Journal of Evaluation. Or a serious drop in clients from a particular ethnic group may result in the administrator of a program immediately replacing the director of professional services, whereas the evaluator's reaction may be to do a study to determine why the drop occurred. As with all relations between program staff and evaluators in general, negotiation of these matters is essential.

A warning: There are many aspects of program management and administration (such as complying with tax regulations and employment laws or negotiating union contracts) that few evaluators have any special competence to assess. In fact, evaluators trained in social science disciplines and (especially) those primarily involved in academic careers may be unqualified to manage anything. It is wise to keep in mind that the evaluator's role, even when sharing information from an MIS, is not to join the administrators in the running of the organization.

In the remainder of this chapter, we concentrate on the concepts and methods pertinent to monitoring program process and program outcome. It is in these areas that the competencies of persons trained in social research are most relevant. Because most program monitoring approaches emphasize process information, we give it especial attention by separately discussing the service utilization component and the organizational component of program process, drawing on the distinctions we have used for defining program theory.

MONITORING SERVICE UTILIZATION

In Chapter 4, we discussed how essential it is to define target populations carefully in planning, designing, and implementing programs. But, having done so, it is also important to know the extent to which the intended targets actually receive program services. Target participation concerns both program managers and sponsors. Managing a project effectively requires that target participation be kept at an acceptable level and corrective action be taken if it falls below that level. From the viewpoint of program sponsors, target participation is a key measure of a program's vitality and the demand for its services.

Monitoring of service utilization is particularly critical for interventions in which program participation is voluntary or in which participants must learn new procedures, change their habits, or take instruction. For example, community mental health centers designed to provide a broad range of services often fail to attract a significant proportion of those persons who may benefit from their services. Even patients who have been recently discharged from psychiatric hospitals and encouraged to make use of the services of community mental health centers often fail to contact the centers (Rossi, Fisher, and Willis, 1986). Similarly, a program designed to provide information to prospective home buyers might find that few persons seek the services offered. Hence, program developers need to be concerned with how best to motivate potential targets to seek out the program and participate in it. Depending on the particular case, they might, for example, need to build outreach efforts into the program or pay special attention to the geographical placement of program sites (Boruch, Dennis, and Carter-Greer, 1988).

One of the most useful tools in designing a scheme for monitoring service utilization is a careful description of the program's service utilization plan, as described above (see Exhibit 6-A). The service utilization plan, recall, is a detailed depiction of the sequence of events through which the target population is expected to make contact with the program, become engaged, and maintain involvement through completion of the intended services. A full articulation of a program's service utilization plan will identify the junctures in the process that are most critical to the program's success in serving the target population and, therefore, most important to monitor for purposes of evaluation, management, or accountability. Moreover, a good service utilization plan will be sufficiently specific about what is expected to happen at each juncture, and what the undesirable alternatives are, to guide the selection of measures or performance indicators that can be used to monitor those events.

Coverage and Bias

Service utilization issues typically break down into questions about *coverage* and *bias.* Whereas coverage refers to the extent to which participation by the target population achieves the levels specified in the program design, bias is the degree to which some subgroups participate in greater proportions than others. Clearly, coverage and bias are related. A program that reaches all projected participants and no others is obviously not biased in its coverage. But because few social programs ever achieve total, exact coverage, bias is typically an issue.

Bias can arise out of self-selection; that is, some subgroups may voluntarily participate more frequently than others. It can also derive from program actions. For instance, a program's personnel may react favorably to some clients while rejecting or discouraging others. One temptation commonly faced by programs is to select the most "success prone" targets. Such "creaming" frequently occurs because of the self-interests of one or more stakeholders (a dramatic example is described in Exhibit 6-F). Finally, bias may result from such unforeseen influences as the location of a program office, which may encourage greater participation by a subgroup that enjoys more convenient access to program activities.

It is usually thought desirable that a program serve a large proportion of the intended targets. The exceptions are those projects whose resources are too limited to provide the appropriate services to more than a portion of the potential targets. In such cases, however, the target definition established during the planning and development of the program probably was not specific enough. Program staff and sponsors may correct this problem by defining the characteristics of the target population more sharply and by using resources more effectively. For example, establishing a health center to provide medical services to persons without regular sources of care may result in such an overwhelming demand that many of those who want services cannot be accommodated. The solution might be to add eligibility criteria that weight such factors as severity of the health problem, family size, age, and income to reduce the size of the target population to manageable proportions while still serving the neediest persons.

The opposite effect, overcoverage, also occurs. For instance, the TV program *Sesame Street* has consistently captured audiences far exceeding the original targets—disadvantaged preschoolers—including children who are not at all disadvantaged and even adults. Because these additional audiences are reached at no additional cost, this overcoverage is not a financial drain. It does, however, thwart one of *Sesame Street's* original goals, which was to lessen the gap in learning between advantaged and disadvantaged children.

In other instances, overcoverage can be costly and problematic. The bilingual programs

EXHIBIT 6-F "Creaming" the Unemployed

When administrators who provide public services choose to provide a disproportionate share of program benefits to the most advantaged segment of the population they serve, they provide grist for the mill of service utilization research. The U.S. Employment Service (USES) offers a clear and significant example of creaming, a practice that has survived half a century of USES expansion, contraction, and reorganization. The USES has as its major aim to provide employers with workers, downplaying the purpose of providing workers with work. This leads the USES to send out the best prospects among the unemployed and to slight the less promising.

It is hardly surprising that USES administrators, a generation after the establishment of the program, stressed the necessity rather than the desirability of an employer-centered service. Its success, by design, depended on serving employers, not the "hard-core" unemployed. As President Johnson's task force on urban employment problems noted some two weeks before the 1965 Watts riots, "We have yet to make any significant progress in reaching and helping the truly 'hard-core' disadvantaged."

SOURCE: Adapted from David B. Robertson, "Program Implementation Versus Program Design," *Policy Study Review,* 1984, 3:391-405.

sponsored by the Department of Education, for instance, have been found to include many students whose primary language is English. Some school systems whose funding from the program depends on the number of children enrolled in bilingual classes have inflated attendance figures by registering inappropriate students. In other cases, schools have used assignment to bilingual instruction as a means of ridding classes of "problem children," thus saturating bilingual classes with disciplinary cases.

The most common coverage problem in social interventions, however, is the failure to achieve full target participation, either because of bias in the way participants are recruited or retained or because potential clients are unaware of the program, unable to use it, or reject it. For example, in most employment training programs only small minorities of those eligible by reason of unemployment ever attempt to join the programs. Similar situations occur in mental health, substance abuse, and numerous other programs (see Exhibit 6-G). We turn now to the question of how program coverage might be measured as a part of program monitoring.

Measuring and Monitoring Coverage

Program managers and sponsors alike need to be concerned with both undercoverage and overcoverage. Undercoverage is measured by the proportion of the targets in need of a program that actually participates in it. Overcoverage is sometimes expressed as the number of program participants who are not in need, compared with the total number not in need in a designated population, and sometimes as the

EXHIBIT 6-G The Coverage of the Food Stamp Program for the Homeless

Based upon a rigorously designed survey of homeless persons sampled from shelters and food kitchens in American cities with a population of 100,000 and over, Burt and Cohen gave some precise dimensions to what we know is true virtually by definition: The homeless live on food intakes that are inadequate both in quantity and in nutritional content. There is no way that a demographic group whose incomes hover slightly above zero can have adequate diets. That the homeless do not starve is largely a tribute to the food kitchens and shelters that provide them with meals at no cost.

Because most homeless persons are eligible by income for food stamps, their participation rates in that program should be high. But they are not—Burt and Cohen reported that only 18% of the persons sampled were receiving food stamps and almost half had never used them. This is largely because certification for food stamps requires passing a means test, a procedure that requires some documentation. This is not easy for many homeless who may not have the required documents, an address to receive the stamps, or the capability to fill out the forms.

Moreover, the food stamp program is based on implicit assumptions that participants can readily acquire their foodstuffs in a local food store, prepare servings on a stove, and store food supplies in their dwellings. These assumptions do not apply to the homeless. Of course, food stores do sell some food items that can be consumed without preparation and, with some ingenuity, a full meal of such foods can be assembled. So some benefit can be obtained by the homeless from food stamps, but for most homeless persons food stamps are relatively useless.

Legislation passed in 1986 allows homeless persons to exchange food stamps for meals offered by nonprofit organizations and made shelter residents in places where meals were served eligible for food stamps. By surveying food providers, shelters, and food kitchens, however, Burt and Cohen found that few meal providers had applied for certification as receivers of food stamps. Of the roughly 3,000 food providers in the sample, only 40 had become authorized.

Furthermore, among those authorized to receive food stamps, the majority had never started to collect food stamps or had started and then abandoned the practice. It made little sense to collect food stamps as payment for meals that otherwise were provided free so that, on the same food lines, food stamp participants were asked to pay for their food with stamps while nonparticipants paid nothing. The only food provider who was able to use the system was one that required either cash payment or labor for meals; for this program, food stamps became a substitute for these payments.

SOURCE: Based on Martha Burt and Barbara Cohen, *Feeding the Homeless: Does the Prepared Meals Provision Help?* Report to Congress on the Prepared Meal Provision, vols. I and II (Washington, DC: Urban Institute, 1988). Reprinted with permission.

number of participants not in need compared with the total number of participants in the program. Generally, it is the latter figure that is important; efficient use of program resources requires both maximizing the number served who are in need and minimizing the number served who are not in need. Efficiency of coverage may be measured by the following formula:

$$\text{Coverage efficiency} = 100 \times \left[\frac{\text{Number in need served}}{\text{Total number in need}} - \frac{\text{Number not in need served}}{\text{Total number served}} \right]$$

This formula yields a positive value of 100 when the actual number served equals the designated target population in need and no inappropriate targets are served. A negative value of 100 occurs if only inappropriate targets are served. Positive and negative values between +100 and −100 indicate the degree of coverage efficiency. For example, if 100 targets need a program in a particular geographical area, and 100 persons are served but only 70 are among those in need, the value obtained by the formula would be +40. If 100 targets need a program, and only 10 of the 100 actually served are appropriate targets, the value obtained would be −80.

This procedure provides a means of estimating the trade-offs in a program that includes inappropriate as well as appropriate targets. The manager of a hypothetical program confronted with a −80 value might, for instance, impose additional selection criteria that eliminated 70 of the 90 inappropriate targets and secure 70 appropriate replacements through an extensive recruitment campaign. The coverage efficiency value would then increase to +60. If the program was inexpensive or if it was either politically unwise or too difficult to impose additional selection criteria to eliminate undercoverage, the manager might elect the option of expanding the program to include all appropriate targets. Assuming the same proportion of inappropriate targets are also served, however, the total number of participants would increase to 1,000!

The problem in measuring coverage is almost always the inability to specify the number in need, that is, the magnitude of the target population. The needs assessment procedures described in Chapter 4, if carried out as an integral part of program planning, usually minimize this problem. In addition, three sources of information can be used to assess the extent to which a program is serving the appropriate target population: program records, surveys of program participants, and community surveys.

Program Records

Almost all programs keep records on targets served. Data from well-maintained record systems—particularly from MISs—can often be used to estimate both program coverage and program bias. For instance, information on the various screening criteria for program intake may be tabulated to determine whether the units served are the ones specified in the program's design. Suppose the targets of a family planning program are women less than 50 years of age who have been residents of the community for at least six months and who have two or more children under age ten. Records of program participants can be examined to see whether the women actually served are within the eligibility limits and the degree to which particular age or parity groups are under- or overrepresented. Such an analysis might also disclose bias in program participation in terms of the eligibility characteristics or combinations of them. Another example involving public shelter utilization by the homeless is described in Exhibit 6-H.

However, programs differ widely in the quality and extensiveness of their records and in the sophistication involved in storing and maintaining them. Moreover, the feasibility of maintaining complete, ongoing record systems for all program participants varies with the nature of the intervention and available resources. In the case of medical and mental

EXHIBIT 6-H Public Shelter Utilization Among Homeless Adults in New York and Philadelphia

The cities of Philadelphia and New York have standardized admission procedures for persons requesting services from city-funded or -operated shelters. All persons admitted to the public shelter system must provide intake information for a computerized registry that includes the client's name, race, date of birth, and gender and must be assessed for substance abuse and mental health problems, medical conditions, and disabilities. A service utilization study conducted by researchers from the University of Pennsylvania analyzed data from this registry for New York City for 1987-1994 (110,604 men and 26,053 women) and Philadelphia for 1991-1994 (12,843 men and 3,592 women).

They found three predominant types of users: (a) the chronically homeless, characterized by very few shelter episodes, but which might last as long as several years; (b) the episodically homeless, characterized by multiple, increasingly shorter stays over a long period; and (c) the transitionally homeless who had one or two stays of short duration within a relatively brief period of time.

The most notable finding was the size and relative resource consumption of the chronically homeless. In New York, for instance, 18% of the shelter users stayed 180 days or more in their first year, consuming 53% of the total number of system days for first-time shelter users, triple the days for their proportionate representation in the shelter population. These long-stay users tended to be older people and to have mental health, substance abuse, and, in some cases, medical problems.

SOURCE: Adapted by permission from Dennis P. Culhane and Randall Kuhn, "Patterns and Determinants of Public Shelter Utilization Among Homeless Adults in New York City and Philadelphia," *Journal of Policy Analysis and Management*, 1998, 17(1):23-43. Copyright © 1998, John Wiley & Sons, Inc.

health systems, for example, sophisticated, computerized management and client information systems have been developed for managed care purposes that would be impractical for many other types of programs.

In measuring target participation, the main concerns are that the data are accurate and reliable. It should be noted that all record systems are subject to some degree of error. Some records will contain incorrect or outdated information, and others will be incomplete. The extent to which unreliable records can be used for decision making depends on the kind and degree of their unreliability and the nature of the decisions in question. Clearly, critical decisions involving significant outcomes require better records than do less weighty decisions. Whereas a decision on whether to continue a project should not be made on the basis of data derived from partly unreliable records, data from the same records may suffice for a decision to change an administrative procedure.

If program records are to serve an important role in decision making on far-reaching issues, it is usually desirable to conduct regular audits of the records. Such audits are similar in intent to those that outside accountants conduct on fiscal records. For example, records might be sampled to determine whether each target has a record, whether records are complete, and whether rules for completing them have been followed.

Surveys

An alternative to using program records to assess target participation is to conduct special surveys of program participants. Sample surveys may be desirable when the required data cannot be obtained as a routine part of program activities or when the size of the target group is large and it is more economical and efficient to undertake a sample survey than to obtain data on all the participants.

For example, a special tutoring project conducted primarily by parents may be set up in only a few schools in a community. Children in all schools may be referred, but the project staff may not have the time or the training to administer appropriate educational skills tests and other such instruments that would document the characteristics of the children referred and enrolled. Lacking such complete records, an evaluation group could administer tests on a sampling basis to estimate the appropriateness of the selection procedures and assess whether the project is serving the designated target population.

When projects are not limited to selected, narrowly defined groups of individuals but instead take in entire communities, the most efficient and sometimes the only way to examine whether the presumed population at need is being reached is to conduct a community survey. Various types of health, educational, recreational, and other human service programs are often community-wide, although their intended target populations may be selected groups, such as delinquent youths, the aged, or women of childbearing age. In such cases, surveys are the major means of assessing whether targets have been reached.

The evaluation of the *Feeling Good* television program illustrates the use of surveys to provide data on a project with a national audience. The program, an experimental production of the Children's Television Workshop (the producer of *Sesame Street*), was designed to motivate adults to engage in preventive health practices. Although it was accessible to homes of all income levels, its primary purpose was to motivate low- income families to improve their health practices. The Gallup organization conducted four national surveys, each of approximately 1,500 adults, at different times during the weeks *Feeling Good* was televised. The data provided estimates of the size of the viewing audiences as well as of the viewers' demographic, socioeconomic, and attitudinal characteristics (Mielke and Swinehart, 1976). The major finding was that the program largely failed to reach the target group, and the program was discontinued.

To measure coverage of Department of Labor programs, such as training and public employment, the department started a periodic national sample survey. The Survey of Income and Program Participation is now carried out by the Bureau of the Census and measures participation in social programs conducted by many federal departments. This large survey, now a three-year panel covering 21,000 households, ascertains through personal interviews whether each adult member of the sampled households has ever participated or is currently participating in any of a number of federal programs. By contrasting program participants with nonparticipants, the survey provides information on the programs' biases in coverage.

In addition, it generates information on the uncovered but eligible target populations.

Assessing Bias: Program Users, Eligibles, and Dropouts

An assessment of bias in program participation can be undertaken by examining differences between individuals who participate in a program and either those who drop out or those who are eligible but do not participate at all. In part, the drop-out rate, or attrition, from a project may be an indicator of clients' dissatisfaction with intervention activities. It also may indicate conditions in the community that militate against full participation. For example, in certain areas lack of adequate transportation may prevent those who are otherwise willing and eligible from participating in a program.

It is important to be able to identify the particular subgroups within the target population who either do not participate at all or do not follow through to full participation. Such information not only is valuable in judging the worth of the effort but also is needed to develop hypotheses about how a project can be modified to attract and retain a larger proportion of the target population. Thus, the qualitative aspects of participation may be important not only for monitoring purposes but also for subsequent program planning.

Data about dropouts may come either from service records or from surveys designed to find nonparticipants. However, community surveys usually are the only feasible means of identifying eligible persons who have not participated in a program. The exception, of course, is when adequate information is available about the entire eligible population prior to the implementation of a project (as in the case of data from a census or screening interview). Comparisons with either data gathered for project-planning purposes or community surveys undertaken during and subsequent to the intervention may employ a variety of analytical approaches, from purely descriptive methods to highly complex models.

In Chapter 11, we describe methods of analyzing the costs and benefits of programs to arrive at measures of economic efficiency. Clearly, for calculating costs it is important to have estimates of the size of populations at need or risk, the groups who start a program but drop out, and the ones who participate to completion. The same data may also be used in estimating benefits. In addition, they are highly useful in judging whether a project should be continued and whether it should be expanded in either the same community or other locations. Furthermore, project staff require this kind of information to meet their managerial and accountability responsibilities. Although data on project participation cannot substitute for knowledge of impact in judging either the efficiency or the effectiveness of projects, there is little point in moving ahead with an impact analysis without an adequate description of the extent of participation by the target population.

MONITORING ORGANIZATIONAL FUNCTIONS

Monitoring the critical organizational functions and activities of a program focuses on whether the program is performing well in managing its efforts and using its resources to accomplish its essential tasks. Chief among those tasks, of course, is delivering the intended services to the target population. In addition, programs have various support functions that must be carried out to maintain the viability and effectiveness of the organization, for example, fund-raising, promotion, advocacy, and governance and management. Program process monitoring seeks to determine whether a program's actual activities and arrangements sufficiently approximate the intended ones.

Once again, program process theory as described in Chapter 3 is a useful tool in designing monitoring procedures. In this instance, what we called the organizational plan is the relevant component (see Exhibit 3-N in Chapter 3). A fully articulated process theory will identify the major program functions, activities, and outputs and show how they are related to each other and to the organizational structures, staffing patterns, and resources of the program. This depiction provides a map to guide the evaluator in identifying the significant program functions and the preconditions for accomplishing them. Program process monitoring then becomes a matter of identifying and measuring those activities and conditions most essential to a program's effective performance of its duties.

Service Delivery Is Fundamental

As mentioned earlier in this chapter, for many programs that fail to show impacts, the problem is a failure to deliver the interventions specified in the program design, a problem generally known as *implementation failure.* There are three kinds of implementation failures: First, no intervention, or not enough, is delivered; second, the wrong intervention is delivered; and third, the intervention is unstandardized or uncontrolled and varies excessively across the target population. In each instance, monitoring the actual delivery of services to identify faults and deficiencies is essential.

"Nonprograms" and Incomplete Intervention

Consider first the problem of the "nonprogram" (Rossi, 1978). McLaughlin (1975) reviewed the evidence on the implementation of Title I of the Elementary and Secondary Education Act, which allocated billions of dollars yearly to aid local schools in overcoming students' poverty-associated educational deprivations. Even though schools had expended the funds, local school authorities were unable to describe their Title I activities in any detail, and few activities could even be identified as educational services delivered to schoolchildren. In short, little evidence could be found that a program existed.

The failure of numerous other programs to deliver services has been documented as well. Datta (1977), for example, reviewed the evaluations on career education programs and found that the designated targets rarely participated in the planned program activities. Similarly, an attempt to evaluate PUSH-EXCEL, a program designed to motivate disadvantaged high school students toward higher levels of academic achievement, disclosed that the program consisted mainly of the distribution of buttons and hortative literature and little else (Murray, 1980).

Instead of not delivering services at all, a delivery system may dilute the intervention so that an insufficient amount reaches the target population. Here the problem may be a lack of commitment on the part of a front-line delivery system, resulting in minimal delivery or "ritual compliance," to the point that the program does not exist. Exhibit 6-I, for instance, expands on an exhibit presented in Chapter 2 to

EXHIBIT 6-1 On the Front Lines: Are Welfare Workers Implementing Policy Reforms?

In the early 1990s the state of California initiated the Work Pays demonstration project, which expanded the state job preparation program (JOBS) and modified AFDC welfare policies to increase the incentives and support for finding employment. The Work Pays demonstration was designed to "substantially change the focus of the AFDC program to promote work over welfare and self-sufficiency over welfare dependence."

The workers in the local welfare offices were a vital link in the implementation of Work Pays. The intake and redetermination interviews they conducted represent virtually the only in-person contact that most clients have with the welfare system. This fact prompted a team of evaluators to study how welfare workers were communicating the Work Pays policies during their interactions with clients.

Using "backwards mapping," the evaluators reasoned that worker-client transactions appropriate to the policy would involve certain "information content" and "use of positive discretion." Information content refers to the explicit messages delivered to clients; it was expected that workers would notify clients about the new program rules for work and earnings, explain opportunities to combine work and welfare to achieve greater self-sufficiency, and inform them about available training and supportive services. Positive discretion relates to the discretion workers have in teaching, socializing, and signaling clients about the expectations and opportunities associated with welfare receipt. Workers were expected to emphasize the new employment rules and benefits during client interviews and communicate the expectation that welfare should serve only as temporary assistance while recipients prepared for work.

To assess the welfare workers' implementation of the new policies, the evaluators observed and analyzed the content of 66 intake or redetermination interviews between workers and clients in four counties included in the Work Pays demonstration. A structured observation form was used to record the frequency with which various topics were discussed and to collect information about the characteristics of the case. These observations were coded on the two dimensions of interest: (a) information content, and (b) positive discretion.

The results, in the words of the evaluators:

> In over 80% of intake and redetermination interviews workers did not provide and interpret information about welfare reforms. Most workers continued a pattern of instrumental transactions that emphasized workers' needs to collect and verify eligibility information. Some workers coped with the new demand by providing information about work-related policies, but routinizing the information and adding it to their standardized, scripted recitations of welfare rules. Others were coping by particularizing their interactions, giving some of their clients some information some of the time, on an ad hoc basis.
>
> These findings suggest that welfare reforms were not fully implemented at the street level in these California counties. Worker-client transactions were consistent with the processing of welfare claims, the enforcement of eligibility rules, and the rationing of scarce resources such as JOBS services; they were poorly aligned with new program objectives emphasizing transitional assistance, work, and self-sufficiency outside the welfare system. (pp. 18-19)

SOURCE: Adapted by permission from Marcia K. Meyers, Bonnie Glaser, and Karin MacDonald, "On the Front Lines of Welfare Delivery: Are Workers Implementing Policy Reforms?" *Journal of Policy Analysis and Management,* 1998, 17(1):1-22. Copyright © 1998, John Wiley & Sons, Inc.

describe the implementation of welfare reform in which welfare workers communicated little to clients about the new policies.

Wrong Intervention

The second category of program failure—namely, delivery of the wrong intervention—can occur in several ways. One is that the mode of delivery negates the intervention. An example is the Performance Contracting experiment, in which private firms contracted to teach mathematics and reading were paid in proportion to pupils' gains in achievement. The companies faced extensive difficulties in delivering the program at school sites. In some sites the school system sabotaged the experiments, and in others the companies were confronted with equipment failures and teacher hostility (Gramlich and Koshel, 1975).

Another way in which wrong intervention can result is when it requires a delivery system that is too sophisticated. There can be a considerable difference between pilot projects and full-scale implementation of sophisticated programs. Interventions that work well in the hands of highly motivated and trained deliverers may end up as failures when administered by staff of a mass delivery system whose training and motivation are less. The field of education again provides an illustration: Teaching methods such as computer-assisted learning or individualized instruction that have worked well within the experimental development centers have not fared as well in ordinary school systems.

The distinction made here between an intervention and its mode of delivery is not always clear-cut. The difference is quite clear in income maintenance programs, in which the "intervention" is the money given to beneficiaries and the delivery modes vary from automatic deposits in savings or checking accounts to hand delivery of cash to recipients. Here the intent of the program is to place money in the hands of recipients; the delivery, whether by electronic transfer or by hand, has little effect on the intervention. In contrast, a counseling program may be handled by retraining existing personnel, hiring counselors, or employing certified psychotherapists. In this case, the distinction between treatment and mode of delivery is fuzzy, because it is generally acknowledged that counseling treatments vary by counselor.

Unstandardized Intervention

The final category of implementation failures includes those that result from unstandardized or uncontrolled interventions. This problem can arise when the design of the program leaves too much discretion in implementation to the delivery system, so that the intervention can vary significantly across sites. Early programs of the Office of Economic Opportunity provide examples. The Community Action Program (CAP) gave local communities considerable discretion in choosing among a variety of actions, requiring only "maximum feasible participation" on the part of the poor. Because of the resulting disparities in the programs of different cities, it is almost impossible to document what CAP's programs accomplished (Vanecko and Jacobs, 1970).

Similarly, Project Head Start gave local communities funds to set up preschool teaching projects for underprivileged children. Across the country, centers varied by sponsoring agencies, coverage, content, staff qualifications, objectives, and a host of other characteristics (Cicirelli, Cooper, and Granger, 1969). Because there is no specified Head Start design, it is not possible to conclude from an evaluation of a sample of projects whether the Head Start concept works. The only generalization that can be made is that some projects are effective and some are ineffective and, among the effective ones, some are more successful than others.

The Delivery System

A program's delivery system can be thought of as a combination of pathways and actions undertaken to provide an intervention (see Chapter 3). It usually consists of a number of separate functions and relationships. As a general rule, it is wise to assess all the elements unless previous experience with certain aspects of the delivery system makes their assessment unnecessary. Two concepts are especially useful for monitoring the performance of a program's delivery system: *specification of services* and *accessibility.*

Specification of Services

For both planning and monitoring purposes, it is desirable to specify the actual services provided in operational (measurable) terms. The first task is to define each kind of service in terms of the activities that take place and the providers who participate. When possible, it is best to separate the various aspects of a program into separate, distinct services. For example, if a project providing technical education for school dropouts includes literacy training, carpentry skills, and a period of on-the-job apprenticeship work, it is advisable to separate these into three services for monitoring purposes. Moreover, for estimating program costs in cost-benefit analyses and for fiscal accountability, it is often important to attach monetary values to different services. This step is important when the costs of several programs will be compared or when the programs receive reimbursement on the basis of the number of units of different services that are provided.

For program monitoring, simple, specific services are easier to identify, count, and record. However, complex elements often are required to design an implementation that is consistent with a program's objectives. For example, a clinic for children may require a physical exam on admission, but the scope of the exam and the tests ordered may depend on the characteristics of each child. Thus, the item "exam" is a service but its components cannot be broken out further without creating a different definition of the service for each child examined. The strategic question is how to strike a balance, defining services so that distinct activities can be identified and counted reliably while, at the same time, the distinctions are meaningful in terms of the program's objectives.

In situations where the nature of the intervention allows a wide range of actions that might be performed, it may be possible to describe services primarily in terms of the general characteristics of the service providers and the time they spend in service activities. For example, if a project places master craftspersons in a low-income community to instruct community members in ways to improve their dwelling units, the craftspersons' specific activities will probably vary greatly from one household to another They may advise one family on how to frame windows and another on how to shore up the foundation of a house. Any monitoring scheme attempting to document such services could only describe the service activities in general terms and by means of examples. It is possible, however, to specify the characteristics of the providers—for example, that they should have five years of experience in home construction and repair and knowledge of carpentry, electrical wiring, foundations, and exterior construction—and the amount of time they spend with each service recipient.

Indeed, services are often defined in terms of units of time, costs, procedures, or products. In a vocational training project, service units may refer to hours of counseling time provided; in a program to foster housing improvement, they may be defined in terms of amounts of building materials provided; in a cottage industry project, service units may refer to activities, such as training sessions on how to operate sewing machines; and in an educational program, the units may be instances of the use of specific curricular materials in classrooms. All these examples require an explicit definition of what constitutes a service and, for that service, what units are appropriate for describing the amount of service.

Accessibility

Accessibility is the extent to which the structural and organizational arrangements facilitate participation in the program. All programs have a strategy of some sort for providing services to the appropriate target population. In some instances, being accessible may simply mean opening an office and operating under the assumption that the designated participants will "naturally" come and make use of the services provided at the site. In other instances, however, ensuring accessibility requires outreach campaigns to recruit participants, transportation to bring persons to the intervention site, and efforts during the intervention to minimize dropouts. For example, in many large cities, special teams are sent out into the streets on very cold nights to persuade homeless persons sleeping in exposed places to spend the night in shelters.

A number of evaluation questions arise in connection with accessibility, some of which relate only to the delivery of services and some of which have parallels in relation to the previously discussed topic of service utilization. Primary is the issue of whether program actions are consistent with the design and intent of the program with regard to facilitating access. For example, is there a Spanish-speaking staff member always available in a mental health center located in an area with a large Hispanic population?

Also, are potential targets matched with the appropriate services? It has been observed, for example, that community members who originally make use of emergency medical care for appropriate purposes may subsequently use them for general medical care. Such misuse of emergency services may be costly and reduce their availability to other community members. A related issue is whether the access strategy encourages differential use by targets from certain social, cultural, and ethnic groups, or is there equal access for all potential targets?

Program Support Functions

Although providing the intended services is presumed to be a program's main function, and one essential to monitor, most programs also perform important support functions that are critical to their ability to maintain themselves and continue to provide service. These functions are of interest to program administrators, of course, but often are also relevant to monitoring by evaluators or outside decisionmakers. Vital support functions may include such activities as fund-raising; public relations to enhance the program's image with potential sponsors, decisionmakers, or the general public; staff training including, possibly, the training of the direct service staff; recruiting and retention of key personnel; developing and maintaining relationships with affiliated programs, referral sources, and the like; obtaining materials required for services; and general advocacy on behalf of the target population served.

Program monitoring schemes can, and often should, incorporate indicators of vital program support functions along with indicators relating to service activities. In form, such indicators and the process for identifying them are no different than for program services. The critical activities first must be identified and described in specific, concrete "output" terms resembling service units, for example, units of fund-raising activity and dollars raised, training sessions, advocacy events, and the like. Measures are then developed that are capable of differentiating good from poor performance and that can be regularly collected. These measures are then included in the program monitoring procedures along with those dealing with other aspects of program performance.

MONITORING PROGRAM OUTCOMES

Outcome monitoring is the routine measurement and reporting of indicators of the results of a program's efforts in the social domain it is accountable for improving (Affholter, 1994). It is important in this context to distinguish between the program's efforts and the resulting improvements (if any) in the target domain. Program outcomes are changes in the social conditions the program addresses that are presumed to result from program actions but are not themselves the program actions. Thus, providing meals to 100 housebound elderly persons is not a program outcome, it is service delivery encompassed within program process. The nutritional effect of those meals on the health of the elderly persons, however, is an outcome, as are any improvements in their morale, perceived quality of life, and risk of injury from attempting to cook for themselves.

A prerequisite for outcome monitoring is identification of the outcomes the program can reasonably be expected to produce. Here, again, a careful articulation of program theory is a very useful tool. In this instance, it is the program impact theory that is relevant. A good impact theory, as described in Chapter 3, will display the chain of outcomes expected to result from program services and be based on detailed input from major stakeholders, consideration of what results are realistic and feasible, and efforts to describe those outcomes in concrete, measurable terms. Another useful feature of well-developed impact theory is that it will distinguish proximal outcomes, those expected to result most immediately from program action, from more distal outcomes that may require more time or a greater cumulation of program effects to attain.

Program outcome monitoring requires that indicators be identified for important program outcomes, starting with the most proximal and covering as many of the more distal ones as is feasible (Exhibit 6-J gives some examples of outcome indicators). This means finding or developing measures that are practical to collect routinely and informative with regard to program performance. The latter requirement is particularly difficult. It is often relatively easy to find indicators of the status of the relevant social condition or target population on an outcome dimension, for instance, the number

EXHIBIT 6-J Examples of Quality-of-Life Changes in Program Participants

Examples of movement toward some desirable change:

■ Condition	A homeless client finding shelter
■ Status	An unemployed client getting a job
■ Behavior	An increase in a juvenile's school attendance
■ Functioning	An increase in a client's coping skills
■ Attitude	An increase in a juvenile's valuing of education
■ Feeling	An increase in a client's sense of belonging
■ Perception	An increase in a client's self-esteem

Examples of movement away from some undesirable change:

■ Condition	Number of nights a homeless person spends on the streets
■ Status	Number of days of work missed by substance-abusing client
■ Behavior	A decrease in the number of times a juvenile skips school
■ Functioning	A decrease in the incidence of a client's fighting with spouse
■ Attitude	A decrease in a juvenile's number of acting-out incidents
■ Feeling	A decrease in a client's feeling of powerlessness over his or her environment
■ Perception	A decrease in a client's negative perception about another ethnic group

SOURCE: Adapted from Lawrence L. Martin and Peter M. Kettner, *Measuring the Performance of Human Service Programs* (Thousand Oaks, CA: Sage, 1996), p. 52.

of children in poverty, the prevalence of drug abuse, the unemployment rate, the reading skills of elementary school students, and the like. The difficulty is in linking *change* in that status specifically with the efforts of the program so that the indicators bear some relation to whatever outcomes the program has actually produced.

The source of this difficulty, as mentioned earlier in this chapter, is that there are usually many influences on a social condition that are not under the program's control. Thus, poverty rates, drug use, unemployment, reading scores, and so forth may change for any number of reasons related to the economy, social trends, and the effects of other programs and policies. Under these circumstances, finding outcome indicators that do a reasonable job of isolating the results attributable to the program in question is often not an easy matter (U.S. General Accounting Office, 1997). Indeed, to isolate program effects in a convincing manner from other influences that might have similar effects requires the special techniques of impact evaluation discussed in Chapters 7-10 of this volume. Because the techniques of impact evaluation are rarely practical for routine, continuing use, outcome monitoring generally will rely on outcome indicators that, at best, are only "outcome related" and, as such, respond to other social influences as well as to the outcomes actually produced by the program.

Guidelines for Outcome Indicators

Nonetheless, there are some guidelines for developing outcome indicators that are as responsive as possible to program effects. One simple point, for instance, is that outcome indicators should be measured only on the members of the target population who actually receive the program services. This means that readily available social indicators for the catchment area served by the program are not good choices for outcome monitoring if they encompass an appreciable number of persons not actually served by the program (although they may be informative supplements to outcome indicators). It also means that those initial program participants who do not actually complete the full prescribed service package should be excluded from the indicator. This is not to say that drop-out rates are unimportant as a measure of program performance, only that they should be assessed as a service utilization issue, not an outcome issue.

Perhaps the most useful technique for focusing outcome indicators on program results is to develop indicators of preprogram to postprogram change whenever possible. For example, it is less informative to know that 40% of the participants in a job training program are employed six months afterward than to know that this represents a change from a preprogram status in which 90% had not held a job for the previous year. One approach to outcome indicators is to define a "success threshold" for program participants and report how many moved from below that threshold to above it after receiving service. Thus, if the threshold is defined as "holding a full-time job continuously for six months," a program might report the proportion of participants falling below that threshold for the year prior to program intake and the proportion of those who were above that threshold during the year after completion of services.

A particularly difficult case for developing outcome indicators with some responsiveness to program-induced change is for preventive programs, whose participants initially are only at risk for a problem rather than actually manifesting the problem. Family preservation programs that intervene when children are judged at risk for being removed from the home illustrate this point. If, after service, 90% of the children are still with their family instead of in foster care, this might appear to indicate a good program outcome. What we do not know is just how much risk there was in the first place that the child would be removed. Perhaps few of these children would actually have been removed from the home in any event, hence the "change" associated with intervention is trivial.

The most interpretable outcome indicators, absent an impact evaluation, are those that involve variables that only the program can affect to any appreciable degree. When these variables also represent outcomes central to the program's mission, they make for an especially informative outcome-monitoring system. Consider, for instance, a city street-cleaning program aimed at picking up litter, leaves, and the like from the municipal streets. Simple before-after photographs of the streets that independent observers rate for cleanliness would yield convincing results. Short of a small hurricane blowing all the litter into the next county, there simply is not much else likely to happen that will clean the streets.

The outcome indicator easiest to link directly to the program's actions is client satisfaction, increasingly called customer satisfaction even in human service programs. Direct ratings by recipients of the benefits they believe

EXHIBIT 6-K Client Satisfaction Survey Items That Relate to Specific Benefits

Client satisfaction surveys typically focus on satisfaction with program services. While a satisfied customer is one sort of program outcome, this alone says little about the specific program benefits the client may have found satisfactory. For client satisfaction surveys to go beyond service issues, they must ask about satisfaction with the *results* of service, that is, satisfaction with particular changes the service might have brought about. Martin and Kettner suggest adding items such as the following to routine client satisfaction surveys:

Service: Information and referral

Question: Has the information and referral program been helpful to you in accessing needed services?

Service: Home-delivered meals

Question: Has the home-delivered meals program been helpful to you in maintaining your health and nutrition?

Service: Counseling

Question: Has the counseling program been helpful to you in coping with the stress in your life?

SOURCE: Adapted from Lawrence L. Martin and Peter M. Kettner, *Measuring the Performance of Human Service Programs* (Thousand Oaks, CA: Sage, 1996), p. 97.

the program provided to them are one form of assessment of outcomes. In addition, creating feelings of satisfaction about the interaction with the program among the participants is a form of outcome, although not one that, in itself, necessarily improves the participants' lives. The more pertinent information comes from participants' reports of whether very specific benefits resulted from program service (see Exhibit 6-K). The limitation of such indicators is that program participants are not always in a position to recognize or acknowledge program benefits, such as drug addicts encouraged to use sterile needles. Alternatively, participants may be able to report on benefits but be reluctant to appear critical and thus overrate them, as might elderly persons asked about the visiting nurses who come to their homes.

Pitfalls in Outcome Monitoring

Because of the dynamic nature of the social conditions typical programs attempt to affect, the limitations of outcome indicators described above, and the pressures on program agencies, there are many pitfalls that are associated with program outcome monitoring. This is not to say that such indicators cannot be a valuable source of information about program performance for program decisionmakers, only that they must be developed and used very carefully.

One important consideration is that any outcome indicator to which program funders or other influential decisionmakers give serious attention will also inevitably receive emphasis from program staff and managers. Thus, if the outcome indicators are not appropriate or fail to cover all important outcomes, program efforts to improve the performance they reflect may distort program activities. Affholter (1994), for instance, describes a situation in which a state used the number of new foster homes licensed as an indicator of increased placements for children with multiple problems. Workers responded by vigorously recruiting and licensing new homes even when the foster parents lacked the specialized skills needed to take hard-to-place children or were not appropriate at all for such children. Thus, the indicator continued to move upward but the actual placement of children in the target population did not actually improve. In education, this is called "teaching to the test." Good outcome indicators, by contrast, must "test to the teaching."

A related problem is the "corruptibility of indicators." This refers to the natural tendency for those whose performance is being evaluated to fudge and pad the indicator whenever possible to make their performance look better than it is. In a program for which the rate of post-program employment among participants is a major outcome indicator, for instance, consider the pressure on the program staff who telephone the participants six months after completion of the program to ascertain their job status. Even with a reasonable effort at honesty, ambiguous cases will be far more likely to be recorded as employment than not. It is usually best for such information to be collected by persons independent from the program if possible. If it is collected internal to the program, it is especially important that careful procedures be used and the results verified in some convincing manner.

Another potential problem area has to do with the interpretation of results on outcome indicators. Given a range of factors other than program performance that may influence those indicators, interpretations made out of context can be very misleading and, even with proper context, can be difficult. To provide suitable context for interpretation, outcome indicators must generally be accompanied by other information that provides a relevant basis for comparison or helps explain potentially anomalous results on the indicator. Outcome indicators are more informative, for instance, if they are examined as part of a time series that shows how the current situation compares with prior periods. It is also pertinent to have information about changes in client mix, demographic trends, and the like as part of the package. Decreased job placement rates, as one example, are more accurately interpreted as a program performance indicator if accompanied by summaries indicating the seriousness of the unemployment problems of the program participants. It may be no reflection on program performance if the placement rate decreases so long as it is clear that the program is working with clients who have fewer job skills and longer unemployment histories.

Similarly, outcome information is often more readily interpreted when accompanied by program process and service utilization information. A favorable job placement rate for clients completing training may, nonetheless, be a matter for concern if, at the same time, monitoring of service utilization shows that training completion rates have dropped to very low levels. The favorable placement rates may only reflect the dropout of all the clients with serious problems, leaving only the "cream of the crop" for the program to place. Incorporating process and utilization information into the interpretation of outcome indicators is especially important when different units, sites, or programs are being compared. It would be neither accurate nor fair to form a negative judgment of one program unit that was lower on an outcome indicator than other program units without considering whether it was dealing with more difficult cases, maintaining lower drop-out rates, or coping with other extenuating factors.

The upshot of these various considerations is that a weak showing on an outcome indicator is not usually a sufficient basis for concluding that program performance is poor. Rather, it should be a signal that further inquiry is needed to determine why the outcome indicator is low. In this way, outcome monitoring is not a substitute for impact evaluation but a preliminary outcome assessment that is capable of giving informative feedback to program decisionmakers, holding programs accountable for showing outcomes, and highlighting where a more probing evaluation approach is needed to contribute the most to program improvement.

COLLECTING DATA FOR MONITORING

A variety of techniques may be used singly and in combination to gather data on program implementation (see King, Morris, and Fitz-Gibbon, 1987; Martin and Kettner, 1996). As in all aspects of evaluation, the particular approaches used must take into account the resources available and the expertise of the evaluator. There may be additional restrictions on data collection, however. One concerns issues of privacy and confidentiality. Program services that depend heavily on person-to-person delivery methods, such as mental health, family planning, and vocational education, cannot be directly observed without violating privacy. In other contexts, self-administered questionnaires might, in theory, be an economical means of studying a program's implementation, but functional illiteracy and cultural norms may prohibit their use.

Several data sources should be considered for program monitoring purposes: data collected directly by the evaluator, program records, and information from program participants or their associates. The approaches used to collect and analyze the data overlap from one data source to the next. A comprehensive monitoring evaluation might include data from all three sources.

Data Collected by the Evaluator

Often critical program monitoring information can be obtained by direct observation of service delivery or other important program functions (Exhibit 6-I, presented earlier, provides an example). Observational methods are feasible whenever the presence of an observer is not obtrusive and the matter at issue is directly observable. In some cases, it can be useful for observers to become, at least for a time, full or partial program participants. Reiss (1971), for example, placed observers in police patrol cars and had them fill out systematic reports of each encounter between the police and citizens in a sample of duty tours. A similar approach was used in the Kansas City Preventive Patrol experiment (Kelling et al., 1974). It is always a question, however, how much the presence of participant observers alters the behavior of program personnel or program participants. Impressionistic evidence from the police studies does not indicate that observers affected the delivery system, because police in the patrol cars soon become accustomed to being observed. Nonetheless, participant-observation methods should be sensitive to the problem of observer effects.

An essential part of any observation effort is a plan for systematically recording the obser-