statistical comparisons are sought, an enormous sample is needed to find subgroups all clearly differentiated by many variables, since the number of possible combinations of dimensions becomes immense with even a few variables. Thirdly, administrative or policy decisions are usually concerned with evaluating only one program or one aspect of a program at a time, and for only one broad category of clientele or for a limited number of categories.

In trying to improve reports and records as input data for evaluation studies, the researcher is frequently deluged with information. The most difficult problem in many agencies is not collection of input data, but reduction of these data. Indeed,

science has been called a kind of shorthand; its task is to reduce an enormous range and complexity of observation to the most useful brief formulations for representing what is most important in the complexity.

There are several methods of reducing the number of input variables in evaluative research on people-changing agencies. The approach discussed in Chapter 6—"Who should be compared?"—is simply to focus on one or two variables at a time when they happen to be problematic—usually the treatment variables—and then to select comparison or control groups presumed to be similar on everything except the variables that concern us. Another procedure is to combine a large variety of information by some system of multivariate weighting, so that one ends up describing subjects and possibly the programs they were in and their performance there as well, all by only one variable, a score. This single score then represents many variables. A third method is one that might precede and lay the groundwork for the first, analyzing subjects and programs with respect to many possible combinations of numerous variables to determine the few combinations that will be most discriminating. This chapter will discuss the last two methods, which will be called, respectively, multivariate scoring and configuration analysis.

Multivariate Scoring

In corrections, the effort to guide decision-makers through scientific research was spearheaded by prediction studies. These attempted to classify offenders into categories that, according to statistics from past experience, have markedly different rates of parole or probation violation, or of recidivism (for histories of these studies, see Manheim and Wilkins, 1955; and Simon, 1971). Evaluation research, as advocated here, goes one step beyond this to determine which types of clientele have more favorable subsequent behavior rates after one type of people-changing program rather than another.

In the earliest parole prediction studies, each category of every variable at all related to outcome was given a "success score" of one or zero according to whether those classifiable in that category had violation rates below or above the average. For example, scores by this method could be tallied as follows: a prisoner with no felony convictions before the one for which he currently is confined receives a score of one, but an inmate

with one or more prior convictions receives a score of zero; if married or widowed the adult prisoner receives another "one," but if single or divorced the prisoner receives a score of zero; from a steady employment record he receives a score of one but with only sporadic and irregular prior employment he receives a score of zero. On these three variables alone, inmates who are married first offenders with steady work records receive a success score of three, and inmates who are divorced recidivists with only sporadic prior employment receive a success score of zero. Inmates with any other combinations of the three variables described receive scores of one or two, depending on the number in which they have the attributes with below average violation rates.

Parole prediction scores at first were based on 20 to 30 variables, so the highest success scores might be over 20; the lowest possible score was, of course, zero. It was discovered that scores derived from many variables differentiated inmates into groups with much more contrasting violation rates than the rates for categories of any isolated variable, such as prior convictions or marital status, taken alone. Thus, in the pioneer study by Burgess with 21 variables, those with success scores of 16 to 21 had a violation rate of only 1.5 percent, while those with scores of four or less had a violation rate of 76 percent (Bruce, *et al*, 1928: 248).

Despite that impressive contrast between extreme score groups, when many factors were used in scoring, most cases fell into score groups near the violation rate of the total sample, which is the probability of violation for the average parolee. Furthermore, when the studies were repeated, the violation rates for the most extreme score groups tended to change and become closer to the rate for the total sample than they were in the original study. This shift is a consequence of regression effects, discussed in Chapter 6.

Subsequent studies showed that if one bases the scoring only on a few of the most differentiating factors which are relatively independent of each other, and which show consistency in their violation rates from one study to the next, the resulting scores based on a few factors will usually have several advantages over scores based on many factors. The scores based on the highly selected few factors may not have quite as extreme contrasts in violation rate between their lowest and their highest categories, but they may have fewer cases in categories with close to the total sample's violation rate. Their categories will also shift

less in relative violation rates from one study to the next (Ohlin, 1951; Simon, 1971).

Some critics of the pioneer studies objected to use of the same weights, either one or zero, for all variables. Thus, having attributes with very low or very high violation rates, such as no prior convictions versus 10 prior convictions, contributed no more to an individual's total score than factors with categories all only slightly above or slightly below the sample's violation rate, for example, IQ. Accordingly, scoring systems were devised with a larger range of weights, so that a subject's score from any specific variable depended on how different the violation rate of the subject's category on that variable was from the violation rate of the total sample. In systems using a large number of variables, however, this did not result in a range of violation rates for total score groups markedly different from the range found with the simpler weighing system.

A more important criticism was that none of these additive multivariate scores takes into account the interaction among variables. For example, criminal record, employment record, and information on addiction to alcohol or narcotics, each taken separately, can be used to classify offenders into groups with markedly contrasting recidivism or violation rates, e.g., first offenders versus recidivists, steady versus irregular workers, or addicts versus nonaddicts. No method of combining information from two or three of these three factors, however will nearly double or triple the extent to which offenders are differentiated into groups with very contrasting violation rates, as compared with the contrast achieved by classifying them on any one of the three variables alone. This is because most of those with a lengthy criminal record will also be in the group with a poor work record, and probably will include a high proportion of those who are addicted. Therefore, while it is appropriate to give a heavy weight to the first of any of these highly differentiating variables by which inmates are classified, less additional weight should be added to the score for each additional variable by which they are further classified.

The foregoing criticism, on the implications of interaction among variables, was answered by use of the discriminant function system of multiple correlation analysis. This rather complex statistical calculation method—now simpler to do because of computers—determines an optimal sequence of diminishing weights for different variables. It also indicates when use of more variables would not add significantly greater differentiation

in violation rates among the various score groups. For example, in a pioneer analysis by this method, Mannheim and Wilkins (1955: 145) produced the following scoring system for classifying inmates of British youth prisons (Borstals) into categories with markedly different "failure rates" on parole:

If evidence of drunkenness, add	24
If any prior offense resulted in fine, add	9
If any prior offense resulted in commitment to prison or training school ("Approved School"), add	8
If home is in an industrial area, add	8
If not living with parent or parents, add	7.5
If any prior offense resulted in probation, add	4

Add an additional weight for the longest period in any one job, weights provided in a scale which runs from zero if longest job is over 18 months to 11.7 if longest job is less than one month.

Although their sample had a failure rate of 43 percent, those with scores of over 40 by the above system had an 87 percent failure, those with scores of 24 to 39.9 had a 66 percent failure rate, those with scores of 15 to 23.9 had a 40 percent failure rate, those with scores of 10 to 14.9 had a 33 percent failure rate, and those with scores of less than 10 had only a 13 percent failure rate.

In 1958, when California quickly moved to the leadership it still retains in correctional research, it brought Leslie Wilkins from Britain as a consultant and there he initiated a distinctive application of discriminant function analysis called the Base Expectancy Method. The first step was to abstract all the information in the files with which to categorize parolees from California correctional institutions, *as of the time they began their institution terms*. This was done separately for males and females, and separately for the Department of Correction cases (adults) and for those committed to the custody of the Youth Authority. No further separation of cases were made, so that each group consisted of all Department of Corrections or all Youth Authority cases of their sex in California paroled in a particular year, except that the Youth Authority at first separated persons on their first parole from those reparaoled after a prior parole violation. From multiple correlation analysis of the file information for each group, a scoring system was calculated (called a "scoring equation") analogous to that illustrated above for the British Borstal cases, and the parole violation rates for each score category were determined.

An example of such a scoring equation is the following, for male Youth Authority wards on their first parole (adapted from Molof, 1967: 36) :

Add a weight for age at first admission, weights provided in a scale which varies by age, from:	
19 years or more, add	384
to	
13 years or less, add	0
If no court commitments prior to commitment to the Youth Authority, add	96
If no known school misbehavior, add	67
If no escapes from incarceration prior to commitment to the Youth Authority, add	66
If one or more offense partners in offense for which committed to the Youth Authority, add	32

The total score an individual offender receives from a table such as the above is known in the California correctional agencies as the "base expectancy score" (usually abbreviated as "B.E. Score"), and the score groups are referred to as "base expectancy categories" (or "B.E. Categories"). The following are the base expectancy categories from the scoring equation above, and the recidivism rates of Youth Authority wards in each of these categories released in 1964:

B.E. Category 1, scores over 545	22% recidivists
B.E. Category 2, scores 419 to 545	35% recidivists
B.E. Category 3, scores 337 to 418	41% recidivists
B.E. Category 4, scores 290 to 336	47% recidivists
B.E. Category 5, scores 195 to 289	56% recidivists
B.E. Category 6, scores below 195	68% recidivists

The "base expectancy" concept was developed as a method of overcoming the difficulty of getting correctional officials to assign offenders to different programs on a random basis. Because each type of institution, and each program within an institution, receives a different mixture of offenders, it is not known whether a difference in violation rate of releasees from one program compared with releasees from another is due to differences in the effects of the programs or to the different selection of offenders each receives.

The term "base expectancy rate," applied to a group of offenders, refers to their expected violation rate when they are first admitted. For this reason, all items of information used to calculate an optimal prediction scoring system for them is restricted to what is known at the time of their admission; deliberately omitted is the additional information available such

as their assignments, performance, escapes, and family communications while confined. If those *in a given base expectancy category* actually have a lower recidivism rate after parole from a particular institution or program than the predicted rate for their category from all institutions and programs in California, then the particular institution or program they were in would be credited with reducing recidivism rates. Thus some programs from which releasees had high recidivism rates might be regarded favorably because their *expected* recidivism rates, according to the B.E. categories of the inmates they received, were even higher than their actual rates. Conversely, low recidivism rates from another program would be unimpressive if it were shown that this program had only cases from whose base expectancies a low recidivism rate would be expected in any program.

In an early application of this system, it was shown that eight different types of facility housing Youth Authority cases during the mid-1950's had parole violation rates ranging from a low of 36 percent for forestry camps to a high of 59 percent for the Nelles School for Boys. The numbers in each base expectancy category among the parolees from each of these facilities were determined, however, and the numbers were multiplied by the expected violation rates for that category, from a base expectancy table for Youth Authority parolees from this period. This yielded an expected violation rate almost exactly the observed rate at five of the eight types of facilities (Beverly, 1961).

The observed rates at the remaining three types of facilities differed from the expected rates. At the forestry camps the rates were lower than expected while at Nelles the violation rates were higher than expected. This suggests what might be called a "group norm effect," evident in many types of people-changing organizations including ordinary public schools. The reason for this is that if individuals of a given performance potential are placed in a group with a high average level of achievement their performance will be better than expected, but if placed in a group with low average performance their achievement will be less than expected.

The third group which had unexpected rates were those not sent to any facilities, but paroled—usually within a month—directly from the two reception centers (one for the northern and one for the southern half of the State) to which youths are delivered from the courts for transfer to other Youth Authority institutions. The violation rate for this group was lower than

that predicted from their base expectancy categories, which suggests that direct release was more favorable for them than any kind of institutionalization.

Interesting additional findings were on the selective impact of various types of facilities on specific base expectancy categories. Wards of the Youth Authority, who by base expectancy rates were good risks, had a significantly higher violation rate than expected if kept in county jails instead of being transferred to the reception centers. Almost all of the better-than-expected outcome for wards paroled from the forestry camps after being transferred there from more traditional correctional institutions was accounted for by the better-than-expected performance of poor risks. This again suggests a group norm effect, or possibly some unusually positive qualities for boys with bad risk attributes which officials might have noted before going against usual odds by sending these boys to forestry camp. Wards of all base expectancy categories who had been transferred directly from the reception centers to the forestry camps had a lower-than-expected violation rate, but this was true especially of the good risks.

The simulation of a controlled experiment made possible by this base expectancy analysis was most impressive. The Youth Authority, however, decided to make a more rigorous test of the impact of forestry camps by randomly selecting only a fraction of the boys for the camps from those deemed eligible for camp assignment. Those chosen became an experimental group while the remainder served as a control group in the institutions. There was no appreciable difference in the subsequent parole violation rates of these two groups, however, each group having about the rate which could have been predicted for them by the base expectancies of their members (Molof, 1967).

While the base expectancy rates still are applied in California as a check on interpretations of outcome, there appears to be much less enthusiasm for them now than there was originally, and they are no longer considered substitutes for experimentation or for other types of comparison grouping. Among the reasons for deemphasis, I believe, were the following:

1. In correctional systems with highly diversified facilities, such as the California youth and adult systems, the distribution of offenders among facilities was never random. Those having traits associated with high violation rates and those having traits associated with low violation rates were already predominantly in different facilities in the samples from which base expectancy

rates were calculated, so that any effects of the facilities on their rates were already largely reflected in the scores assigned to different traits in the scoring equation. The base expectancy scores thereby have diminished utility for assessing the impact of different facilities or of programs, although they may still be useful, especially for assessing the impact of new activities for randomly or somewhat haphazardly selected persons. Thus, with group counseling and with small caseloads on parole, middle base expectancy adult males had somewhat better-than-predicted parole outcomes; their violation rates were still intermediate between those of the best and the worst risk categories, but the latter had the expected rates with or without group counseling, and with small or regular parole caseloads (Harrison and Mueller, 1964; Havel and Sulka, 1962). The Kassebaum, *et al* (1971) controlled experiment with group counseling, which found it had no effect on recidivism, throws considerable doubt on these base-expectancy findings, just as the forestry camp experiment described above dampened the enthusiasm for base expectancy assessment of facilities for youth offenders.

2. All numerical scores calculated from diverse variables, whether by the crude adding of "one" or "zero" for each factor or by sophisticated multiple correlation equations, confound the many sociocultural, psychological, or other explanatory theories that determined the choice of variables to investigate as possibly accounting for differences in outcome rates. A number of quite different combinations of variables may place an offender in a specific score category, especially in the middle risk categories, so that users of the score group data cannot readily know *why* some offenders are classified in one risk group rather than another. This is especially confounded with multiple correlation systems, because laymen cannot understand how the scoring instructions were derived and hence why they add so much more weight for some variables than for others. Policymakers, who must be able to justify their actions if called upon to do so by the governor, legislators, journalists, or others, are understandably reluctant to rely on a system of explanation for outcomes they cannot fully comprehend.

Configuration Analysis

In working with Federal correctional officials during 1958-62, and striving to have an impact on practice, I sought to promote statistical prediction considerations in parole decisions, having

been involved in Illinois parole prediction research in 1950-54. Tremendous resistance to actuarial prediction tables as a source of advice on risks of recidivism was widespread among parole board members—and I met most members of most boards in the United States during 1961-65 through participation in the National Parole Institutes directed by Vincent O'Leary. Despite findings in almost all of the comparisons ever undertaken that actuarial tables make more correct predictions than case studies—whether by psychiatrists, psychologists, sociologists, prison wardens, or fellow prisoners (Manheim and Wilkins, 1955: 158-160, 170, Appendix V; Postman, 1962, Ch. 9; Glaser, 1962: 242-245)—there was much resistance to what was called “letting a statistical clerk decide about human beings.” One commentator observed: “Prisoners were once numbers. Now they may become holes in a card, and so will their relatives” (Evjen, 1962).

It certainly is granted that actuarial tables do not answer such questions as, “How much risk should I, as a judge or a parole board member, take in releasing this offender?” Those tables currently available have just begun to address such questions as whether the offender will be a better risk after longer confinement than he is now (cf. Jaman and Dickover, 1969; Babst, *et al.*, 1973). Also, no tables have yet been prepared, although they are quite feasible, to indicate the risks of different kinds of recidivism—for example, the commission of violent crimes or petty offenses—and this probably is the key consideration of judges and parole boards in deciding whether to order release or confinement. Yet actuarial tables could augment the demonstrably less accurate case studies that judges and parole boards now receive to help them guess these risks. And the tables still would not make the decisions, of course, partly because various humanitarian, public relations, and other considerations are involved, in addition to risk, in each judicial or parole decision, and partly because all offenders have some unique positive and negative features which may persuade a decision-maker to go against the statistical odds in a particular case.

Some years ago it occurred to me that, since prediction tables are intended only as advisory devices rather than decision machines, there might be less resistance to them if the officials could see how the risk estimations were reached. Instead of advising:

Offender A is an 80 percent violation risk category because he is in a score group of which, in the past experience, 80 percent of the members violated parole.

The advice from a prediction table might be in the form :

Offender A is in an 80 percent violation risk category because he is a recidivist auto thief, 18 to 21 years old, who never held a job as long as a month, and has a history of disruptive use of alcohol; experience shows that 80 percent of parolees with these attributes violated parole.

It would be even more useful, of course, to be able to add that 50 percent violated by new nonviolent felonies, 10 percent by violent felonies, and 20 percent by petty or technical infractions, and that these figures remain about the same for up to 2 years of confinement, increase thereafter about 10 percent per year up to 6 years' confinement, then decline about 5 percent per year of additional confinement. It would also be well to qualify this advice by data on the variation in rates with different assignments and performances in the institution. All of such actuarial risk data could be calculated from the records of 10 or 15 years' experience in large correctional systems, such as those in California or New York, or the Federal system.

Even with the statistical advice above, decisions might go against the odds in many—if not most—cases, but the decisions would at least be grounded in better knowledge of how the odds were reached. In the illustration above, for example, while Offender A may be in an 80 percent violation risk group because of the attributes indicated, the circumstances of the offenses, his record in prison employment, the loyalty and stability of his spouse, or innumerable other factors might lead a parole board member to regard him as a better risk than the odds indicated—one of the 20 percent of his type who do not recidivate. Or the circumstances of the offenses, for example, like one lad I recall who always abandoned the cars near where he had "borrowed" them and in better condition than when he had stolen them, may make a long penalty seem improper purely from the standpoint of abstract justice. In other cases there may be special reasons for more pessimism than the odds suggest. The risk data are "caution" or "proceed with care" suggestions, rather than "stop" and "go" signs.

The main point is that when a source of statistical risk estimation is presented in words, its relevance to a particular case can be assessed more readily, especially by nonresearchers, than if the basis of the risk estimate is only presented as a score. Similarly, if groupings of inmates for base expectancy analysis in evaluative research are presented in verbal categories, actual rates higher or lower than expected when this category of cases

is in a particular program can be interpreted much more readily than when the base expectancy derivation is a numerical score representing diverse mixtures of score sources. With the more readily interpretable actuarial risk data expressed verbally, but derived from precoded input data to replace narrative reports (as noted in the preceding chapter), it might be appropriate to return to a narrative report, but this time only as a brief commentary on the *fit* of the actuarial assessment of each particular case.

The reason for developing multivariate scoring in the first place was the belief that many factors must be taken into account to produce an optimal separation of clients into risk groups. Multivariate analysis suggested, however, that if one considered the few most salient variables in the optimal sequence and weighting, 6 to 10 variables might yield more adequate risk differentiations than cruder systems with 20 or more variables. Yet using even a few variables to classify all cases by verbal attributes implies their division into two or more categories on one variable (e.g., type of offense), then dividing each of these categories into two or more subcategories by another factor (e.g., prior convictions) then dividing each subcategory into subsubcategories by a third factor (e.g., age), and thus having an enormous number of groupings with less than half a dozen variables. This is known as the method of "manifold classification," and is illustrated in Table 9.1 and in the upper part of Figure 9.1.

In this table and diagram, using only three variables and with each only a dichotomy, we create eight final subdivisions, which we shall call "Types." If we were to divide each of these by a fourth dichotomous variable there would be 16 types; if one variable of the three in Table 9.1—for example, age—were divided into three categories, there would be 12 types. As it stands in Table 9.1, however, the second type has only about 50 cases. Further subdivision, therefore, would create types with so few cases as to make their percentage of success affected largely by chance fluctuations, and hence less dependable than a percentage from a larger sample (even 50 is too small). Nevertheless, with only three variables, Table 9.1 presents a classification of types of offenders with appreciable contrast in rates of success (success being defined here as not being reimprisoned).

Another approach to subdivision into verbally identifiable types is to assume that once a population is divided into risk

Table 9.1.—A Manifold Classification Table for Typing Offenders by Eight Categories of Recidivism Risk

(Based on a 10% sample of 1956 releaseses from U.S. Federal prisons: N=1015)

Variables Employed and Their Categories:

A. Prior Institutional Commitments: 1. None; 2. One or More

B. Age at Release: 1. 31 or older; 2. 30 or younger

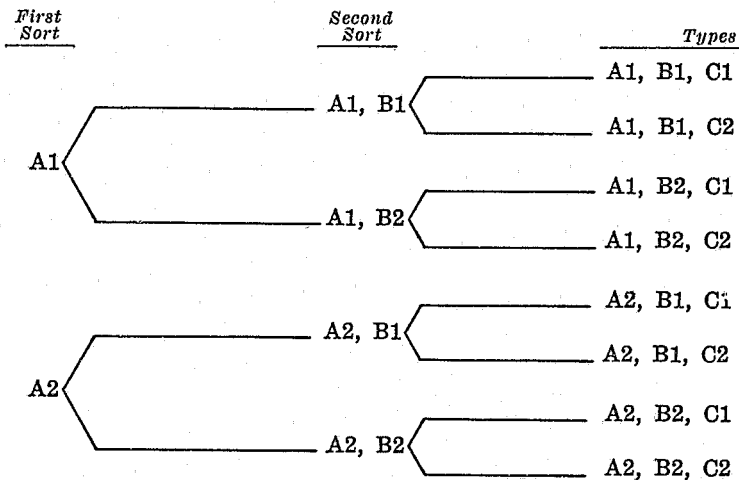
C. Longest Job While in Free Community: 1. 1 or more years; 2. Under 1 year

<i>Types</i>	<i>Percent of total releaseses</i>	<i>Percent not reimprisoned in 3 years</i>
1. A1, B1, C1: No prior institutional commitment, age 31 or older, 1 or more years at longest job	8	86
2. A1, B1, C2: No prior institutional commitment, age 31 or older, under 1 year at longest job	5	75
3. A1, B2, C1: No prior institutional commitment, age 30 or younger, 1 or more years at longest job	11	81
4. A1, B2, C2: No prior institutional commitment, age 30 or younger, under 1 year at longest job	9	62
5. A2, B1, C1: One or more prior institutional commitments, age 31 or older, 1 or more years at longest job	9	76
6. A2, B1, C2: One or more prior institutional commitments, age 31 or older, under 1 year at longest job	20	60
7. A2, B2, C1: One or more prior institutional commitments, age 30 or younger, 1 or more years at longest job	9	69
8. A2, B2, C2: One or more prior institutional commitments, age 30 or younger, under 1 year at longest job	29	54
All Types	100	65

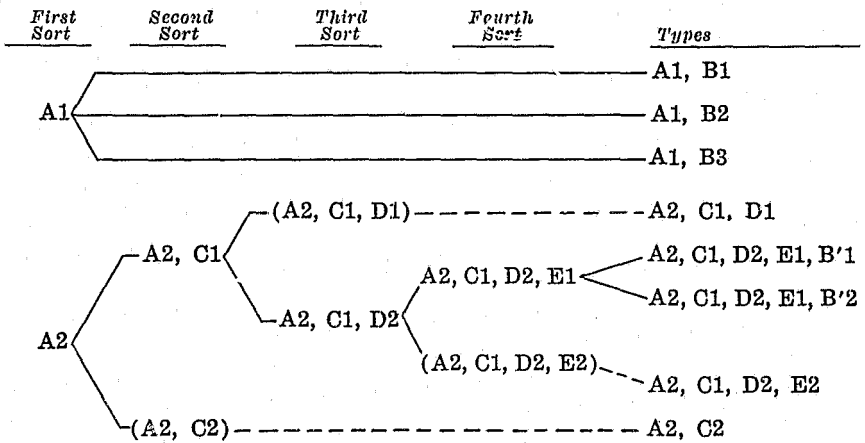
Based on Glaser, 1964:296, but with some estimations based on unpublished printouts of reimprisonment rates for all categories of each variable taken separately, as this manifold classification was not actually attempted on these data.

Figure 9.1.—Sorting Patterns for the Manifold Classification in Table 9.1 and the Configuration Analysis in Table 9.2

1. *Manifold Classification:*



2. Configuration Analysis:



Parentheses indicate sort at which final type was reached for one category when other category was not yet at final type, and broken lines connect these with the same entry in the "Types" column.

Solid lines connect product of one sorting to the next sort, or to the "Types" column when there are no further sortings of any of the products of a given sort.

groups by one variable, the optimal variable for further differentiating the resulting categories to identify contrasting risk groups might be different for each of the categories from the first variable. This is shown in Table 9.2 and in the lower part of Figure 9.1, in what has been called a "Configuration Analysis." It is from the same sample as Table 9.1, and with the first division (A1, A2) also made by the most differentiating of all the separate variables for this sample, Number of Prior Institutional Commitments.

In Table 9.2, those with *no* prior institutional commitments were found most differentiated into risk groups by an age trichotomy, but the two-thirds of the sample *with* prior institutional commitments were most differentiated into risk groups by their prison adjustment. Apparently, while many prisoners with exceptional difficulty in adapting to prison life on their first commitment are deterred from further serious crime by this experience, those who have such difficulties even when institutionalized for the second or subsequent time will probably also have difficulties in the free community and be reconfined after their next release.

Since the previously confined individuals with no record of unusually serious problems in prison adjustment evident in the

Table 9.2.—A Configuration Table for Typing Offenders by Eight Categories of Recidivism Risk

(Based on a 10% sample of 1956 releasees from U.S. Federal prisons: N=1015)

Variables Employed and Their Categories:

- A. Prior Institutional Commitments: 1. None; 2. One or More
- B. Age at Release: 1. 36 or older; 2. 24 through 35; 3. 23 or younger
- B' Age at Release: 1. 31 or older; 2. 30 or younger
- C. Prison Adjustment: 1. Satisfactory; 2. Unsatisfactory reported
- D. Longest Job in Free Community: 1. 4 or more years; 2. Under 4 years
- E. Prior Paroles Violated (anywhere): 1. None; 2. One or more

<i>Types</i>	<i>Percent of total releasees</i>	<i>Percent not reimprisoned in 3 years</i>
1. A1, B1: No prior institutional commitment, age 36 or older	7	93
2. A1, B2: No prior institutional commitment, age 24 through 35	16	78
3. A1, B3: No prior institutional commitment, age 23 or younger	10	64
4. A2, C1, D1: One or more prior institutional commitments, Satisfactory prison adjustment, 4 or more years at longest job	6	79
5. A2, C1, D2, E1, B'1: One or more prior institutional commitments, Satisfactory prison adjustment, Less than 4 years at longest job, No prior parole violation, 31 or older	16	72
6. A2, C1, D2, E1, B'2: Same as #5, but 30 or younger	18	60
7. A2, C1, D2, E2: Same as #5, but regardless of age, violated one or more prior paroles	12	51
8. A2, C2: One or more prior institutional commitments, Unsatisfactory prison adjustment	15	42
All Types	100	65

Based on Glaser, 1964:296

files comprised over 62 percent of the sample, they could be subdivided further. Longest job in the free community was highly differentiating here, but only in separating out as a highly successful small group the 6 percent of the same who had been both institutionalized once before and held a single job in the free community for 4 or more years (generally between a juvenile or youthful criminal record and their much later current imprisonment). The remaining 52 percent with prior incarceration and satisfactory prison adjustment, but without this exceptional employment record, were most differentiated further into a poor risk group of 12 percent of the sample who had violated one or more paroles, either from the current sentence or from previous confinements, including those from institutions for juveniles. Only for the residual of the sample now left was age the factor associated with the most contrasting reimprisonment rates, but for this group the optimal age group division was into those above and below 30 years.

Both the manifold classification and the configuration analysis shown here, from the same sample, produced eight final risk categories or types. A comparison of Tables 9.1 and 9.2 shows that the eight types from the configuration analysis have appreciably more contrast in reimprisonment rate than the eight types from the manifold classification.

The configuration analysis employs five variables, but not for all sortings, and it includes one variable as trichotomy, whereas the manifold classification can use only three dichotomies if it is to end with only eight types. Of course, further sorting would have been feasible in either analysis had the sample been larger, but the prospect of being limited by few cases in one or more types before many variables are used is especially great with manifold classification. Because of the interrelationships among variables, however, the contrast in success rates produced by an additional sorting is likely to diminish relatively rapidly with either method, but less quickly with configuration analysis than with manifold classification. The interrelationships among variables that will result in a combination of 5 to 10 usually predicting as well or better than 20 to 30 were discussed and illustrated in pointing out the advantages of multiple correlation over simple additive points for multivariate scoring.

The tree-like branching system of successive subdivision to produce the configuration analysis presented in Table 9.2 and Figure 9.1 was performed crudely with punched card accounting machines rather than computers in this 1958-62 study. The procedure was simply to sort all cases on every item on which information was available (about 55), to determine which differentiated the cases into categories of more than an arbitrary minimum number of cases with the largest contrast in reimprisonment rate. Each category from this sorting was then divided successively by all the remaining variables to determine which divided it into the most contrasting categories, and this continued until there were no categories left with more than twice the minimum acceptable number of cases. The sorting was also terminated if no significant contrast in success rates could be found with further sortings.

The arbitrary minimum number of cases set for a category in Table 9.2 was 60, actually much too small a figure for confidence that the findings would not be appreciably affected by chance fluctuations. The intention at this point was only to demonstrate an alternative method of combining input information for actuarial and evaluative research. The major reason for not

having more cases was that the information had to be gleaned from case files of predominantly narrative content, a tedious and inaccurate process that can be eliminated with what Chapter 8 called "automating input data."

The first major step in making configuration analysis procedures more mathematically sophisticated was the Wilkins and MacNaughton-Smith (1964) use in parole prediction of two statistical procedures—Association Analysis and Predictive Attribute Analysis—developed in plant ecology. About the same time, Sonquist and Morgan (1964) developed the Automatic Interaction Detector (abbreviated "AID") computer program, which searches for tree-like configurations of attributes that have maximum predictive power (usually defining this power technically as explaining the maximum percentage of variance in the criterion). Some years later Sonquist (1970) reported application of AID to a variety of social policy problems. To my knowledge its first application to correctional prediction was Newman's (1972) analysis of recidivism by misdemeanor offenders released from the Los Angeles County Jail, although his outputs turn out to be closer to a manifold classification than a configuration analysis, as these have been distinguished here.

At any rate, with modern computer technology it is possible to determine risk groups in terms of verbally describable categories at least as adequately, from a statistical prediction standpoint, as with multivariate scoring by multiple correlation. Babst, *et al*, (1968, 1971) found multiple regression and a crude configural analysis equally predictive and equally stable. Simon (1972) found that a variety of sophisticated methods all "work equally well." Of course, with all methods there are risks of regression effects; these effects make the differentiation of risk groups in future samples less contrasting than they appeared to be in past samples, so that predictions of highly favorable or unfavorable prospects prove to be exaggerated. This problem can be diminished with research samples of past cases sufficiently large to be broken into separate studies for successive periods, so that one may then use for prediction only those types of differentiations that showed stability from one period to the next in the past.

As input and outcome data on people-changing experience are both increasingly automated through computer-stored record systems, and as sophisticated configuration analysis also is done with computers, actuarial risk information on verbally designated client categories should become more routinely available as an

aid in evaluation research. As implied at several points in this volume, the most fruitful assessments of people-changing efforts are likely to be the extent to which they alter risks for specific types of clientele, rather than for every type of client. Routinization of configuration analysis should enhance our ability to identify the types of client for which each type of program is most profitable and efficient from a cost-benefit standpoint.

Conclusion

Methods of consolidating statistical input information for evaluative research in people-changing agencies can be classified as either multivariate scoring or configuration analysis.

Numerical scoring procedures transform diverse information suggesting the causes of past success or failure of people-changing efforts into numbers on a prognostic scale. Such scales, known as prediction tables or equations, generally provide much more accurate forecasts of the behavior of clients of people-changing agencies than can be procured from case-study prognoses. Administrators, however, are reluctant to rely on these predictive scores, either as sources of advice in case decisions or as base expectancies for assessment of treatment programs. One reason for this disuse is the failure of the tables and equations to address all of the types of predictive judgment that enter into most policy decisions, although such judgments could be usefully guided by appropriate actuarial prediction studies. Some of these tables, notably those used as base expectancies, reflect the disadvantages of correlational rather than experimental methods of investigating effectiveness, especially where the possibility of quasi-experimentation has not been created by many policy changes and much policy inconsistency. A more important reason for this disuse is the fact that a numerical score masks the diverse information by which it is determined. This makes it difficult for administrators to assess how well a statistical prognosis fits an individual case, and why its weighting system is presumed to yield the best predictions.

Configuration analysis also reveals the most prognostic combination of statistical information, but it identifies the information utilized by verbal labels. This clear designation of the basis for its actuarial prediction enhances the prospect that such prognostic devices will be taken into account in policy decisions. When operations officials can readily comprehend the source of a predictive classification in treatment evaluations,

they can assess the relevance of statistical advice on a program's probable effectiveness in a particular case. They can also call for new types of statistical tabulation, to test any hunches they may have as to a more adequate generalization on what programs are most effective for particular types of clientele. Thus a major advantage of configuration analysis is that it permits an interaction of explanation with statistical prediction and evaluation. Consequently, configuration procedures should help to routinize an actuarial approach to obtaining optimum prescriptions in people-changing endeavors.

CHAPTER 10

EVALUATION RESPONSIBILITIES: WHO SHOULD EVALUATE WHOM?

In many people-changing organizations the term "research" is considered synonymous with "investigation," and both are viewed as threatening. Nevertheless, with the expansion of government and foundation funding for these agencies, especially since the early 1960's, a frequent requirement for grant applications has been that they include a procedure for evaluation. It is often insisted that the budget of each treatment agency provide a certain percentage for evaluation.

Relatively little can be learned from most of these evaluation efforts, even in cases where a project is labeled an "experiment." In the lexicon of many, the term "experiment" is confused with "innovation." Anything that can be sold as "new" is given a label which suggests that it will be scientifically appraised, even though no control or comparison group measurements are undertaken. Frequently—my impression is usually—the research consists only of describing what has been done, rather than of determining its consequences. Research employees are required to assist in public relations, to count services rendered clients, or to do miscellaneous nonresearch functions, including augmentation of treatment staff, rather than evaluate as this concept is used here.

In this age of "grantsmanship," as Campbell (1969) points out, every proposal is presented as certain to be successful. But, he also asserts, "If the political and administrative system has committed itself in advance to the correctness and efficacy of its reforms, it cannot tolerate learning of failure."

I have been repeatedly informed by research personnel at correctional and addiction treatment organizations, both in the United States and abroad, of suppression of their followup studies by top officials of their agency. Nevertheless, one organization, the California Youth Authority, has been publishing evaluative

research studies regularly since the inception of its research office 15 years ago. Doubtless this has been facilitated by its having long continuity in agency direction under Heman G. Stark and Allen F. Breed, and one Director of Research continuously, Dr. Keith S. Griffiths, but I shall suggest an additional reason for their achievements. Other correctional agencies—for example, the California, Massachusetts, Wisconsin, and District of Columbia Departments of Correction, and the Los Angeles County Department of Probation—have had periods of releasing followup studies interspersed with periods of either not conducting such research or of not publishing the results of what they had done. Their shifts to and from bona-fide evaluative research seem to have been largely a function of changes in leadership of the organization, and occasionally in the research staff. To some extent, however, I think it was a function of the range and scope of activities the various researchers and agencies proposed to evaluate.

It is my impression that the most regularly suppressed data are on overall effectiveness of an entire system, and on system-wide programs that were specially funded on the basis of great promises. I would contend that there is wisdom in not publishing overall effectiveness data on entire people-changing organizations or programs that include diverse clientele and a variety of markedly different facilities, just as there is wisdom in not publishing overall death rates for general hospitals. In all these cases, the outcome is dependent primarily on input. One would not expect the same death rate, for example, in a hospital prominent for its eye and ear operations or its plastic surgery that one would expect in a hospital predominantly engaged in cancer research or heart surgery.

What can be published most readily and usefully, both for people-changing organizations and for medical hospitals, are the outcome rates for alternative treatments for a given type of condition. This may strongly indicate the preferred treatment, and might guide investment in facilities or staff for one type of treatment rather than another. Such findings would not necessarily be immediately conclusive for all cases, however; their publication might only promote considerable discussion, replication, and further analysis to validate initial findings—perhaps to specify more precisely the types of clientele or circumstance to which the conclusions are most applicable.

An extension of this medical analogy suggests that routinization of evaluative research will be maximized with three types

of allocation of research responsibility: inhouse trials, outside monitoring, and hierarchical auditing. A fourth type, the outside agency project, which is different from outside monitoring, will also be considered. These outside projects are usually nonroutine and poorly evaluated, but they can be a definite source of evaluative research and they could conceivably be routinized as a more dependable evaluative procedure.

Inhouse Trials

The California Youth Authority has been able to maintain evaluative research without as much interruption as has occurred in other people-changing organizations, largely I believe, because it has persistently started its innovations as limited trials. These were generally controlled experiments in only one or two of its facilities or districts. Such trials have assessed, much more rigorously than is customary in corrections, psychiatric teams for youth institutions, diagnosis and prognosis by a personality inventory, intensive small-unit institutional treatment for younger wards, intensive short-term institutional treatment for older wards, intensive differential treatment in the community, special narcotic addict counseling and naline testing, intensive reading instruction, job placement programs, small parole caseloads, community-involving parole centers, and other variations in treatment practice. Research staff were involved with operations personnel in planning the trials so that their consequences could be measured. It is also noteworthy that these projects were not launched with great fanfare, and they were called experiments or trials from the start so that there was no promise of demonstrable effectiveness.

As a result of this approach, the Youth Authority has learned from its trials. Most produced only negative findings, identifying only practices that had no effect on outcome, but these results were beneficial since they curtailed further investment in the kinds of programs that were tried but proved fruitless. Many studies had mixed results, often predominantly negative but sometimes positive, such as specifiable groups among the clients showing beneficial consequences from some features of the trial programs. Negative data often suggested why a measure had been ineffective and indicated the problems which the programs neglected. This was beneficial because it contributed to the design of subsequent trials, notably the community parole center programs (Seckel, *et al*, 1973). These seem to be more

successful in difficult settings than previous parole service projects because they adapt more extensively to sociocultural variations among communities, they link parole more to both the institutions and the community, and they provide parolees with greater variety and attractiveness of legitimate opportunities for success and excitement than these youths have previously experienced. The advantages of this continual inhouse trial policy become even more evident when the youth authority's record is contrasted with that of other organizations with different approaches to correctional change.

The California Department of Corrections, when it was also initiating its Research Division, in 1958-61, launched with great fanfare a program of group counseling of inmates by line staff. This program was encouraged at all of its facilities, though stressed more at some than at others. It was publicized at national correctional meetings and in the mass media as a major advance toward the rehabilitation ideal. Special Federal grants were received to undertake some controlled experimentation with variations of this program, such as size of group, hours per week, and type or ratio of staff supplied for counseling, but a maximum push for these types of programs did not await conclusion of the experiments. Despite accumulating research evidence that the major effects of the programs, if any, were to alleviate tensions of institutional life rather than to alter recidivism rates, about a decade elapsed before a marked shift occurred in the Department's promotion of group counseling at closed institutions (in contrast to halfway houses) as a primary rehabilitative measure.

It has been demonstrated repeatedly that legislators and the mass media are much more supportive of failure in new programs frankly introduced as limited tests or experiments, than of failure in large-scale programs introduced with great promise as sweeping solutions to difficult problems. If clear progress does not materialize, a backlash reaction can ensue that generates demand for a complete shakeup, as has occurred repeatedly with delinquency prevention and narcotic treatment crash programs initiated as massive solutions to problems in some localities, and has sometimes been experienced by the community mental health movement. This explains why overall outcome rates or rates for widespread, highly publicized programs tend to be suppressed or distorted, especially by organizations with leadership that has concentrated on salesmanship.

When programs are introduced only as limited trials and

negative findings are frankly admitted and honestly interpreted, the organization is likely to retain support despite failures. Furthermore, with firm evidence on the relative effectiveness of alternative policies and programs, an organization can answer critics by showing that it is following the best policies thus far demonstrated, and trying others. This certainly has been the record on the whole of the Youth Authority. Indeed, it appears that valid reports can be expected only when an agency has long been committed to testing all alternatives under its control, rather than claiming that all of its programs are beneficial.

Outside Agency Projects

An alternative to inhouse trials, for piecemeal and cumulative learning about the effectiveness of innovations, is to have new programs introduced by a contracting agency—such as a university, a clinic, or a religiously sponsored residential treatment home—working with the clients or potential clients of a more permanently established government or private organization (such as a court, a department of corrections, or a department of mental health).

The advantage of contracting with an outside agency is that it supplements the routine staff of a treatment organization, often with persons having an expertise, a relationship to clients, or a commitment and dedication to a particular treatment method or to research that traditional organizations lack. In addition, of course, these projects often are funded by grants from a Federal or other government source, or from a foundation, thus supplementing the services that an organization can provide from its own resources. There are four variations, however, in the commitment of the "inside" (host or sponsor) organization and of the outside agency to action and to research, which greatly affect whether outside agency projects contribute to the growth of knowledge.

The first variation occurs when neither the sponsoring organization nor the outside agency is as much interested in research as in action. Occasionally they are frank about this, and emphasize that a project is "a demonstration project rather than a research project." The most that can be demonstrated without systematic followup of clients or minimally, systematic recording of in-treatment failures, is simply that the project can operate, be accepted, attract clients, attract public support, and even point to some "success cases." This may be enough to

satisfy many, even though there is no record of the failure and success *rates*, or evidence that the successes would not have occurred without the project.

The second variation is the main source of failure in efforts of funding agencies to procure valid evaluations; this occurs when the funding agency is committed to evaluation but the outside agency is strongly committed only to action. Programs that demonstrate they can involve the community, or give people a sense of love or a feeling of serenity, or propagate a particular philosophy of life or religion, may thereby be sufficiently evaluated for those participating in them or supporting them. Promotion of a particular treatment ideology, or of a total "way of life" is the primary concern of those involved in many such treatment agencies (e.g., Synanon and many other mutual aid antidrug or antidelinquency organizations). Organizations of this type that are successful in getting people with drug addiction, delinquency, or mental depression and anxiety to participate fully in them doubtless may help the participants; the evaluation problems for a funding office, especially a unit of government, are to procure statistics on what type of persons these organizations can attract, hold, and change, and for how long.

Funding organizations that rely on grantees of this type to evaluate themselves generate the kinds of evaluation which were described in the opening paragraphs of this chapter. Such nonevaluation by personnel funded through budgets that provide an evaluation staff is concentrated in organizations committed more to action and self-perpetuation than to evaluation. One of the common ploys of these organizations, often advocated seriously and not regarded as deception, is not to count as failure anyone who engages in drug use, crime, or other deviant behavior after he leaves the organization. Yet the organizations expel anyone who engages in these activities after he is admitted, so they can have no failures in their organizations simply by their definition of membership in them. A client also frequently undergoes a long trial period of participating and receiving support and services while regarded only as an applicant and not as a member. Data are then not provided on the applicants who are denied membership because they relapse to their former deviant behavior, or on members who simply leave the organization, the "splittees."

The results of these systems of assessment become especially confounding when they are mixed with evaluations derived from

other procedures. Thus Efren Ramírez, a very inspiring psychiatrist who has certainly helped many addicts in the centers predominantly operated by ex-addicts he established in Puerto Rico, asserted:

The usual relapse rate among addicts treated at such centers as the Federal Hospital in Lexington, Kentucky, and Fort Worth, Texas, has been about 92 percent. Even the most advanced experimental centers in the United States average a relapse of 70 to 75 percent. However, at the Addiction Research Center in Río Piedras, Puerto Rico, the relapse rate is 5.6 percent. During the past 3½ years 124 heroin addicts completed treatment at the Center, with only seven becoming readdicted (from Vocational Rehabilitation Administration, 1966:171).

When Dr. Ramírez was brought to New York City to found that municipality's Addiction Service Agency, a large bar graph based on the above statistics was prominently displayed in his office, and he cited them in his successful applications for extensive support from the State and Federal governments.

What the quoted Ramírez figures fail to mention, however, is revealed by further perusal of his account of the Río Piedras programs. The 124 who form the base for his percentage are those who completed the reentry phase of this program, which was preceded by an induction phase, and an intensive treatment phase. Each of the three phases normally lasts many months and can last more than a year. Ramírez mentions a patient series of 1,800 in the quoted report, so the 124 who survive through the reentry phase represent less than 7 percent of the series! And the followup time was variable, so that for some of the 124 it could be a matter of only a few weeks.

Even the 7 percent figure cited above may be too high. In the Río Piedras program the induction phase begins with a street or community clinic encounter subphase, advances to a day-night care center subphase, and culminates in a physical and mental detoxification center subphase, all before entrance into the intensive treatment phase. It is not clear where the patient series counting begins, but one Puerto Rican professor advised me that the 1,800 in this series would not include everyone who initiated the encounter subphase, and that they might total 10,000, since for every one who entered the day-night care center subphase, four or five might advance no further than the casual encounters where there is no listing. At any rate, it is clear that the Phoenix Houses which Dr. Ramirez started could not solve New York City's addiction problem as sweepingly as his bar

graphs suggested, despite the many virtues of these houses. His was but the first of a series of relatively short-lived incumbencies in the directorship of the Addiction Services Agency.

Failure to count dropouts in determining success rates is most dramatic at addiction treatment agencies, but occurs also in other people-changing efforts. Lerman (1968) reports that a private residential center for boys which he studied in New York rejects 17 applicants for every one it admits, and subsequently expels 31 percent of the admittees "resisting treatment" before they complete the center's program, which has an average duration of 16 months. An evaluation that does not take into account the rejectees and expellees could clearly be misleading.

The third variation in outside agency projects is that in which the agency is primarily committed to research, but the sponsoring organization is primarily concerned with action—or perhaps mainly with the appearance of action in order to maintain public support. This has been the bane of many professors, graduate students, and research corporations who have undertaken to organize and evaluate experimental programs at people-changing establishments.

These experimental programs usually begin with a research plan the researchers thought was agreed upon. Later they find that they cannot carry out the planned research design because either operations personnel or top officials interfere with it, either by direct action, such as transferring experimental clients to control programs or vice versa or cutting off funds. My colleague, LaMar R. Empey, and his associate, Maynard L. Erickson, provide a vivid account of how this occurred in the famous Provo Project (1972: Ch. 8). They conclude:

Prevailing beliefs about the best way to help delinquents and criminals determine what any decision-maker will see in a new program, and whether he will accept or reject it

. . . if any correctional enterprise departs too far from current beliefs and practices, it may not receive official blessing if it fits too well with existing patterns, it is not likely to be very innovative. Even if official approval and funding for an innovation can be obtained, it must still surmount the high degree of inertia that is associated with existing budgetary commitments, staffing patterns and organizational arrangements. High-level policy-makers, no less than their underlings, are often severely constrained by existing institutional patterns

. . . there is serious reason to question whether, in a highly political world, the policy-maker, even the practitioner, can ever be expected to live with the tentativity that is required in scientific work, especially where that work is conducted in the goldfish bowl in which correctional programs operate. While the success and prestige of the scientist require it, those of the policy-maker (and the politician) do not.

In their dealings with county and staff officials, experimental staff usually tried to be as candid as possible regarding the controversial methods that were being used . . . and the likelihood that they would provide some kind of correctional panacea. Since, in the past, few programs had ever been highly successful, that fact was freely admitted. Much to the chagrin of the investigators, however, this candor boomeranged. Even among State officials, where greater understanding might have been expected since they were in the correctional business themselves, objective information was used, not in the spirit in which it was presented, but as a weapon against the Experiment. It was as though a scientist studying cancer was being punished because he expressed reservations about the chances he would find a cure. The lesson that was learned is that any investigator who openly shares information according to scientific rules may, according to political rules, find himself hoisted on his own petard (Empey and Erickson, 1972:170-171).^a

Usually the researcher shares with the policymakers of the host organization some concern that public support be maintained. Eagerness to cultivate good will, awareness that they are guests in someone else's facility, and recognition that ultimate responsibility for the fate of the clients still rests with the host organization, also motivate many researchers to "bend over backwards" to go along with requests of facility administrators, funding offices, or top policymakers in order to keep their research project going. The result of too much compromise often is a research enterprise that cannot be evaluated, because no research design was consistently followed or because planned data-gathering was not completed adequately if to do so would interfere with the demands of operations personnel. Fortunately, before its funds were cut off the Provo Project was able to handle enough offenders in an experimental program to permit achievement of a significant evaluation by following up experimental subjects and controls 4 years later.

These problems not only beset agency researchers, but also occur when the staff of research offices within a large organiza-

^a Reproduced with permission from *The Provo Experiment: Evaluating Community Control of Delinquency*. Copyright: Lexington, Mass. Lexington Books, D.C. Heath and Co., 1972. Further reproduction prohibited without permission of copyright holder.

tion undertake inhouse trials. In both outside and inhouse projects, a change of personnel at the top of an organization may eliminate the backing the researchers thought they had at the beginning of an experiment. More often, in my experience, the independence of local operations personnel from top-level control of their daily routines leads to failure to follow research design procedures that deviate from what is most convenient or traditional, so that research design is not followed as exactly as the research office presumes. This is especially frequent when experimental and control groups in the same facility are supposed to receive different services. Frequently these deviations from plan are not discovered until it is too late to correct them, and perhaps they are never discovered.

The experiences described here are especially probable when a research project is begun with a plan that is vague, when the project plan is not well communicated to operations personnel at all levels, and when the duration of the research restrictions is not clearly specified in advance and insisted upon thereafter.

The fourth variation in outside agency projects occurs when the host organization and the research agency are both committed primarily to research, that is, to expanding knowledge, rather than just to supplying treatment services. It is only when this consensus prevails throughout the duration of a study that the enterprise can appropriately be called "an action research project" or "a demonstrable research project." Such a condition is especially probable if:

(a) The theory underlying the research design is clearly stated and the research design is explicitly deduced from the theory, so that the reasons for insisting that the design be meticulously followed can be evident to top-level officials of the host organization.

(b) The exact requirements of the research design are spelled out, including the demands and restrictions they impose on operations personnel, and the duration of these requirements.

(c) The foregoing requirements are made a matter of formal contract, as a written and binding obligation, preferably with compensatory payments to the researchers if the conditions are violated.

(d) The contract is distributed to all officials and researchers who make decisions or actions that could have

a serious effect on the project's conformity with the research design.

After learning of the importance of measures such as those above from the experience at Provo, Empey carefully saw that most of these measures were followed in his Silverlake experiment (Empey and Lubeck, 1971). It represents a procedural model for outside-agency action research projects. If this model is followed, organizations can contract for the equivalent of inhouse trials and derive the benefits of such trials, even when the organizations themselves do not have the research personnel to conduct adequate assessment of the trials.

There may be two disadvantages to an outside agency project, as compared to an inhouse trial. The first is that since it is directed by an outside agency, after the outsiders depart, the staff of the sponsoring organization may not have the commitment, the capacity, or the backing to carry out the experimental programs should they prove effective. The programs that the project initiates die when the project terminates. Sometimes this is justifiable, however, since projects cannot be evaluated until some followup and analysis period has elapsed.

If there is a commitment to continue experimental programs before the evaluations are in, there is likely to be pressure to suppress or ignore findings that are unfavorable to the programs. Organizations tend to become committed to their activities, which represent jobs and budget both rewarding to those in the activities and enhancing to the status of the agency or division director who supervises them. Experimental programs thus may die more slowly than they should, as appears to have been the case with some of the heavy investments in group counseling within adult penal and narcotics treatment institutions in California. Alternative approaches are not tested as soon as is warranted whenever negative evidence about a current approach is suppressed or ignored. The detachment of outside research agencies may thus be an advantage rather than a handicap in many cases.

A second disadvantage to outside agency projects may be the readiness with which shifts occur in the commitment to research on the part of either the outside agency or the host organization, if they are not under the same direction. Should shifts of interest in research occur, as indicated by our typology of four variations, the project may well cease to be an action research project, unless all the contractual conditions suggested above have been

clearly established. Even then, the contracts may not be enforceable without mutual good will. Presumably, an organization which has a large investment and tenured personnel in its own research office may not abandon research so readily. Such reluctance has unfortunately not persisted in most agencies familiar to me, particularly where the heads of organizations change whenever there is a change in the political party in power. An ostensible disadvantage of outside research projects may turn out to be an asset for society.

The independence of outside researchers, especially when they are tenured university faculty, may be essential to the ultimate dissemination of the findings of an experiment that an administration decides to squelch. Dissemination can only occur if the experiment has already proceeded far enough for meaningful results, if the followup may be adequate, and if the analysis may be completed despite political interference with the experimental plan. If Empey and Erickson had been State or county employees instead of university professors, we possibly would not have gained the knowledge provided by the Provo Project report (Empey and Erickson, 1972).

It is possible that a long-term arrangement may be institutionalized in some jurisdictions, whereby a university or a research firm or foundation regularly conducts action research projects for a treatment organization. Continuity of relationships in such cases may cause the impact of research on practice to be greater than it would be in contractual relationships on a project-by-project basis.

When outside agency research projects are contracted on the basis of bidding for each project some economies may be achieved, but there is a severe risk that inept and incompetent research may be done, primarily for profit. It is especially important in hiring senior researchers, whether as inhouse research directors or as outside contractors, that they be evaluated by the quality of the projects they have completed *through the final report stage*. Experience in research, or in people-changing activities of the type being researched, or simply in writing impressionistic commentary on research or practice, is often extremely irrelevant as compared with research competence demonstrated by satisfactory reports on completed research. Also, academic training that does not involve such research, and especially training or experience that is more in clinical than in research activities, may be irrelevant, and can even be a handicap to evaluative research performance.

When outside agency projects become more demonstration than research in orientation and evaluative research is desired, even for evaluation of a people-changing organization's own treatment program, a relationship to outside researchers different from that of the outside research project may be desirable. This is the next topic for our consideration.

Outside Monitoring

Many treatment funding organizations simply do not have the personnel to evaluate the projects they support. They may also be caught in a conflict of interest if they evaluate projects they fund, because the failure of projects they have financed might be construed as evidence of their poor judgment in disbursement of funds. Accordingly, it is often appropriate for a funding agency to issue two contracts at once to two different agencies: one for a people-changing endeavor, and another for research to evaluate this endeavor. Both such outside monitoring and outside agency projects, in contrast to inhouse trials, have sometimes facetiously been referred to as "outhouse research."

Shortly after he and his wife, Dr. Marie E. Nyswander, initiated the methadone maintenance approach to treatment of heroin addiction, Dr. Vincent P. Dole of Rockefeller University successfully urged that his program at Rockefeller be evaluated by an independent agency, and that this should be the model for evaluation of all addiction treatment programs (Dole and Warner, 1967). His funding source, the New York State Narcotic Addiction Control Commission, accordingly also funded the monitoring of his operation by Dr. Frances Gearing of the Columbia University School of Public Health. It was in addition stipulated that her program be supervised by an independent Evaluation Committee composed of representatives of a variety of treatment agencies, universities, and research offices in the area. This committee meets about monthly, discusses monitoring plans and operations with her, and reviews draft reports before they are finally promulgated. For many years it has included some severe critics of methadone maintenance as well as passionate protagonists of the treatment.

An even larger outside monitoring operation is that established at the Texas Christian University Institute of Behavioral Research, directed by Professor S. B. Sells, which evaluates over 50 different addiction treatment agencies financed by the National Institute of Mental Health. This center receives from

each of these agencies a one-sheet precoded form for each addict on his admission to treatment, and another one-sheet form every 2 months thereafter or at the termination of treatment, whichever comes first. These reports permit the University's Institute at Fort Worth to provide increasingly longer followup studies on larger numbers of addicts in more diverse programs and settings than has heretofore been possible. For example, as cited in Chapter 4, their early reports indicated that only 29 percent of the addicts admitted to therapeutic community programs were retained in treatment for as long as a year, as compared to 65 percent retention for a year or more in methadone maintenance programs (Joe, *et al*, 1972: 30). Since, as indicated earlier, therapeutic communities are prominent among the operators of outside agency projects which do not get evaluated meaningfully, this kind of monitoring can be an important source of information.

In New York, the State Narcotic Addiction Control Commission has subsidized over 40 "accredited agencies," mostly therapeutic communities. It receives admission and termination of treatment reports for each client of these agencies. A 1969 tabulation of retention rates for a number of such agencies revealed that one politically active organization which had its representative attend court trials to urge judges to release addicts to it on probation, would in 6 months lose over 80 percent of those it admitted. Indeed, most of them departed in the first few weeks. Yet this agency had its residents carrying coffins in front of the hotel where the governor was making a speech, since they alleged that the addicts in their care would die if the agency's State funding were cut.

Any system of outside monitoring that depends on special forms sent to it from the monitored agencies has the problems of automating input data described in Chapter 8. A form filled out to be mailed to a distant office and not used again at the location where it is filled out is likely to be completed carelessly. There will be no feedback if entries are in error, especially if no spaces on the forms are left blank.

One method of alleviating these monitoring problems is the employment of field supervisors to train staff of the monitored agencies in filling out the forms correctly, to check on the way they are filled out, and to make descriptive reports on the treatment services and operations of the monitored agencies. This description of agencies is necessary to know the kinds of agencies and treatments to which the retention and other evaluative

statistics apply. It should be noted that such field surveillance is done for the Institute at Fort Worth. In addition to retention data, its forms include a variety of background information on those admitted to treatment, as well as notations on employment, arrest, participation in treatment, and other aspects of performance which will ultimately permit evaluation from much more complex standpoints than mere retention.

Despite the field supervision and the threat of termination of funding if reporting is unsatisfactory, many treatment agencies cooperate poorly with outside monitors. The agencies frequently receive funds, accreditation, or both from several different sources, each of which requires regular submission of reports on every client. In addition, the agencies usually have reports and record forms they develop themselves and use in their operations. They are thus overburdened by paperwork and perform it resentfully and poorly. This suggests the advantage of evaluation that integrates operations and research forms as proposed in Chapter 8, or of monitoring by unobtrusive observations of evaluative data supplied inadvertently by treatment agencies in the course of their operations routines. Such monitoring is often feasible by supervising, licensing, or security service organizations, especially those that have some authority over the treatment agencies for reasons other than evaluation of effectiveness.

Another style of outside monitoring consists of the research organization assigning its own staff to the people-changing agency, to observe and report what occurs there. This type of arrangement is especially important where the phenomena to be evaluated are modes of collective interaction at the agency, or where it is crucial that the rules and procedures of an experimental design be scrupulously followed. These concerns prevail, for example, in the project at the Connecticut School for Boys, where a major disturbance led to the entire institution's being reorganized as a behavior modification and "creation of new settings" experiment directed by the Yale University Psycho-Educational Clinic. The university maintains a graduate research assistant as observer in each of the school's eight cottages, in addition to total project supervision by its faculty members (Dean and Repucci, 1974). More varied and diffuse monitoring has been undertaken by the Center for Criminal Justice of Harvard Law School to assess the total consequences for the juvenile justice system of drastic changes in juvenile corrections

by the State of Massachusetts, including the closing of all State correctional institutions for juveniles.

Hierarchical Auditing

A major theme of this manual, most explicitly expressed in Chapter 4 on cost-benefit analysis, is that evaluative research should be routinized much as accounting systems are routinized in business. Of course, people-changing agencies have bookkeeping systems to account for their receipt and disbursement of money, but they do not prepare figures on their effectiveness as regularly and precisely as they prepare their financial statements, or as business enterprises prepare profit-and-loss statements.

As should be abundantly clear from this manual, the effectiveness of people-changing cannot be ascertained with nearly as much precision as profits in most businesses. Furthermore, since effectiveness is necessary for officials to have job security and for funding in treatment agencies, and since variations of client input may be the main source of variation in treatment outcome, it is probable that organizations themselves will often not measure their overall effectiveness dependably. Parallels exist when corporations that sell stock to the public would cause the value of their stock to go down if they failed to report regular profits, and for this reason the law does not trust them completely; they are required to have their financial statements checked by licensed independent auditors. Similarly, banks and savings and loan associations are subject to monitoring, auditing, and inspection by Federal and State agencies.

Our discussion of inhouse trials suggested, essentially, that any large organization evaluate those subordinate units and programs among which it must choose in allocating its resources. Thus, while the overall effectiveness of any treatment organization may often not be measured dependably by the organization itself, the organization may be able to estimate accurately the relative effectiveness of its component or subordinate programs, activities, or units for particular types of client. Each organization in the hierarchy of government makes decisions on the allocation of resources among its subordinate units to achieve a maximum utility. An organization should therefore have an objective interest in knowing the relative effectiveness of the different fund recipient components among which it is free to alter its allocations of resources.

It is accordingly proposed that any people-changing agencies

receiving public funding should be to some extent audited for effectiveness by a governmental unit above it in the hierarchy of which it is a part. Applying this principle to a large State, one would expect the department of corrections to make some gross assessments of the overall effectiveness for particular types of offenders of the separate jails and probation systems which it operates, supervises, or subsidizes. Similarly, a State department of mental health might assess the relative effectiveness for particular types of patients of local hospitals, clinics, and community health programs which it subsidizes or even licenses. Finally, State legislative or executive budget offices, and possibly Federal funding offices, might wish to assess the effectiveness of the State departments.

There are two main procedures available for such auditing of people-changing effectiveness. The first, parallel to the auditing of banks by Federal agencies and of stock exchanges, brokers, and corporations by the Securities and Exchange Commission, is simply to set some standards for record-keeping by the audited agencies and to inspect the records. A second and less obstrusive method is available to State and Federal governments in the auditing of people-changing agencies that must, for their own operational needs, submit their clients' fingerprints to State and Federal criminal identification offices.

As indicated in Chapter 7, these fingerprint collecting offices must supply current criminal record data on past releases to research units of correctional agencies if those units are to measure outcome with maximum feasible thoroughness and efficiency. The fingerprint offices themselves can also utilize these records to make independent assessments of the overall effectiveness of judicial sentences and correctional programs for various gross categories of offenders. This consists essentially of compiling statistics on the client flow among all agencies which submit fingerprints when they receive or release a client. This has already been initiated, to a limited extent, by the "Careers in Crime" statistics of the *Crime in the United States* publication of the FBI.

Such fingerprint flow compilation concentrates the power to publish so much critical and diversely interpretable data in one agency that the compilation and publication preferably should be supervised by an independent board of experts. At both the Federal and State levels, board members could be drawn from diverse autonomous agencies, such as universities and research foundations, and various statistical offices of the State or of

large cities. An excellent model for this is the already cited Evaluation Committee of the Methadone Maintenance Evaluation Project supervising the methadone maintenance monitoring directed by Dr. Gearing of the Columbia University School of Public Health.

Such a board would meet regularly to assess the client flow compilation plans and progress of the fingerprint collection office that it supervised. The board would examine drafts of statistical reports and commentaries before their issuance, and would release as part of each report a statement of evaluative comment on the report's accuracy and on its implications. They might well assess the findings from the standpoint of our final concern: how can new knowledge be integrated and coordinated with prior evidence and inference?

Conclusion

Evaluation has been routinely called for by government and foundation offices funding people-changing agencies, but usually has been supplied by these agencies irregularly, inadequately, or not at all. Frequently the findings of completed evaluative research have been suppressed by administrators, who felt threatened by research conclusions.

This fear of rigorous evaluation reflects the political atmosphere of program-funding, in which administrators must "sell" their programs as though they were infallible. This fear also reflects the tendency in political debates to call for global evaluations of total agencies, although the relative effectiveness of different total agencies is often determined primarily by the type of clientele they receive. As repeatedly stressed in this manual, what is most useful is evaluation of the relative effectiveness of alternative types of people-changing activity for a given type of client. Four patterns of allocating evaluative research responsibility were distinguished, each with special implications for fostering the routinization of program assessment.

An approach to evaluation that often serves both the public and people-changing agencies extremely well, consists of *inhouse trials*. In this approach, ideally, the agency tries out new programs, at first only on a small scale and with as close to a controlled experimental design as possible, for rigorous evaluation. While such investigations often yield negative assessments, the conclusion that a program is ineffective is much less costly to society after only a small-scale trial than after large-scale

and long-term use of the program without much objective evaluation. Indeed, one may often learn from predominantly negative results how to modify a program, or how to focus it on a more limited clientele, in order to make it more effective in subsequent trials. It is eminently desirable that a regular part of any agency's annual budget be allocated to inhouse trials, for continuous small-scale experiments with the program innovations it is considering. If inhouse trials are a routine policy and there is regular publication of findings, requests for expansion of programs can be made on the basis of research evidence that the programs have been effective.

Enterprises initially described as experiments, or as "demonstration research" or "action research," but in fact never evaluated, are most frequently *outside agency projects*. Evaluation in any adequate sense of the term usually does not occur when the outside agency and the sponsors are not oriented more to research than to action, do not consistently maintain this research emphasis, or do not share a clear consensus on the purpose and procedure of the research.

There are many potential advantages to *outside monitoring*, in which an organization specializing in research is hired to evaluate activities of a people-changing agency. This arrangement, if the research organization is truly independent, eliminates the conflict of interest involved when a treatment agency is asked to evaluate itself. Such independence may be questionable if the people-changing agency being evaluated hires the monitoring organization; it is less questionable if the office which funds the agency also hires an independent research organization to do the monitoring. Frequently there are problems of non-cooperation of the people-changing agency with the organization that is to monitor it, especially when such evaluation is initiated after the treatment agency is long established and previously has been autonomous. Ideally, the outside monitoring should be contracted at the start of a people-changing program, so that the independent assessment is clearly understood to be a condition of the funding arrangements from the start. In addition, an independent board should oversee the monitoring operations and comment on the monitoring organization's reports.

Hierarchical auditing occurs when large units of government undertake monitoring of people-changing activities by smaller and more localized government units. Such a pattern of evaluation is analogous to Federal and State government supervision of local financial establishments. Ideally, such auditing should

be done by unobtrusive measures, and should be supervised by an independent board. Whether or not there should also be hierarchical control of local people-changing agencies to maximize effectiveness, rather than decentralization and local control, is a complex issue not addressed here. Central monitoring of local treatment agencies by a higher government office and publication of findings may provide useful information input for public opinion, regardless of whether the higher office has regulatory power or not.

CHAPTER 11

MAKING KNOWLEDGE CUMULATIVE: PERPETUAL INVENTORIES OF WHAT WE THINK WE KNOW

One official of a large correctional agency defined research as spending \$200,000 of some other organization's money to produce a 2-inch thick report that is distributed when it is 2 years out of date. This pattern is a tradition of the "project" approach that is justifiably criticized. In contrast, this manual emphasizes making research a continuous concomitant of operations.

Many research inquiries must still be organized on a project basis, but it is important to try to design long-run projects as successive stages, each of which produces useful feedback as soon as possible. It is preferable to have numerous short reports than one monumental report several years out of date—if one cannot have both. Some of the major correctional research projects of the 1960's were 6 to 12 years from inception to publication, with most of this period elapsing after the data-gathering had been completed. The delay was certainly one factor in limiting their impact. Unpublished interim reports to officials and advisors sometimes are of mutual benefit, and help to maintain interest in the project. Furthermore, since the secret of getting good writing completed is to start writing early, but be willing and able to rewrite repeatedly, interim reports often accelerate and improve final reports.

Frequent publication of short articles while also desirable, sometimes blinds readers to the major patterns in knowledge expansion. Not only is it difficult to keep up with all the published literature, but the many small articles we read in a given field often have formulations of problems and of findings that are uncoordinated and not comparable.

Perhaps the least satisfactory solution to this problem was that sponsored by the Federal Government in the 1960's. This

consisted of huge collections of abstracts of publications and summaries of research in progress grouped together by broad topics or simply arranged alphabetically by author. While such collecting sounds like a rational idea, the resulting large "phone books" on current literature were much less useful than their proponents imagined. Their mixtures were of such uneven quality and under such broad headings that one simply had to look through too much to find useful items.

A second approach to summarizing accumulated knowledge on the effectiveness of people-changing measures has not only reflected research imperfectly, but has sometimes been anti-thetical to it. This consists of publishing manuals of standards, position statements, or model legislation that presumably reflect the current consensus of experts. Consensus on what is most effective in these enterprises does not depend on rigorous research, but unfortunately, on drifts of popularity and on persuasive rhetoric in the drafting committee. By fostering illusions on knowledge, these formulations often impede the growth of knowledge.

A third type of solution is represented by highly edited selections, such as those in the publication, *Crime and Delinquency Literature*, issued by the Information Center of the National Council on Crime and Delinquency. While this is a collection of abstracts, it is small and well selected. Each issue also contains one article, and this is frequently an attempt to summarize the total literature on some topic generally related to issues of policy or practice.

The solution I would propose for any people-changing field is a synthesis of the manual of standards and the literature summary approach. Instead of setting forth statements as to consensus on optimal practice, it would set forth propositions as the hypotheses currently most supported by available theory and research, then summarize and cite this support, as well as opposing views and data, for each proposition. It could be developed by a committee or board representing leadership in both operations and research, possibly with some appropriate full-time staff. If this were codified and updated regularly, but well edited for cogency and conciseness, new research might be oriented specifically to testing, expanding, or revising the propositions in such an inventory of the current state of knowledge, and new publications would link their propositions to the current inventory.

A model for the proposed inventory is provided in a section

of my Federal correctional study entitled "Conclusions of this Project as Hypotheses for New Research" (Glaser, 1964: 504-513). Some of the 89 propositions presented there probably are too specific in the rates they present on matters in which exact figures are never known, and estimates soon become dated, but relationships are persistent. The following is an example of how such a propositional inventory might be presented:

Proposition:

The proportion of releasees returned to prison tends to be higher:

a. where probation is used extensively, so that only the worst risks go to prison (although this may make long-run recidivism rates of all felons lower);

b. where parole is used extensively, so that many poor risk parolees are released on a trial basis;

c. where a large proportion of parolees are returned to prison when they have violated parole regulations but have not been charged with or convicted of new felonies;

d. where there is a high overall crime rate in the communities to which prisoners are released, so that there is high prospect of the releasee coming from and going to highly criminogenic circumstances.

Evidence:

This proposition rests primarily on inference because the fingerprint followup studies that could test it conclusively have not been systematically undertaken. There is evidence that for adult offenders, probation is granted primarily to first offenders (England, 1957) and its recidivism rates are low, so if they are less frequently in prison the recidivism rates of prisoners should become higher. Similarly, extensive use of parole creates a prospect of a high rate of reimprisonment for parole violation, especially when the policy is to reimprison readily for technical violation instead of just when new felonies are charged. Finally, it is presumed that the causes of crime and of reimprisonment are related, so that reimprisonment rates should be higher when crime rate, especially felony rates, are higher.

Although there has been no systematic test of the above, scattered studies are suggestive. (Glaser, 1969: 11) reports in the 1950's reimprisonment rates of 31 percent for Pennsyl-

vania and for Federal prisoners, 38 percent for Washington (counting "wanted status" with reimprisonment), 39 percent for Wisconsin, and 44 percent for New York. Kassebaum and coworkers (1971) found 51 percent reimprisonment of releasees during the 1960's from one California prison. According to section (a) of the above proposition, California was one of the leading States in use of probation for felons, with 51 percent receiving it in the mid-sixties, compared to less than 40 percent for the Federal court system in this period (Kassebaum, *et al*, 1971: 289). In section (b) the data are not so consistent. Washington and California were the only States in the above collection paroling over 90 percent of their releasees; the Federal prisons paroled only 31 percent, plus a smaller percentage on "mandatory release" somewhat resembling parole. According to section (c), the proportion of those reimprisoned who were only returned for technical violations, without a new felony conviction, ranged from a low of only 10 percent for the Federal prisoners to a high of 84 percent for New York. A recent California study (Spencer and Beracochea, 1972) showed that the State's high reimprisonment rate for women parolees, almost twice the national average, was due almost entirely to their high rate of return to prison for technical violations without new felonies. Finally, according to section (d) and apparently the most important item, California has usually had the highest Index Crime rates of any State in the Nation, with a rate of 2,826 per 100,000 in 1966, compared to New York's 2,400, Washington's 1,579, Pennsylvania's 965, Wisconsin's 891, and a United States rate of 1,656.

Although the proposition above has aroused interest and been cited, any thorough utilization of our FBI fingerprint records to check it could produce much more fruitful sentence and parole policy propositions. This checking should strive to identify total system and long-term recidivism rates associated with various mixes of probation and parole, preferably using some specific types of offense and criminal record separately. It would attempt long-run cost-benefit assessments of various total criminal justice system policies, and perhaps make some differentiations for various regions of the country, or use metropolitan and rural areas separately. These would provide benchmarks, so that deviations from the predominant recidivism and re-

imprisonment rates associated with particular sentencing and parole policies in some localities could be investigated to determine what caused them to have atypical patterns. (For proposals on the development of a still more systemic assessment of rates, see Klein, *et al*, 1971.)

Many other propositions more relevant to specific practices now pursued might also be formulated, with summaries of the available evidence and argument about them. Because of this focusing of attention on them, they might be more carefully tested. Some of the immense variety of propositions of this sort which are possible in penology are presented below, without the imperfect but suggestive evidence that now supports them.

A. Promoting the isolation of inmates from each other fosters rehabilitation, when this is done by providing housing which facilitates an inmate's achievement of privacy when he or she desires it.

B. Inmate pressures on other inmates to avoid communication with staff varies directly with the extent to which there is an impersonal and authoritarian orientation of staff to inmates.

C. The more ritualistic and routinized the duties of an employee become in dealing with offenders, the more he is inclined to become authoritative and punitive toward them.

D. Custody grading systems impede rehabilitation if they provide freedom only for play activities and as reward only for avoiding disorder, rather than rewards more relevant to postrelease needs (e.g., more release money or more outside contacts) for self-regulation more relevant to post-release needs (e.g., education and training achievements).

E. Not training in vocational skills so much as habituation of inmates to regularity in constructive and rewarded employment, and anticriminal personal influences of work supervisors, are the main contribution of work in prison to reduction of recidivism rates.

F. Prison education is associated with reduced recidivism rates only when it is extensive enough to alter grade level attainment appreciably; small doses of prison education are associated with higher crime rates.

G. The maximum rehabilitative impact of wages paid for inmate work in prison occurs when wages may be earned at any kind of work, the rate of pay is a function of performance, and the amount spent while in prison and when first on parole, is considerably limited.

H. Reimprisonment rates are reduced most by release through a few months in a community residential center before parole, provided the offender is neither a first offender with strong family support nor a clearly habitual offender.

I. The most unfavorable postrelease residential arrangement, in terms of postrelease failure rates, is that in which the ex-prisoner lives alone.

The propositions above may or may not be valid; doubtless many of them could be improved by revision, and could be replaced or supplemented by still more useful formulations. The main point is that by reducing our most confident theories and beliefs to their central law-like propositions, we direct empirical studies to these generalizations and increase the prospect of their becoming knowledge. Scientific knowledge is theory which has been tested and found valid. Validity, however, is never absolute and final, without the possibility of qualification or even rejection. Instead, comments on validity are expressions regarding the degree of confidence which evidence has thus far inspired.

Carole H. Weiss observes, "Evaluators complain about many things, but their most common complaint is that their findings are ignored" (1972: 319). If the conclusions of evaluative research studies are focused on the validation or revision of items in a standard and widely available propositional inventory, it will be more difficult for administrators or for other researchers to ignore them.

Propositional inventories should point up issues, rather than restricting themselves to statements about which there is consensus. They should summarize all the principal argument and evidence on controversial matters in order to highlight contentions urgently in need of further test. This would promote what Cain and Hollister called "intentional experimentation," when they noted:

It is not unfair . . . to characterize the approach to social action programs . . . in the past as one of serial experimentation through program failure. A program is built around a single concept, eventually it is realized that it does not work, so the program is scrapped (or allowed to fade away) and a new program and concept is tried. Certainly serial experimentation through failure is the hard way to learn. An intentionally experimental approach would allow us to learn faster by trying alternative concepts *simultaneously* and would make it more likely that we could determine not only

that a particular concept failed, but also *why* it failed. (In Rossi and Williams, 1972:135.)

A final observation should be made about a possible limitation or disadvantage of a propositional inventory. As Kuhn (1962) pointed out, the major advances in scientific knowledge have not come merely from testing and sharpening prior propositions (which he calls "normal science"), but by making imaginative new kinds of formulations that prove valid (which he calls "scientific revolutions"). The theory of universal gravitation, the germ theory of disease, the theory of biological evolution, the theory of mass-energy convertability, and the theory of space-time relativity were all breakthroughs not so much because they were more precise tests of the hypotheses than being debated, but because they proposed more profound and fruitful hypotheses.

Conclusion

Evaluations too often are formulated in noncomparable terms, are reported in widely scattered publications, and consequently, provide only disconnected bits of knowledge that are non-cumulative. This problem has not been diminished as much by compiling huge collections of abstracts, or appointing committees to draft manuals of standards, as by publishing highly selected abstracts and periodic articles summarizing knowledge and theory on specific topics. An even greater cumulative benefit could result from the periodic compilation of carefully edited inventories of propositions on people-changing issues, each followed by a summary of the theory and research relevant to it. If we routinize evaluative research, and address it to testing items in a propositional inventory, yet do not make it so routine and so restricted to the inventory's items as to rule out imaginative new ideas and experiments, we shall maximize the effectiveness of people-changing endeavors.

REFERENCES

- Adams, Stuart. *Effectiveness of Interview Therapy with Older Youth Authority Wards: An Interim Evaluation of the PICO Project*. Research Report No. 20. Sacramento, CA: Youth Authority, 1961. Reprinted in Johnson et al, eds. (q.v. intra), 1970.
- . Measurement of effectiveness and efficiency in corrections. In: Glaser, Daniel, ed. *Handbook of Criminology*. Rand-McNally, forthcoming.
- Adams, Stuart; Heaton, Wanda S.; Monos, Dmitri; and Upchurch, Margaret. *Parole Performance Trends Among Community Treatment Center Releasees*. Research Report No. 7. Washington, D.C.: District of Columbia Department of Corrections, 1968.
- American Friends Service Committee. *Struggle for Justice*. New York: Hill and Wang, 1971.
- Babst, Dean V.; Gottfredson, Don M.; and Ballard, Jr., Kelly B. Comparison of multiple regression and configuration analysis techniques for developing base expectancy tables. *Journal of Research in Crime and Delinquency*, 5:72-80, January 1968.
- Babst, Dean V.; Koval, Mary; Neithercutt, Marcus G.; and Gottfredson, Don M. Relationship of time served to parole outcome for different classifications of burglars. *Journal of Research in Crime and Delinquency*, in press.
- Babst, Dean V.; Inciardi, James A.; and Jaman, Dorothy R. The uses of configural analysis in parole prediction research. *Canadian Journal of Criminology and Corrections*, 13:1-9, July 1971.
- Becker, Howard S. *Outsiders*. New York: Free Press, 1963.
- Beverly, Robert F. *Analysis of Parole Performance by Institution of Release*. Research Report No. 22. Sacramento, CA: Youth Authority, 1961.
- Bruce, Andrew A.; Burgess, Ernest W.; and Harno, Albert J. *The Workings of the Indeterminate-Sentence Law and the Parole System in Illinois*. Springfield, IL: Board of Parole, 1928.
- Campbell, Donald T. Reforms as experiments. *American Psychologist*, 24:409-429, April 1969. Reprinted in Weiss, 1971; and in revised form in Caro, 1971.
- Campbell, Donald T., and Stanley, Julian C. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally, 1966. Reprinted from Gage, N.L., ed. *Handbook of Research on Teaching*. Chicago: Rand-McNally, 1963.
- Cantor, Barbara, and Adams, Stuart. *The Cost of Correcting Youthful Offenders*. Research Report No. 6. Washington, D.C.: District of Columbia Department of Corrections, 1968.

- Caro, Francis G. *Readings in Evaluation Research*. New York: Russel Sage Foundation, 1971.
- Carter, Robert M. The presentence report and the decision-making process. *Journal of Research in Crime and Delinquency*, 4:203-211, July 1967. Reprinted in Carter and Wilkins (q.v. *intra*), 1970.
- Carter, Robert M., and Wilkins, Leslie T. *Probation and Parole*. New York: Wiley, 1970.
- Dean, Charles W., and Repucci, N. Dickon. Juvenile correctional institutions. In: Glaser, Daniel, ed. *Handbook of Criminology*. Rand-McNally, forthcoming.
- Dole, Vincent P., and Warner, Alan. Evaluation of narcotics treatment programs. *American Journal of Public Health*, 57:2000-2008, November 1967.
- Doshay, Lewis J. *The Boy Sex Offender and His Later Career*. New York: Grune Stratton, 1943.
- Empey, LaMar T., and Erickson, Maynard L. *The Provo Experiment*. Boston: Heath, 1972.
- Empey, LaMar T., and Lubeck, Steven G. *The Silverlake Experiment*. Chicago: Aldine-Atherton, 1971.
- England, Ralph. What is responsible for satisfactory probation and post-probation outcome? *Journal of Criminal Law, Criminology and Police Science*, 47:667-677, April 1957.
- Evjen, Victor H. Current thinking on parole prediction tables. *Crime and Delinquency*, 8:215-224, July 1962. Reprinted in Carter and Wilkins (q.v. *supra*), 1970.
- Fiedler, F. E., and Bass, A. R. *Delinquency, Confinement and Interpersonal Perception*. Technical Report No. 6. Urbana, IL: University of Illinois Group Effectiveness Laboratory, 1959.
- Ford Foundation. *Dealing With Drug Abuse*. New York: Praeger, 1972.
- Gearing, Frances R. "Evaluation of Methadone Maintenance Treatment for Heroin Addiction: A Progress Report." Paper presented at Epidemiology Section of the American Public Health Association, 1969.
- Glaser, Daniel. Testing correctional decisions. *Journal of Criminal Law, Criminology and Police Science*, 45:679-684, March-April 1955.
- . Released offender statistics: A proposal for a national program. *American Journal of Correction*, 19:15-17, 25, March-April 1957.
- . Prediction tables as accounting devices for judges and parole boards. *Crime and Delinquency*, 8:239-258, July 1962.
- . *The Effectiveness of a Prison Parole System*. Indianapolis: Bobbs-Merrill, 1964.
- . Correctional research: An elusive paradise. *Journal of Research in Crime and Delinquency*, 2:1-11, January 1965. Reprinted in Carter and Wilkins (q.v. *supra*), 1970; and Radzinowicz and Wolfgang (q.v. *infra*), 1971.
- . Automated research and correctional practice. *California Youth Authority Quarterly*, 18:24-31, Winter 1965.
- . National goals and indicators for the reduction of crime and delinquency. *Annals of the American Academy of Political and Social Science*, 371:104-126, May 1967.
- . *The Effectiveness of a Prison and Parole System*. Abridged edition. Indianapolis: Bobbs-Merrill, 1969.
- . *Social Deviance*. Chicago: Markham, 1971.

- . *Adult Crime and Social Policy*. Englewood Cliffs, NJ: Prentice-Hall, 1972.
- . ed., *Handbook of Criminology*. Chicago: Rand-McNally, forthcoming.
- Glaser, Daniel, and O'Leary, Vincent. *Personal Characteristics and Parole Outcome*. Washington, D.C.: U.S. Government Printing Office, 1966.
- Glueck, Sheldon, and Eleanor. *Five Hundred Criminal Careers*. New York: Knopf, 1930.
- . *Juvenile Delinquents Grown Up*. New York: Commonwealth Fund, 1940.
- . *Unraveling Juvenile Delinquency*. New York: Commonwealth Fund, 1950.
- . *Delinquents and Nondelinquents in Perspective*. Cambridge, MA: Harvard University Press, 1968.
- Goldenberg, I. Ira. *Build Me a Mountain: Youth, Poverty and the Creation of New Settings*. Cambridge, MA: M.I.T. Press, 1971.
- Grant, J. Douglas, and Marguerite Q. A group dynamics approach to the treatment of nonconformists in the Navy. *Annals of the American Academy of Political and Social Science*, 322:126-135, March 1959.
- Guttman, Evelyn S. *Effects of Short-Term Psychiatric Treatment on Boys in Two California Youth Authority Institutions*. Research Report No. 36. Sacramento, CA: Youth Authority, 1963.
- Hall, Reis H.; Milazzo, Mildred; and Posner, Judy. *A Descriptive and Comparative Study of Recidivism in Pre-Release Guidance Center Releasees*. Washington, D.C.: U.S. Bureau of Prisons, 1966.
- Harrison, Robert M., and Mueller, Paul F.C. *Clue Hunting About Group Counseling and Parole Outcome*. Research Report No. 11. Sacramento, CA: Department of Corrections, 1964.
- Havel, Joan, and Sulka, Elaine. *Special Intensive Parole Unit, Phase Three*. Research Report No. 3. Sacramento, CA: Department of Corrections, 1962.
- Huff, Darrell. *How to Lie With Statistics*. New York: Norton, 1954.
- Irwin, John. *The Felon*. Englewood Cliffs, NJ: Prentice-Hall, 1970.
- Jaman, Dorothy R., and Dickover, Robert N. *A Study of Parole Outcome as a Function of Time Served*. Research Report No. 35. Sacramento, CA: Department of Corrections, 1969.
- Jessness, Carl F. The Preston Typology Study: An experiment with differential treatment in an institution. *Journal of Research in Crime and Delinquency*, 8:38-52, January 1971.
- Joe, G. W.; Person, Jr., Philip; Sells, S. B.; and Retka, R. L. *An Evaluative Study of Methadone and Drug-Free Therapies for Opiate Addiction*. Preliminary draft. Fort Worth, TX: Institute of Behavioral Research, Texas Christian University, 1972.
- Johnson, Bertram M. *Parole Performance of the First Year's Releasees, Parole Research Project: Evaluation of Reduced Caseloads*. Research Report No. 27. Sacramento, CA: Youth Authority, 1962.
- Johnston, Norman; Savitz, Leonard; and Wolfgang, Marvin E., eds. *The Sociology of Punishment and Correction*. Second edition. New York: Wiley, 1970.
- Kassebaum, Gene; Ward, David A.; and Wilner, Daniel M. *Prison Treatment and Parole Survival*. New York: Wiley, 1971.
- Kish, Leslie. *Survey Sampling*. New York: Wiley, 1967.

- Klein, Malcom W.; Kobrin, Solomon; McEachern, A. W.; and Sigurdson, Herbert R. System rates: An approach to comprehensive criminal justice planning. *Crime and Delinquency*, 4:355-372, October 1971.
- Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.
- Law Enforcement Assistance Administration. *The St. Louis Detoxification and Diagnostic Evaluation Center*. Washington, D.C.: U.S. Government Printing Office, 1970.
- Leiberg, Leon G. *Project Crossroads*. Washington, D.C.: National Committee for Children and Youth, 1971.
- Lerman, Paul. Evaluative studies of institutions for delinquents. *Social Work*, 13:55-64, July 1968.
- . *Delinquency and Social Policy*. New York: Praeger, 1970.
- McCall, George J., and Simmons, J. L. *Identities and Interactions*. New York: Free Press, 1966.
- McGlothlin, William H.; Tabbush, Victor C.; Chambers, Carl D.; and Jamison, Kay. "Alternative Approaches to Opiate Addiction Control: Costs, Benefits and Potential." Los Angeles: Dept. of Psychology, UCLA. Report for Bureau of Narcotics and Dangerous Drugs, U.S. Department of Justice, 1972. (Mimeo.)
- McGrath, S. D.; O'Brien, P. F.; Power, F. J.; and Shea, J. R. Nicotinamide treatment of schizophrenia. *Schizophrenia Bulletin*, 5:74-76, Spring 1972.
- Mannheim, Hermann, and Wilkins, Leslie T. *Prediction Methods in Relation to Borstal Training*. London: Her Majesty's Stationery Office, 1955.
- May, Edgar. *Narcotic Addiction and Control in Great Britain*. New York: Ford Foundation, 1972.
- Meehl, Paul E., and Rosen, Albert. Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin*, 52:194-215, May 1955.
- Merton, Robert K. *Social Theory and Social Function*. Revised edition. New York: Free Press, 1957.
- Miles, Arthur P. The utility of case records in probation and parole. *Journal of Criminal Law, Criminology and Police Science*, 56:285-293, September 1965. Reprinted in Carter and Wilkins, 1970.
- Molof, Martin J. *Forestry Camp Study: Comparison of Recidivism Rates of Camp-Eligible Boys Randomly Assigned to Camps and to Institutional Programs*. Research Report No. 53. Sacramento, CA: Youth Authority, 1967.
- National Conference on Parole. *Parole in Principle and Practice*. New York: National Probation and Parole Association, 1957.
- Newman, J. Robert. *The Prediction of Recidivism for Misdemeanant Offenders Released from the Los Angeles County Jail*. Los Angeles: Public Systems Research Institute, University of Southern California, 1972.
- New York State Division for Youth. *Characteristics of Delinquent Youths at Various Stages of the Treatment Process*. Albany, NY: State Division for Youth, 1970.
- Ohlin, Lloyd E. *Selection for Parole*. New York: Russell Sage Foundation, 1951.
- . The routinization of correctional change. *Journal of Criminal Law and Criminology*, 45:400-411, Nov.-Dec. 1954.

- Palmer, Theodore B.; Neto, Virginia V.; Johns, Dennis A.; Turner, James K.; and Pearson, John W. *Community Treatment Project, Seventh Progress Report, Part 1*. CTP Research Report No. 9, Part 1. Sacramento, CA: Youth Authority, 1968.
- Postman, J. *Psychology in the Making*. New York: Knopf, 1962.
- Radzinowicz, Leon, and Wolfgang, Marvin E. *Crime and Justice*. Volume III, The Criminal in Confinement. New York: Basic Books, 1971.
- Robins, Lee N. *Deviant Children Grown Up*. Baltimore: Williams and Wilkins, 1966.
- Robins, Lee N., and Hill, Shirley Y. Assessing the contributions of family structure, class and peer groups to juvenile delinquency. *Journal of Criminal Law, Criminology and Police Science*, 57:325-331, September 1966.
- Robins, Lee N., and Murphy, George E. Drug use in a normal population of young Negro men. *American Journal of Public Health*, 57:1580-1596, September 1967.
- Roethlisberger, F. J., and Dickson, William J. *Management and the Worker*. Cambridge, MA: Harvard University Press, 1939.
- Rosenthal, Robert. *Experimental Effects in Behavioral Research*. New York: Appleton-Century-Crofts, 1966.
- Rossi, Peter H., and Williams, Walter. *Discussion Papers: Institute for Research on Poverty*. Madison: University of Wisconsin, 1969. Reprinted in Rossi, Peter H., and Williams, Walter, eds. *Evaluating Social Programs*. New York: Seminar Press, 1972.
- Ryder, Norman B. The cohort as a concept in the study of social change. *American Sociological Review*, 30:843-861, December 1965.
- Schur, Edwin M. *Labeling Deviant Behavior*. New York: Harper and Row, 1971.
- Seckel, Joachim P. *Experiments in Group Counseling at Two Youth Authority Institutions*. Research Report No. 46. Sacramento, CA: Youth Authority, 1965.
- Seckel, Joachim; Pond, Esther M.; and Davis, Carolyn B. *A Comparative Study of the Community Parole Center Program*. Research Report No. 63. Sacramento, CA: Youth Authority, 1973.
- Sellin, Thorsten. Recidivism and maturation. *National Probation and Parole Association Journal*, 4:241-250, July 1958.
- Simon, Frances H. *Prediction Methods in Criminology*. London: Her Majesty's Stationery Store, 1971.
- . Statistical methods of making prediction instruments. *Journal of Research in Crime and Delinquency*, 9:46-53, January 1972.
- Snow, Mary. Maturing out of narcotic addiction in New York City. *International Journal of the Addictions*, in press.
- Sonquist, John A. *Multivariate Model Building*. Ann Arbor, MI: Institute for Social Research, University of Michigan, 1970.
- Sonquist, John A., and Morgan, J. N. *The Detection of Interaction Effects: A Report on a Computer Program for the Selection of Optimal Combination of Exploratory Variables*. Monograph No. 35. Ann Arbor, MI: Institute of Social Research, University of Michigan, 1964.
- Spencer, Carol, and Berecochea, John E. *Recidivism Among Women Parolees: A Long-Term Survey*. Research Report No. 47. Sacramento, CA: Department of Corrections, 1972.
- Suziedelas, Antanas. *Program Participation*. Washington, D.C.: U.S. Bureau of Prisons, 1963.

- Vera Institute of Justice. *The Manhattan Court Employment Project*. New York: The Institute, 1970.
- Vocational Rehabilitation Administration. *Rehabilitating the Narcotic Addict: Report of a Conference*. Washington, D.C.: U.S. Government Printing Office, 1966.
- Warren, Marguerite Q. *The Community Treatment Project After 5 Years*. Sacramento, CA: Youth Authority, 1966.
- Webb, Eugene J.; Campbell, Donald T.; Schwartz, Richard D.; and Sechrest, Lee. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand-McNally, 1966.
- Weiss, Carol H. *Evaluating Action Programs*. Boston: Allyn and Bacon, 1972.
- Wenk, Ernst A.; Robison, James O.; and Smith, Gerald W. Can violence be predicted? *Crime and Delinquency*, 18:393-402, October 1972.
- Wilkins, Leslie T. *Social Deviance*. Englewood Cliffs, NJ: Prentice-Hall, 1965.
- . *Evaluation of Penal Measures*. New York: Random House, 1969.
- Wilkins, Leslie T., and MacNaughton-Smith, P. New prediction and classification methods in criminology. *Journal of Research in Crime and Delinquency*, 1:19-32, January 1964. Reprinted in Carter and Wilkins (q.v. supra), 1970; Johnston, et al. (q.v. supra), 1970.
- Winick, Charles. Maturing out of narcotic addiction. *United Nations Bulletin on Narcotics*, 14 (1):1-7, 1962.
- Wolfgang, Marvin E.; Figlio, Robert M.; and Sellin, Thorsten. *Delinquency in a Birth Cohort*. Chicago: University of Chicago Press, 1972.
- Zeisel, Hans. *Say it With Figures*. Fifth edition. New York: Harper and Row, 1968.

APPENDIX A: A PRECODED FORM AS INITIAL REPLACEMENT FOR NARRATIVE REPORT

New York State
Narcotic Addiction Control Commission
TR-17 (3/70)

Appendix A: Page 1

INTRAMURAL CASE SUMMARY
(Type or Print Only)

See Instructions on Form TR-17.1

1. Subject's Name: Last		First	Middle	2. (1-3)	3. Current Facility (4-6)	4. NACC No. (7-13)
5. No. Prior Releases This Commitment (14) 0 None 1 One 2 Two 3 Three or more				6. Commission Determination Code (15)		
7. Purpose of Summary (16)						
1 Assign to Community Based Service		2 Recommended for NACC Methadone Maintenance		5 3-Month Review		
3 Interfacility Transfer		4 Affiliated Program		6 12-Month Review		
				7 Extended Pass to Aftercare		
				8 Other, specify _____		

Reason for Transfer

- | | |
|-------------------------------------|-------------------------|
| 0 Not applicable, not a transfer | 5 Attempted absence |
| 1 Resident does not accept program | 6 Returned from absence |
| 2 Resident physically abusive | 7 Positive urine |
| 3 Move to facility nearer to family | 8 Other, specify _____ |
| 4 Transferred to "opening" facility | |

Program Participation in Intramural Facilities

(Check one on each line)

Degree of Participation

Intramural Programs	Type of Participation	Degree of Participation						
		Does Not Apply	Did Not Participate	Below Average		Average		Above Average
		9	0	1	2	3	4	5
8. Individual Counseling	Involvement							(17)
	Attendance							(18)
9. Group Counseling	Involvement							(19)
	Attendance							(20)

10. Total Number of Months Participate in Group Counseling This stay: (21-22)	11. Group Counseling Level at this Time (23)
_____	0 Not stratified therapy
	1 Beginning "C" Group
	2 Intermediate "B" Group
	3 Final "A" Group or Pass Group

Comments on Above

INTRAMURAL CASE SUMMARY

Subject's Name: Last, First Initial		NACC No.
<p>12. <u>Resident's Emotional State</u> (24)</p> <p>1 Given to radical changes in mood 2 Fairly stable 3 Stable 4 Constricted</p>	<p>13. <u>Amount of Correspondence Initiated by Resident</u> (25)</p> <p>1 Below Average 2 Average 3 Above Average</p>	
<p>14. <u>Resident Family Contacts While in Facility</u> (phone, mail or visit) (26)</p> <p>1 Frequent 2 Moderate 3 Infrequent 4 None 5 No family</p>	<p>15. <u>Behavior on Pass or Work Release</u> (27)</p> <p>0 Never released 1 Very dependable 2 Fairly dependable 3 Not dependable</p>	
<p>16. <u>Ability to Work with Others</u> (28)</p> <p>1 Overly shy and frightened of others 2 Shy, but can work with others if called to 3 Works well with others 4 Overbearing, but can accommodate 5 Very difficult to work with - cannot accommodate</p>	<p>17. <u>Evidence of Psychiatric Disorder</u> (29)</p> <p>0 None 1 Mild 2 Serious</p> <p>Psychiatric diagnosis _____ _____ _____</p>	
<p>18. <u>Quality of Resident's Reactions</u> (30)</p> <p>1 Overreactive 2 Somewhat appropriate to situation 3 Reactions usually appropriate to situation 4 Somewhat appropriate but emotional reactions flat 5 Emotional reactions flat</p>	<p>19. <u>Quality of Resident's Relationships</u>(31)</p> <p>1 Unable to trust others or form friendships 2 Does not trust others, but is attempting to form friendships 3 Stays in clique friendships 4 Able to form trusting relationships outside of clique 5 Trusts and makes friendships with everyone who will listen to him</p>	

Comments on Above

INTRAMURAL CASE SUMMARY

Subject's Name: Last	First Initial	NACC No.
----------------------	---------------	----------

20. Understanding of Past Experiences (32)

- | | |
|---|--|
| 1 Blames self unreasonable amount | 4 Beginning to understand but still projecting |
| 2 Beginning to understand but still tending to blame self | 5 Projects blame unreasonable amount |
| 3 Understands role of self and environment in past | |

PSYCHO-SEXUAL FACTORS

21. Ability to Communicate About Sex (33)

- 1 Feels free to discuss sex
- 2 Blocks discussions about sex

22. Sublimation (34)

- 1 Sublimates sexual energies in constructive channels
- 2 Sublimates sexual energies in negative channels
- 3 Unable to sublimate sexual energies

23. Homosexuality (35)

- 0 None, 1 Overt 2 Covert

24. External Sexual Image (36)

Effeminate		Masculine		
1	2	3	4	5

25. Acceptance of Sexual Image (37)

- 1 Functional
- 2 Changing positively
- 3 Changing negatively
- 4 Clearly non-functional

Relationship and Responsibility	Recent Rating					Progress Since Adm.		
	Below Average		Above Average			No Prog.	Some Prog.	Definite Progress
	1	2	3	4	5	1	2	3
26. <u>Interpersonal relationships-staff</u> (38)								(41)
27. <u>Interpersonal relationships-peer</u> (39)								(42)
28. <u>Responsibility for person and property</u> (40)								(43)
29. <u>Participation in Special Treatment Programs in Facility</u> (44)								

- 1 Participates in special voluntary groups
- 2 Is in a special ward (e.g., for young people)
- 3 Resident in special leadership position
- 4 Other special treatment

Comments on Above

INTRAMURAL CASE SUMMARY

<u>Subject's Name:</u> Last	First Initial	NACC No.
-----------------------------	---------------	----------

PRE-RELEASE PLAN

(To be completed only for those considered for release)

<u>30. Attitude towards Aftercare (45)</u> 1 Willing to cooperate and try abstinence 2 Mixed feelings about aftercare 3 Clearly negative and hostile	<u>31. Work Plan upon Release (46)</u> 1 No definite job available 2 Plans to attend school 3 Has job, not related to center training 4 Has job, related to center training 5 Other, specify _____
---	---

<u>32. Living Arrangements upon Release (47)</u> 0 Indefinite 1 Alone 2 With Friends 3 With Parents	4 With spouse 5 At NACC Community Based Center 6 Other, specify _____
---	---

<u>33. Circle Major Problem Faced at Release (48-49)</u> 1 Outstanding warrant 2 Needs residence 3 Needs financial assistance 4 Hostility of family 5 Lack of family interest	6 Continued need for counseling 7 Continued need for voc/ed training 8 Continued need for special med/dent. care 9 Continued need for psychiatric treatment 10 Other, specify _____
--	---

<u>34. Estimate Probability of Absence or Being Returned to Intramural Status Within One Year after Release (50)</u> 1 Less than 25% 2 25-49% chance 3 50-74% chance 4 75% or more chance

Comments on Above

INTRAMURAL CASE SUMMARY

<u>Subject's Name: Last</u>		<u>First Initial</u>	<u>NACC No.</u>
<u>Unusual Incidents This Stay</u> (Circle those appropriate to case)			
<u>35. Physically Abusive (51)</u>	<u>36. Segregation (52)</u>	<u>37. Disturbances (53)</u>	
0 None 1 With Other residents 2 With Counselors 3 With Officers 4 To self	0 None 1 Once 2 Twice 3 Three or more times	0 None 1 Provoked Grp Acting Out 2 Participated in Group Acting Out	
<u>38. Abscondence (54)</u>		<u>39. Abscondence Attempts (53)</u>	
0 None 1 First 2 Second 3 Third		0 None 1 First 2 Second 3 Three or more	
<u>40. How Returned from Abscondence (56)</u>			
0 Does not apply 1 By own volition 2 All by authorities 3 Once by own volition, all others by authorities 4 Twice by own volition, all others by authorities			
<u>41. Current Physical Health (57)</u>		<u>42. Verbal Skills (58)</u>	
1 No serious health problem 2 Handicapped, but can work 3 Handicapped, cannot work Medical Diagnosis, if applicable		1 Is comfortable with square talk 2 Beginning to drop street jargon, but still uncomfortable in square talk 3 Generally unwilling to abandon jargon	
<u>Comments on Above</u>			

INTRAMURAL CASE SUMMARY

Subject's Name: Last	First Initial	NACC No.
----------------------	---------------	----------

43. High School Diploma or College Course Work (59)	44. I.Q. Rating (60)
1 High School Equivalency from NACC 2 High School Equivalency or diploma from elsewhere 3 College correspondence course from NACC 4 College courses elsewhere	1 Above Average 2 Average 3 Below Average 4 Not Tested Test(s) Used _____

Program Participation in Intramural Facilities

Intramural Programs	Type of Participation	Degree of Participation (check one on each line)							
		Does Not Apply	Did Not Partici-pate	Below Average			Average		Above Average
		9	0	1	2	3	4	5	
45. Academic Education	Attendance								(61)
	Progress								(62)
46. Vocational Training	Attendance								(63)
	Progress								(64)

Grade Achievement Level

47. (Circle one in each column)	Reading		Arithmetic		Last Grade Completed (69)
	At (65) Admission	Recent Retest (66)	At (67) Admission	Recent Retest (68)	
Not tested	0	0	0	0	0
3rd grade or less (0-3.4)	1	1	1	1	1
4th grade (3.5-4.4)	2	2	2	2	2
5th grade (4.5-5.4)	3	3	3	3	3
6th grade (5.5-6.4)	4	4	4	4	4
7th grade (6.5-7.4)	5	5	5	5	5
8th grade (7.5-8.4)	6	6	6	6	6
9th grade (8.5-9.4)	7	7	7	7	7
10th grade (9.5-10.4)	8	8	8	8	8
11th grade or more (10.5)	9	9	9	9	9

Comments on Above

INTRAMURAL CASE SUMMARY

Subject's Name: Last First Initial NACC No.

Program Participation in Intramural Facilities

Intramural Programs Participation		(Check one on each line) Degree of Participation							
		Does Not Apply	Did Not Participate	Below Average			Above Average		
				1	2	3	4	5	
48. Therapeutic Recreation	Attendance								(70)
	Involvement								(71)
49. Takes responsibility for Work Assignment									(72)
50. Uses free time constructively									(73)

Comments on Above

Completed By: _____ (74-76) _____ Date _____
 Approved: _____ Associate N.R.C. _____ Date _____
 Approved: _____ Director _____ (77-80) _____ Date _____

NEW YORK STATE
NARCOTIC ADDICTION CONTROL COMMISSION
REPORT OF ADMISSION TO TREATMENT AND/OR SERVICE PROGRAM

TYPE OF PROGRAM	
<input type="checkbox"/>	DEMONSTRATION
<input type="checkbox"/>	METHADONE
<input type="checkbox"/>	YOUTHFUL DRUG ABUSE

AGENCY		AGENCY CODE	
1. I.D. NO.:		2. Subject's Name (Last, First, Middle)	
5. Subject's Address (Street, Number, City, County, State)		3. Sex: Male <input type="checkbox"/> Female <input type="checkbox"/>	
8. Place of Birth (State or Country)		4. Date Admitted: ___/___/___	
50 <input type="checkbox"/> New York State 54 <input type="checkbox"/> Puerto Rico		6. First Admissions: 1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No	
9. Father - Country of Birth		7. Date of Birth: ___/___/___	
49 <input type="checkbox"/> U.S. 54 <input type="checkbox"/> Puerto Rico		49 <input type="checkbox"/> U.S. 54 <input type="checkbox"/> Puerto Rico	
Other (Specify) _____ Code _____		Other (Specify) _____ Code _____	
11. Maiden Name:		12. U.S. Citizen: 1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No	
14. Race: 1 <input type="checkbox"/> White 3 <input type="checkbox"/> Other 2 <input type="checkbox"/> Black		13. U.S. Veterans: 1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No	
15. Puerto Rican: 1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No		If yes, Separation or Discharge Date: ___/___/___	
16. Marital Status: 1 <input type="checkbox"/> Single 2 <input type="checkbox"/> Married 3 <input type="checkbox"/> Widowed 4 <input type="checkbox"/> Separated 5 <input type="checkbox"/> Divorced 6 <input type="checkbox"/> Unascertained 7 <input type="checkbox"/> Common-law		Type of Discharge: 1 <input type="checkbox"/> Separated, not discharged 2 <input type="checkbox"/> Honorable 3 <input type="checkbox"/> Dishonorable 4 <input type="checkbox"/> Bad Conduct 5 <input type="checkbox"/> Undesirable 6 <input type="checkbox"/> General 7 <input type="checkbox"/> Other	
17. Education (Check highest grade completed)		18. Religion:	
0 <input type="checkbox"/> None 1 <input type="checkbox"/> Grades 1-6 2 <input type="checkbox"/> Grade 7 3 <input type="checkbox"/> Grade 8 4 <input type="checkbox"/> Grade 9 5 <input type="checkbox"/> Grade 10 6 <input type="checkbox"/> Grade 11		1 <input type="checkbox"/> Protestant 2 <input type="checkbox"/> Roman Catholic 3 <input type="checkbox"/> Jewish 4 <input type="checkbox"/> Greek Orthodox 5 <input type="checkbox"/> Islam 6 <input type="checkbox"/> Other 7 <input type="checkbox"/> None 8 <input type="checkbox"/> Unascert.	
7 <input type="checkbox"/> High School Graduate 8 <input type="checkbox"/> Some College 9 <input type="checkbox"/> College Graduate A <input type="checkbox"/> Advanced degree courses B <input type="checkbox"/> Ungraded classes C <input type="checkbox"/> Unascertained		20. Spouse's First Name: _____	
19. Occupations: _____ Code _____		20. Social Security Number: _____	
22. Referred By: 1 <input type="checkbox"/> Self 2 <input type="checkbox"/> Relative 3 <input type="checkbox"/> Friend 4 <input type="checkbox"/> Physician 5 <input type="checkbox"/> Private Agency 6 <input type="checkbox"/> Public Agency		Name of Public or Private Agency: _____	
23. Household Composition (Mark all which apply)			
Subject lives: 1 <input type="checkbox"/> Alone 2 <input type="checkbox"/> With Parents 4 <input type="checkbox"/> In an Institution		1 <input type="checkbox"/> With Spouse 2 <input type="checkbox"/> With Children 4 <input type="checkbox"/> With Other Relatives	
2 <input type="checkbox"/> With Male Friend 3 <input type="checkbox"/> With Female Friend 4 <input type="checkbox"/> With Others			
24. If answer to #23 above is other than "Alone" or "In an institution": Do any of these people use drugs? 1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No			
25. If yes, who? (Mark all which apply and indicate if the person is currently in treatment)			
1 <input type="checkbox"/> Parent(s) 2 <input type="checkbox"/> Spouse		3 <input type="checkbox"/> One or more children 4 <input type="checkbox"/> Other relatives	
5 <input type="checkbox"/> Male Friend 6 <input type="checkbox"/> Female Friend 7 <input type="checkbox"/> Other			
26. Previous treatment for drug use received throughout drug use history:			
1 <input type="checkbox"/> N.A.C.C. 2 <input type="checkbox"/> NYS M.H. Hospital 4 <input type="checkbox"/> Fed. Narc. Hospital		1 <input type="checkbox"/> V.A. Hospital 2 <input type="checkbox"/> Gen. Hospital 4 <input type="checkbox"/> Pvt. Hospital	
3 <input type="checkbox"/> Outpatient Clinic 2 <input type="checkbox"/> Pvt. Physician 4 <input type="checkbox"/> Other (Specify below) 8 <input type="checkbox"/> None			
27. Age first illegally used any drug? _____		28. Since onset of drug use, what was the longest period of voluntary abstinence? _____	
What was first drug? _____		What was drug? _____	
29. Age first illegally used any narcotic drug? _____			
30. Total number of detoxifications in hospitals, jails, lockups, etc. _____		31. Total detoxifications not in hospitals, jails, lockups, etc. _____	

32. Drug use prior to admission: Indicate frequency by placing a check in appropriate box beside drug name:

Primary Drug used: Heroin Other (specify) _____

	FREQUENCY OF USE DURING LAST THREE MONTHS		
	DAILY	LESS THAN DAILY	NONE
Heroin	1	2	0
Methadone	1	2	0
Codaine	1	2	0
Other Narcotics	1	2	0
Cocaine	1	2	0
Barbiturates and other Sedatives	1	2	0
Amphetamines and similar agents	1	2	0
Psychedellics	1	2	0
Marijuana/Hazhish	1	2	0
Other (Specify)	1	2	0

33. Total estimated cost to subject per day: (During last month on street)

a. Cost of primary drug \$ _____ b. Total cost of all drugs used \$ _____
(Most frequent)

34. Arrest History: Ever arrested? 1 Yes 2 No

If yes, a. Total number _____

b. Total number for violations of the drug laws _____

c. Age first arrested _____

35. Subject will be participating in your program as:

1 A resident or in-patient

2 An Outpatient

36. Mother's first name: _____

37. Alcohol: 1 No problem 2 Previous problem 3 Current Problem

Define problem (any indication of drunk arrests, loss of time on job or marital discord due to excessive drinking):

38. Work History:

0 Never worked

1 Worked in past

Longest period in ANY job months

Length of time since last employment months

39. Employment Pattern (One month prior to Admission):

1 Employed more than 30 hours per week (include self-employed).

2 Employed up to 30 hours per week (include self-employed).

3 Not known to be employed

4 In school or training program

5 Housewife

6 Incapacitated, specify _____

7 In institution, specify _____

If employed, present income \$ _____ (weekly)

40. Welfare History:

1 Currently on

2 Previously on

3 Never on

PREPARED BY: _____ TITLE: _____

DATE: _____

AFTERCARE CHRONOLOGICAL SUMMARY

REASON FOR SUBMISSION OF REPORT

1	<input type="checkbox"/>	Quarterly Summary	2
2	<input type="checkbox"/>	Discharge (requires TR-27)	2
3	<input type="checkbox"/>	Death (" ")	2
4	<input type="checkbox"/>	Return to residency or Police Custody	4
5	<input type="checkbox"/>	Change of NAO (requires TR-27 or TR-105)	1
6	<input type="checkbox"/>	Report of Abscondence (requires TR-27)	2

No. of Copies

Facility _____

Report Period: from _____ to _____

NAO _____

NAME	Last	First	Middle	SEX	NAOC I.D. No.	Birthdate	Date Cert.	Date Exp.	Date Released to A/C	Rec. Released from
CURRENT AFTERCARE STATUS										
MHL #		No. Arrests Prior to Cert.		No. Arrests Since Cert.		CURRENT LEGAL MARITAL STATUS Check one				
1 <input type="checkbox"/>		2 <input type="checkbox"/>		3 <input type="checkbox"/>		4 <input type="checkbox"/>		5 <input type="checkbox"/>		6 <input type="checkbox"/>
Single		Married		Widowed		Separated		Divorced		Unascertained

1. CURRENT ADDRESS: Give name of institution if applicable
 Street Address _____ Apt. No. _____ City _____ County _____ State _____ Zip _____ Telephone No. _____

Number of different residences during this report period _____

2. CURRENT RESIDENTIAL SITUATION: Check all applicable

Client Lives:	1 <input type="checkbox"/>	Alone	1 <input type="checkbox"/>	With Spouse	1 <input type="checkbox"/>	With Male Friend(s)
	2 <input type="checkbox"/>	With Parents	2 <input type="checkbox"/>	With Children	2 <input type="checkbox"/>	With Female Friend(s)
	4 <input type="checkbox"/>	In an Institution	4 <input type="checkbox"/>	With Other Relatives	4 <input type="checkbox"/>	Transient or Unstable

Give name of individual(s) with whom resides _____

3. CONTACTS (No Phone Calls): _____ No. of Visits _____

Dates of Office Visits with Client _____

Dates of Office Visits with Other Persons eg. Family _____

Dates of Field Visits to Home of Client _____

Specify Other Field Visits and Dates _____

4. REPORTING RECORD: Check one on each line

Present Schedule 1 Weekly 2 Bi-Weekly 3 Monthly 4 Other (specify) _____

Attendance 1 All 2 More than half 3 Less than half 4 Very few

Dates of unexcused failures to report: _____ TOTAL _____

5. TYPES OF STATUS DURING REPORT PERIOD Check all applicable

1 <input type="checkbox"/>	Field Service	1 <input type="checkbox"/>	Halfway House	1 <input type="checkbox"/>	Outside Methadone Maintenance	_____	Name of Program _____
2 <input type="checkbox"/>	Community Care	2 <input type="checkbox"/>	NAOC Methadone Maintenance	2 <input type="checkbox"/>	Outside Abstinence Program	_____	_____
							4 <input type="checkbox"/> Other Chemotherapy _____

6. TREATMENT INVOLVEMENT: For each modality, indicate extent of involvement in all programs by the following code:

0 = None 1 = Less than once a month 2 = Once a month or more but less than once a week
 3 = Once or twice a week 4 = Three or more times a week 5 = Not applicable

PROGRAM: 1 _____ Individual Counseling 2 _____ Group Therapy 3 _____ Encounter Group 4 _____ Psychotherapy
 5 _____ Academic Education 6 _____ Vocational Guidance & Training 7 _____ Music & Art 8 _____ Recreation
 9 Other Specify _____

7. CURRENT SELF-SUFFICIENCY: Check Main Support

1 Supports Self Only 2 Supports Self & Others - give number of persons dependent on client including self _____

3 Supported by Parents - give parent's total net weekly income \$ _____. Give total number of persons dependent on that income including the wage earner(s) _____

4 Supported by Spouse - give spouse's total net weekly income \$ _____. Give total number of persons dependent on that income including spouse _____

5 Public Assistance 6 Unemployment Insurance 7 Other Specify _____

8. CURRENT EMPLOYMENT SITUATION:

Was client referred to the placement officer during this report period 1 Yes 2 No 3 Not Applicable Present Employer's Name _____

Was client employed this period 1 Yes 2 No Address _____

Was employment verified 1 Yes 2 No Job Title _____

If yes, specify source _____

Number of jobs held this report period _____ Dates of last employment: From _____ To _____

Is client's status known to employer 1 Yes 2 No

Number of hours worked last week _____ Net weekly salary \$ _____ Hourly wage \$ _____

NAOC TR-195
6/72

9. IF CLIENT DID NOT WORK MORE THAN 30 HOURS LAST WEEK INDICATE REASON: Check one

- | | | |
|---|---|---|
| 1 <input type="checkbox"/> In school | 7 <input type="checkbox"/> Did not report to work | 13 <input type="checkbox"/> In training program |
| 2 <input type="checkbox"/> Housewife | 8 <input type="checkbox"/> Did not seek employment | 14 <input type="checkbox"/> No reason indicated |
| 3 <input type="checkbox"/> Incapacitated | 9 <input type="checkbox"/> Laid off due to lack of work | 15 <input type="checkbox"/> Other-specify _____ |
| 4 <input type="checkbox"/> Temporary return to residence | 10 <input type="checkbox"/> In aftercare less than one week | 16 <input type="checkbox"/> Quit |
| 5 <input type="checkbox"/> Unknown because of failure to report | 11 <input type="checkbox"/> In daily or residential treatment program | 17 <input type="checkbox"/> Was fired |
| 6 <input type="checkbox"/> Unsuccessful in finding employment | 12 <input type="checkbox"/> Worked less than 30 hours or had no fixed hours (incl. self-employed) | |

10. ARREST DETAILS THIS PERIOD:

DATE of ARREST	COUNTY	CHARGE (Final If Known)	DISPOSITION OR ADJOURNMENT DATE	DETAINER WARRANT DATE	DESCRIPTION OF OFFENSE (From Arresting Officer or Court Papers)

11. DRUG USE DURING THIS PERIOD: Each drug should be coded for frequency of use: 0=None 1=Less than once a week 2=Once a week 3=Several times a week 4=Daily; If taken by doctor's prescription, check applicable box.

Drug	Frequency	MD Rx	Drug	Frequency	MD Rx
Heroin			Marijuana/Hashish		
Medadone			Cocaine		
Psychotogens (LSD, Etc.)			Alcohol Intoxication		
Other Opiates (specify)			Barbiturates (specify)		
Amphetamines, (specify)			Other drugs (specify)		

12. INDICATE BASIS FOR DETERMINATION OF CURRENT DRUG USE: Check all applicable
 Urine 1 Yes 2 No Physical Examination 1 Yes 2 No Interview 1 Yes 2 No

13. DATES OF ANY TEMPORARY RETURNS THIS PERIOD AND REASON:

SUMMARY OF AFTERCARE SUPERVISION -- IN PARAGRAPH FORM PROVIDE CONCISE CASE SUMMARY IDENTIFYING PROBLEMS, PROGRESS DURING REPORT PERIOD, FAMILY RELATIONSHIPS, COOPERATION WITH NAO, ASSOCIATES, TREATMENT GOALS, ETC., ALSO LIST ALL ADDRESSES WHERE CLIENT HAS LIVED DURING THIS REPORT PERIOD. (Please type or print - use separate sheet if necessary)

NAO _____ Social Security Number _____ Date _____

SR, NAO _____ Date _____

IF CLIENT IS LOST TO CONTACT AT THE TIME OF THIS REPORT, GIVE DATE OF CHANGE OF STATUS _____

IF CLIENT IS BEING DISCHARGED, GIVE DATE ON WHICH HE IS TERMINATED AND ATTACH FORM TR-1950 _____

TR-195
6/72

APPENDIX C: FORMS FROM THE U.S. BUREAU OF PRISONS RAPS-2 PROJECT

Appendix C: Sheet 1

PROGRAM ANALYSIS SHEET													
1. REGISTER NUMBER			2. NAME (LAST, FIRST, MIDDLE INITIAL)				3. EFFECTIVE DATE						
4. INSTITUTION NAME			5. INSTITUTION CODE				6. NEXT REVIEW DATE						
8. RATING		7. AGE		8. PRIOR COMMITMENTS		9. SENTENCE		10. PRIORITY		11. CUSTODY ASSIGNED			
IF FIRST ENTER 1		IF UNDER 30 ENTER 1		IF NONE ENTER 0		IF FJDA, YCA, OR NARA ENTER		IF FIRST ENTER 1		IF MINIMUM ENTER 1			
IF SECOND ENTER 2		IF 30 TO 45 ENTER 2		IF ONE ENTER 1		IF 18 MONTHS TO 15 YEARS OR 4208(A)(1) OR (A)(2) ENTER 2		IF SECOND ENTER 2		IF MEDIUM ENTER 2			
IF THIRD ENTER 3		IF OVER 45 ENTER 3		IF OVER ONE ENTER 2		IF LESS THAN 18 MONTHS OR OVER 15 YEARS ENTER 3		IF THIRD ENTER 3		IF CLOSE ENTER 3			
						IF EXCLUDABLE OFFENSE ENTER 4 (INCOME TAX, IMMIGRATION LAWS, LIQUOR LAW, SELECTIVE SERVICE ACT, SITUATION BREACH OF TRUST)							
<p>For lines 12-21, enter the appropriate NEED LEVEL and in the boxes marked activity, enter the activities which are planned for the individual. If the activity which is planned has a constraint the constrain must be entered in the box marked constraint to the right of the activity.</p>													
		NEED LEVEL		ACTIVITIES PLANNED				CONSTRAINTS PREVENTING PLAN					
		0 - NONE 1 - LOW 2 - HIGH		49 - EDUCATION - ABE 50 - EDUCATION - GED 51 - RECREATION 52 - VOCATIONAL TRAINING 53 - ON-THE-JOB TRAINING 54 - INDUSTRY 55 - PSYCHOTHERAPY (INDIV.) 56 - PSYCHOTHERAPY (GROUP)				57 - COUNSELING (INDIV.) 58 - COUNSELING (GROUP) 59 - CORRECTIONAL COUNSELING 60 - HEALTH SERVICES 61 - VOLUNTARY GROUPS 62 - WORK RELEASE 63 - STUDY RELEASE 64 - GENERAL MAINTENANCE 65 - CTC 66 - OTHER _____					
				67 - CUSTODY REASONS 68 - LACK PROGRAM 69 - PRIORITY 70 - PROGRAM FILLED 71 - TIME TOO SHORT 72 - TEMPORARILY CLOSED 73 - UNQUALIFIED 74 - OTHER _____									
CORRECTIONAL FACTORS				ACT 1	CNST 1	ACT 2	CNST 2	ACT 3	CNST 3	ACT 4	CNST 4	ACT 5	CNST 5
ENVIRONMENT	12. ECONOMIC STATUS												
	13. FAMILY CONDITIONS												
HEALTH	14. MENTAL HEALTH												
	15. PHYSICAL HEALTH												
SKILLS	16. EDUCATIONAL												
	17. VOCATIONAL												
CHARACTER	18. SELF-CONTROL												
	19. INTER-PERS. RELATIONS												
TRAITS	20. STANDARDS & VALUES												
	21. ASPIRATIONS												

BP-6.1 (NOV.71) Mechanical

BUREAU OF PRISONS
INMATE ACTIVITY REPORT

INSTRUCTIONS

To use as an ENROLLMENT form ----- Complete items 1-10 ONLY.
 To use as a COMPLETION form ----- Complete items 1-7 AND items 11, 12, and 13.
 To use as a WITHDRAWAL form ----- Complete items 7-7 AND items 11, 12, 13, and 14.

COMPLETE ALL ITEMS IN THIS SECTION.	1. REGISTER NUMBER			
	2. INMATE NAME (LAST, FIRST, MIDDLE)			
	3. INSTITUTION CODE (E.G. ATLANTA IS 131)			
	4. TYPE OF REPORT 2=ENROLLMENT 3=COMPLETION 4=WITHDRAWAL			
	5. DATE INMATE ENROLLED IN THIS ACTIVITY (MONTH, DAY, YEAR)			
	6. ACTIVITY NUMBER			
	49=EDUCATION--ARE 50=EDUCATION-GED AND OTHER 51=RECREATION 52=VOCATIONAL TRAINING 53=ON-THE-JOB TRAINING 54=INDUSTRY 55=PSYCHOTHERAPY (INDIV.) 56=PSYCHOTHERAPY (GROUP) 57=COUNSELING (INDIV.)		58=COUNSELING (GROUP) 59=CORRECTIONAL COUNSELING 60=HEALTH SERVICES 61=VOLUNTARY GROUP 62=WORK RELEASE 63=STUDY RELEASE 64=GENERAL MAINTENANCE 65=OTC 66=OTHER _____	
7. DEPARTMENT RESPONSIBLE FOR THIS ACTIVITY				
01=EDUCATION 02=CORRECTIONAL SERVICES 03=CASE MANAGEMENT 04=CHAPLAINS 05=MEDICAL 06=BUSINESS		07=NARA 08=DRUG ABUSE PROGRAM 09=MECHANICAL SERVICES 10=INDUSTRY 11=COMMUNITY VOLUNTEERS 12=FOOD SERVICES 13=OTHER _____		
COMPLETE THIS SECTION FOR ENROLLMENTS ONLY	8. PROGRAM TITLE			
	9. COUNSELOR STATUS		1=FULL-TIME	2=PART-TIME 3=OTHER
	10. CUSTODY AT ENROLLMENT		1-MINIMUM	2-MEDIUM 3-CLOSE
	11. DATE OF COMPLETION OR WITHDRAWAL (MONTH, DAY, YEAR)			
COMPLETE THIS SECTION FOR COMPLETIONS OR WITHDRAWALS ONLY.	12. CUSTODY AT COMPLETION OR WITHDRAWAL		1-MINIMUM	2-MEDIUM 3-CLOSE
	13. TOTAL INMATE HOURS AND MINUTES INVOLVED IN ACTIVITY			
	14. IF WITHDRAWAL, INDICATE REASON			
	1=RELEASED 2=TRANSFERRED 3=PROGRAM CHANGE 4=INMATE REQUEST		5=PROGRAM DISCONTINUED 6=CONTROL PURPOSES 7=INSTITUTIONAL NEEDS 8=OTHER _____	

BP 6-2 (NOV71) MECHANICAL

DHEW Publication No. (HSM) 73-9123
Printed 1973

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