KEY CONCEPTS FOR CHAPTER 6

Program monitoring The systematic documentation of aspects of program performance that are indicative of whether the program is functioning as intended or according to some appropriate standard. Monitoring generally involves program performance related to program process, program outcomes, or both.

Accountability The responsibility of program staff to provide evidence to stakeholders and sponsors that a program is effective and in conformity with its coverage, service, legal, and fiscal requirements.

Administrative standards Stipulated achievement levels set by program administrators or other responsible parties, for example, intake for 90% of the referrals within one month. These levels may be set on the basis of past experience, the performance of comparable programs, or professional judgment.

Process evaluation A form of program monitoring designed to determine whether the program is delivered as intended to the targeted recipients. Also known as implementation assessment.

Management information system (MIS) A data system, usually computerized, that routinely collects and reports information about the delivery of services to clients and, often, billing, costs, diagnostic and demographic information, and outcome status.

Performance measurement The collection, reporting, and interpretation of performance indicators related to how well programs perform, particularly with regard to the delivery of service (outputs) and achievement of results (outcomes).

Outcome monitoring The measurement and reporting of indicators of the status of the social conditions the program is accountable for improving.

Implementation failure The program does not adequately perform the activities specified in the program design that are assumed to be necessary for bringing about the intended social improvements. It includes situations in which no service, not enough service, or the wrong service is delivered, or the service varies excessively across the target population.

Accessibility The extent to which the structural and organizational arrangements facilitate participation in the program.

Coverage The extent to which a program reaches its intended target population.

Bias in coverage The extent to which subgroups of a target population participate differentially in a program.

CHAPTER 6

MONITORING PROGRAM PROCESS AND PERFORMANCE

In previous chapters, we discussed the ways in which evaluators can assess the nature of the social problem targeted by a program and the quality of the theory inherent in a program about how the program activities will ameliorate that problem. To be effective in bringing about the desired improvements in social conditions, of course, a program needs more than a good plan of attack, although that is an essential precondition. Most important, the program must implement its plan; that is, it must actually carry out the intended functions in the intended way.

Although implementing a program concept may seem straightforward, in practice it is often very difficult. Social programs typically must contend with many adverse influences that can compromise even well-intentioned attempts to conduct program business appropriately. The result can easily be substantial discrepancies between the program as intended and the program actually implemented.

An important evaluation function, therefore, is to assess program implementation: the program activities that actually take place and the services that are actually delivered in routine program operation. Program monitoring and related procedures are the means by which the evaluator investigates these issues.

Program monitoring is usually directed at one or more of three key questions: (a) whether a program is reaching the appropriate target population, (b) whether its service delivery and support functions are consistent with program design specifications or other appropriate standards, and (c) whether positive changes appear among the program participants and social conditions the program addresses. Monitoring may also examine what resources are being, or have been, expended in the conduct of the program.

Program monitoring is an essential evaluation activity. It is the principal tool for formative evaluation designed to provide feedback for program improvement and is especially applicable to relatively new programs attempting to establish their organization, clientele, and services. Also, adequate monitoring (process evaluation) is a vital complement to impact evaluation, helping distinguish cases of poor program implementation from ineffective intervention concepts.

Program monitoring also informs policymakers, program sponsors, and other stakeholders about how well programs perform their intended functions. Increasingly, some form of program performance monitoring is being required by government and nonprofit agencies as a way of demonstrating accountability to the public and the program stakeholders.

After signing a new bill, President Kennedy is reputed to have said to his aides, "Now that this bill is the law of the land, let's hope we can get our government to carry it out." Both those in high places and those on the front lines are often justified in being skeptical about the chances that a social program will be appropriately implemented. Many steps are required to take a program from concept to full operation, and much effort is needed to keep it true to its original design and purposes. Thus, whether any program is fully carried out as envisioned by its sponsors and managers is always problematic.

One important and useful form of evaluation, therefore, is devoted to describing how a program is operating and assessing how well it is performing its intended functions. This form of evaluation does not represent a single distinct evaluation procedure but, rather, a family of approaches, concepts, and methods that are used in different contexts and for different purposes. The defining theme of this form of evaluation is a focus on the enacted program itself—its operations, activities, functions, performance, component parts, resources, and so forth. There is no widely accepted label for this family of evaluation approaches, but because it mainly involves measuring and recording information about the operation of the program, we will refer to it generally as program monitoring.

WHAT IS PROGRAM MONITORING?

Program monitoring is the systematic documentation of key aspects of program performance that are indicative of whether the program is functioning as intended or according to some appropriate standard. It generally involves program performance in the domain of service utilization, program organization, and/or outcomes. Monitoring service utilization consists of examining the extent to which the intended target population receives the intended services. Monitoring program organization requires comparison of the plan for what the program should be doing, especially with regard to providing services, and what is actually done. Monitoring program outcome entails a survey of the status of program participants after they have received service to determine if it is in line with what the program intended to accomplish.

In addition to these primary domains, program monitoring may include information about resource expenditures that bear on whether the benefits of a program justify its cost. Monitoring also may include an assessment of whether program activities comply with legal and regulatory requirements—for example, whether affirmative action requirements have been met in the recruitment of staff.

More specifically, program monitoring schemes are designed to answer such evaluation questions as these:

How many persons are receiving services?

Are those receiving services the intended targets?

Are they receiving the proper amount, type, and quality of services?

Are there targets who are not receiving services?

Are members of the target population aware of the program?

Are necessary program functions being performed adequately?

Is program staffing sufficient in numbers and competencies for the functions that must be performed?

Is the program well organized? Do staff work well with each other?

Does the program coordinate effectively with the other programs and agencies with which it must interact?

Are program resources, facilities, and funding adequate to support important program functions?

Are program resources used effectively and efficiently?

Are costs per service unit delivered reasonable?

Is the program in compliance with requirements imposed by its governing board, funding agencies, and higher-level administration?

Is the program in compliance with applicable professional and legal standards?

Is program performance at some program sites or locales significantly better or poorer than at others?

Are participants satisfied with their interactions with program personnel and procedures?

Are participants satisfied with the services they receive?

Do participants engage in appropriate follow-up behavior after service?

Are participants' conditions, status, or functioning satisfactory in areas the service addresses after service is completed?

Do participants retain satisfactory conditions, status, or functioning for an appropriate period after completion of services?

For any particular program, of course, more specialized versions of these questions will be at issue. In a Head Start early-education program, for instance, the questions would involve pertinent characteristics of the target children, the teachers and aides, the classroom facilities and materials, the instructional and recreational activities, the parents' attitudes toward the program, the language and social skills of the children, and so forth. Nonetheless, this list of questions serves to characterize the general nature of the issues that program monitoring typically investigates.

It is especially important to recognize the evaluative themes in program monitoring questions such as those listed above. Virtually all involve words such as *appropriate, adequate, sufficient, satisfactory, reasonable, intended,* and other phrasing that indicates that an evaluative judgment is required. To answer these questions, therefore, the evaluator or other responsible parties must not only describe the program performance but assess whether it is satisfactory. This, in turn, requires that there be some basis for making a judgment, that is, some defensible criteria or standards to apply. In situations where such criteria are not already articulated and endorsed, the evaluator may find that establishing workable criteria is as difficult as determining program performance on the pertinent dimensions.

There are several approaches to the matter of setting criteria for program performance. Moreover, different approaches will likely apply to different dimensions of program performance because the considerations that go into defining, say, what constitutes an appropriate number of clients served are quite different from those pertinent to deciding what constitutes adequate program resources. This said, however, the approach to the criterion issue that has the broadest scope and most general utility in program monitoring is the application of program theory as described previously in Chapter 3.

Recall that program theory, as we presented it, is divided into program process theory and program impact theory. Program process theory is formulated to describe the program as intended in a form that virtually constitutes a plan or blueprint for what the program is expected to do and how, as a result, targets will receive appropriate services. Program impact theory is formulated to describe what outcomes are expected to follow from effective service and why. Furthermore, these formulations, properly done, build on a needs assessment (whether systematic or informal) and thus connect the program design with the social conditions the program is intended to ameliorate. And, of course, the process through which they are derived and adopted usually involves both input and ultimate endorsement by the major stakeholders. Program theory thus has a certain authority in delineating what a program "should" be doing and, correspondingly, what constitutes adequate performance.

Program monitoring, therefore, can be built on the scaffolding of program theory, especially process theory. Program process theory describes the critical components, functions, and relationships that are assumed to be necessary for an effective program, because that is its primary purpose. This information identifies the aspects of program performance that are most important to monitor. As a program blueprint, however, process theory also gives some indication of what level of performance is intended and, thus, provides some basis for assessing whether actual performance measures up.

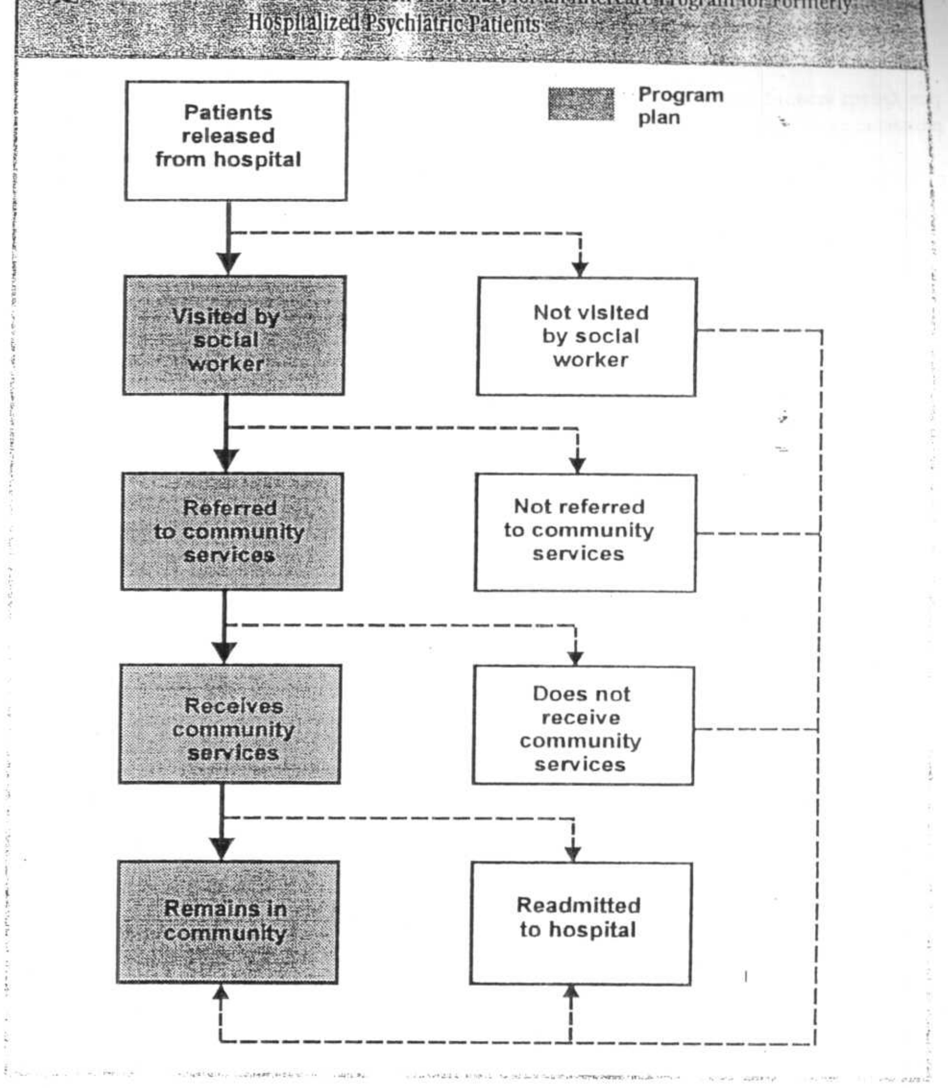
An example will perhaps clarify the relationship between program process theory and the assessment of program performance through a monitoring scheme. Exhibit 3-M in Chapter 3 illustrated the service utilization component of program process theory for an aftercare program for released psychiatric patients. For convenience, this is reproduced in this chapter as Exhibit 6-A. This flowchart depicts, step by step, the interactions and experiences patients released from the hospital are supposed to have as a result of program service.

A thorough monitoring procedure should report on each important aspect of service utilization. The first role of the service utilization flowchart in Exhibit 6-A, therefore, is to identify the important events so that information can be collected about them. Program monitoring would then document in some systematic manner what actually happened at each step. A monitoring procedure for this aftercare program, for instance, might report how many patients were released from the hospital each month, what proportion were visited by a social worker, how many were referred to services and which services, how many actually received those services, and so forth.

The second function of the service utilization flowchart is to indicate just what *should* happen at each step. If what is supposed to happen does not happen, that indicates poor program performance. In practice, of course, the critical events will not occur in an all-or-none fashion but will be attained to some higher or lower degree. Thus some, but not all, of the released patients will receive visits from the social worker, some will be referred to services, and so forth. Moreover, there may be important quality dimensions. For instance, it would not represent good program performance if a released patient was referred to several community services but they were not appropriate to his or her needs.

The service utilization plan in Exhibit 6-A only tells us categorically what is supposed to happen, which provides some basis for assessing performance but does not tell us how much must be done, or how well, to constitute good performance. For that, we need additional criteria that parallel the information the monitoring procedure provides. That is, if the monitoring procedure reports that 63% of the released patients are visited by a social worker within two weeks of release, we cannot evaluate that performance without some standard that tells us what percentage is "good." Is 63% a poor performance, given that we might expect 100% to be desirable, or is it a very impressive performance with a clientele that is difficult to locate and serve?

EXHIBIT 6-A Service Utilization Flowchart for an Aftercare Program for Formerly Hospitalized Psychiatric Patients


Patients released from hospital
▨ Program plan
Visited by social worker
Not visited by social worker
Referred to community services
Not referred to community services
Receives community services
Does not receive community services
Remains in community
Readmitted to hospital

The most common and widely applicable criteria for such situations are simply *administrative standards* or objectives, that is, stipulated achievement levels set by program administrators or other responsible parties. For example, a program director and staff may commit to attaining 80% completion rates for services or to having 60% of the program participants permanently employed six months after receiving the program's job training. For the aftercare program above, it might be that the administrative target is to have 75% of the patients visited within two weeks of hospital release. Thus, the 63% found with program monitoring shows a subpar performance that, nonetheless, is not too far below the mark.

Administrative standards and objectives for program performance may be set on the basis of past experience, the performance of comparable programs, or simply the professional judgment of program managers or advisers. If reasonably justified, however, they can provide meaningful standards against which to assess observed program performance. In a related vein, some aspects of program performance may fall under applicable legal, ethical, or professional standards. The "standards of care" adopted in medical practice for treating common ailments, for instance, provide an essential set of criteria against which to assess program performance in health care settings. Similarly, a program of children's protective services has legal requirements to meet with regard to how it handles cases of possible child abuse or neglect.

Some recognition must also be given to the fact that, in practice, the assessment of particular dimensions of program performance is often not based on specific, predetermined criteria but represents an after-the-fact judgment call. This is the "I'll know it when I see it" school of thought on what constitutes good program performance. An evaluator who collects program monitoring data on, say, the proportion of high-risk adolescents who recall seeing program-sponsored antidrug media messages may find program staff and other key stakeholders rather vague and inconsistent in their views of what an acceptable proportion would be. If the results come in at 50%, however, a consensus may arise that this is rather good considering the nature of the population, even though some stakeholders might have reported much higher expectations prior to seeing the data. On the other hand, 5% might strike all stakeholders as distressingly low.

The example above makes use of the service utilization component of program process theory. Very similar considerations apply to the organizational component of the process theory. A depiction of the organizational plan for the aftercare program was presented in Exhibit 3-N in Chapter 3. Looking back at it will reveal that it, too, identifies dimensions of program performance that can be monitored and assessed against appropriate standards. Under that plan, for instance, case managers are expected to interview clients and families, assess service needs, make referrals to services, and so forth. A program monitoring procedure would document what was done under each of those categories and provide that information for assessment.

Program impact theory, on the other hand, serves a somewhat different role in relation to program monitoring than program process theory. Impact theory identifies the outcomes that are expected to result from the program and, therefore, gives guidance to any attempt to monitor the status, condition, or functioning of program participants on relevant outcome dimensions. Not all program monitoring schemes include outcome indicators, in part because the data can be difficult to collect. Moreover, describing service recipients with regard to their status on relevant outcome indicators does not tell us what effects (or impact) the program has had on those dimensions, only what participants' overall level is on them. (The next four chapters of this volume discuss the special demands of impact assessment.) Nonetheless, as will be discussed later in this chapter, there are good reasons for some program monitoring schemes to track outcome data and assess them, like process data, against administrative objectives and other such applicable standards.

Common Forms of Program Monitoring

Monitoring and assessment of program performance are quite common in program evaluation, but the approaches used are rather varied and there is little uniformity in the terminology for the different variants. The commonality among these variants is a focus on indicators (qualitative or quantitative) of how well the program performs its critical functions. An assessment of this sort may be conducted as a one-shot endeavor or may be continuous so that information is produced regularly over an extended period of time. It may be conducted by an outside evaluator or an evaluator employed within the program agency and may, indeed, be set up as a management tool with little involvement by professional evaluators. Moreover, its purpose may be to provide feedback for managerial purposes, to demonstrate accountability to sponsors and decisionmakers, to provide a freestanding process evaluation, or to augment an impact evaluation. Amid this variety, we distinguish three principal forms of program monitoring, which are described briefly below.

Process or Implementation Evaluation

Evaluators often distinguish between process (or implementation) evaluation and outcome (or impact) evaluation. *Process evaluation*, in Scheirer's (1994) words, "verifies what the program is and whether or not it is delivered as intended to the targeted recipients." It does not, however, attempt to assess the effects of the program on those recipients—that is the province of impact evaluation. Process evaluation is typically conducted by evaluation specialists as a separate project that may involve program personnel but is not integrated into their daily routine. When completed and, often, while under way, process evaluation generally provides information about program performance to program managers and other stakeholders, but is not a regular and continuing part of management information systems (MISs). Exhibit 6-B describes a process evaluation of an integrated services program for children.

As an evaluation approach, process evaluation plays two major roles. First, it can stand alone as an evaluation of a program in circumstances where the only questions at issue are

EXHIBIT 6-B Process Evaluation to Assess Integrated Services for Children

Many analysts have observed that the traditional system of categorical funding for children's services, with funds allocated to respond to specific problems under strict rules regarding eligibility and expenditures, has not served children's needs well. The critics argue that this system fragments services and inhibits collaboration between programs that might otherwise lead to more effective services.

In 1991 the Robert Wood Johnson Foundation launched the Child Health Initiative to test the feasibility of achieving systemic changes through the integration of children's services and finances. Specifically, the initiative called for the development of the following components:

- A decategorization mechanism that would pool existing categorical program funds and create a single children's health fund.
- A care coordination procedure using case management that would use the pooled funds to provide comprehensive and continuous care for needy children.
- A monitoring system that would identify the health and related needs of children in the community and the gaps in existing services.

Nine sites across the country were selected to launch demonstration programs. The Institute for Health Policy Studies, University of California, San Francisco conducted an evaluation of these programs with two major goals: (a) to gauge the degree to which the implementation of the projects was consistent with the original planning objectives (fidelity to the model), and (b) to assess the extent to which each of the major program components was implemented.

In the first year, the evaluation focused on the political, organizational, and design phase of program development. During subsequent years, the focus turned to implementation and preliminary outcomes. A combination of methods was used, including site visits, written surveys completed by the program managers, in-depth interviews of key participants, focus groups of service providers and clients, and reviews of project-related documents.

The evaluation found that most of the nine sites experienced some degree of success in implementing the monitoring and care coordination components, but none was able to implement decategorization. The general find- ings for each component were as follows:

- Decategorization—several sites successfully created small pools of flexible funds but these were from sources other than categorical program funds. No site was able to fully implement decategorization under the definitions originally adopted.
- Care coordination—was implemented successfully by most of the sites at the client level through case management but there was generally less coordination at the system level.
- Monitoring—the sites encountered a number of barriers in successfully completing this task but most instituted some appropriate process.

SOURCE: Adapted from Claire Brindis, Dana C. Hughes, Neal Halfon, and Paul W. Newacheck, "The Use of Formative Evaluation to Assess Integrated Services for Children," *Evaluation & the Health Professions,* 1998, 21(1):66-90.

about the effectiveness of program operations, service delivery, and other such matters. A stand-alone process evaluation might be appropriate for a relatively new program, for instance, to answer questions about how well it has established its operations and services. Program process is often the focus of formative evaluation designed to provide useful feedback to managers and sponsors of new programs. A process evaluation might also be called for in the case of a more established program when questions arise about how well it is organized, the quality of its services, or the success with which it is reaching the target population. A process evaluation may also constitute the major evaluation approach to a program charged with delivering a service known or presumed effective so that the most significant performance issue is whether that service is being delivered properly. In a managed care environment, for instance, process evaluation may be employed to assess whether the prescribed medical treatment protocols are being followed for patients in different diagnostic categories.

Process or implementation evaluation is also often carried out in conjunction with an impact evaluation. Indeed, it is generally not advisable to conduct an impact evaluation without including at least a minimal process evaluation. A precondition for impact on the social conditions a program addresses is that the program actually be implemented in a manner that could plausibly affect those conditions. Because maintaining an operational program and delivering appropriate services on an ongoing basis are formidable challenges in many human service arenas, it is not generally wise to take program implementation for granted. A full impact evaluation, therefore, generally includes a process component to determine what quality and quantity of services the program provides so that this information can be integrated with findings on what impact those services have.

Routine Program Monitoring and Management Information-Systems

Continuous monitoring of indicators of selected aspects of program process can be a useful tool for effective management of social programs by providing regular feedback about how well the program is performing its critical functions. Such feedback allows managers to take corrective action when problems arise and can also provide stakeholders with regular assessments of program performance. For these reasons, a form of process assessment is often integrated into the routine information systems of social programs so that appropriate data are obtained, compiled, and periodically summarized for review. In such cases, process evaluation becomes coextensive with the MISs in human service programs. Exhibit 6-C describes an MIS that was developed for a marital and family counseling program.

MISs routinely provide information on a client-by-client basis about services provided, staff providing the services, diagnosis or reasons for program participation, sociodemographic data, treatments and their costs, outcome status, and so on. Some of the systems bill clients (or funders), issue payments for services, and store other information, such as a client's treatment history and current participation in other programs. MISs have supplanted process evaluations in many instances because much of the information that would be gathered for process evaluation is available in the program's MIS. Even when a program's MIS is not configured to completely fulfill the requirements of a thoroughgoing process evaluation, it may nonetheless provide a large portion of the information an evaluator needs for such purposes. MISs can thus supply data that can be used by both managers and evaluators.

EXHIBIT 6-C An Integrated Information System for a Family and Marriage Counseling Agency in Israel

The Marital and Family Counselling Agency is run under the joint auspices of the Welfare Department of the Tel Aviv municipality and the Bob Shapell School of Social Work at Tel Aviv University. The agency provides marital and family counseling and community services for the Jewish, Moslem, and Christian residents of one of the poorest sections of Tel Aviv.

The integrated information system developed for the agency is designed to follow up clients from the moment they request help to the end of treatment. It is intended to serve the agency and the individual counselors by monitoring the process and outcomes of treatment and providing the data needed to make organizational and clinical decisions. To accomplish this, data are collected on three forms and then programmed into the computerized information system. The data elements include

- Background data provided by the client, for example, sociodemographic characteristics, medical and psychological treatment history, the problems for which they are seeking help, the urgency of those problems, their expectations from treatment, and how they found out about the clinic.
- The McMaster Clinical Rating Scale, a standardized scale that monitors families on the basis of six dimensions of family functioning and overall family health; the counselors fill out this form once a month for each client.
- A retrospective evaluation form filled out after treatment is completed, one by the counselors and another by the clients. This includes, for example, factual questions about the treatment such as its duration, the problems dealt with, the degree to which the client and counselor agreed on the problems, whether there were issues not addressed and why, retrospective assessments of the process and evaluations of improvement in the presented problems and the McMaster areas of functioning, client and counselor satisfaction with the process and outcomes.

The counselors can enter and retrieve data from this system whenever they wish and are given a graph of each client's status every three months to support clinical decisions. Also, reports are generated for the clinic management. For example, a report of the distribution of clients by ethnic group led to the development of a program located within Arab community centers to better reach that population. Other management reports describe the ways and times at which treatment is terminated, the problems that brought clients to the agency, and the percentage of people who applied for treatment but did not show up for the first session. The information system has also been used for research purposes. For example, studies were conducted on the predictors of treatment success, the comparative perceptions by clients and counselors of the treatment process and outcomes, and gender differences in presenting problems.

SOURCE: Adapted from Rivka Savaya, "The Potential and Utilization of an Integrated Information System at a Family and Marriage Counselling Agency in Israel," *Evaluation and Program Planning,* 1998, 21(1):11-20.

Performance Measurement and Monitoring

Increased public and political demands for accountability from social service agencies in recent years have brought forth a variety of initiatives to require such agencies to demonstrate that their programs accomplish something worthwhile. The most far-reaching of these initiatives is the Government Performance and Results Act of 1993 (GPRA), which requires federal agencies to identify the goals of their programs and report on their results in attaining those goals. Recognizing that this will be difficult for many agencies, GPRA provided for a seven-year implementation period with all agencies required to institute regular reporting by fiscal year 2000. More than 70 pilot projects have been launched under this act to provide experience with the concepts and procedures involved (Martin and Kettner, 1996; U.S. General Accounting Office, 1997).

In addition, many of the federal block grant programs require performance measurement and reporting, and a number of state legislatures have imposed similar requirements on their state agencies (Hatry, 1997). In the 1990s, the Governmental Accounting Standards Board (GASB), a private organization that sets the accounting standards for state and local governments, began working on "service efforts and accomplishments" (SEA) reporting. If such reporting becomes mandatory, as is expected within several years, all state and local government agencies will be required to identify measures of performance and report results on them (Martin and Kettner, 1996). Many major nonprofit agencies are also pressing forward with performance measurement initiatives (Plantz, Greenway, and Hendricks, 1997). The United Way of America has produced materials for its regional chapters and member agencies to use in developing performance monitoring, and similar efforts have been made by such organizations as Boy Scouts of America, Girls Incorporated, the Family Services Association of America, and Goodwill Industries International. Managed care agencies in health and mental health services have been particularly active in developing performance monitoring systems as part of their cost-control and quality assurance efforts.

The performance measurement schemes emerging from these various initiatives (also known as performance monitoring and outcome monitoring) have much in common with those for process evaluation and MISs. Like process evaluation, they involve collection, reporting, and interpretation of information relevant to how well programs perform certain critical functions, particularly the delivery of service. And, like MISs, performance measurement is intended to be a routine and continuing program activity that will improve management and yield regular reports of program accomplishments. Compared with these other program monitoring approaches, however, performance measurement strategies orient especially toward assessment of program outcomes, that is, the results of services. An example of performance measurement for a family crisis program appears in Exhibit 6-D.

In particular, performance measurement schemes distinguish program *outcomes* from program *outputs.* Program outputs are the products or services delivered to program participants or other such activities viewed as part of the program's contribution to society. Measures of output, for example, would relate to

EXHIBIT 6-D Outcome Monitoring for a Family Crisis Program

Many states buy crisis intervention services. The concept is to identify and help families where child abuse or neglect has occurred as a result of a temporary crisis. Florida began an Intensive Crisis Counseling Program (ICCP) as a demonstration in one site more than ten years ago. Under contract, professional counselors enter the home and work intensively with the family for relatively short periods to resolve the crisis, to remove the risk of subsequent or continued abuse or neglect, and thereby to avert a placement in emergency shelter or foster care.

Florida has used an outcome monitoring system for ICCP since its inception. The most significant outcome indicators are counts of the families that remain intact at case closures and those for which children were removed and placed in shelter or foster care. Overall, these data showed that 80% or more of the families served remained intact at case closure. This apparent success was one of the factors that has encouraged Florida to gradually expand ICCP over the years.

The ICCP outcome monitoring system recently began to report data by individual contract providers. This has shown that there is much variation in performance–from 70% to 93% of the families served by different providers remained intact at case closure and even wider variation was found for family status three months after cases were closed.

One use of the findings from the outcome monitoring system was that administrative staff decided to investigate providers showing poorer outcomes. In the case of one of the poorest performers, for instance, they discovered that the program had evolved into a service where the provider was available 24 hours via a telephone hotline but no longer provided in-home service beyond an initial assessment.

SOURCE: Adapted from Dennis P. Affholter, "Outcome Monitoring," in *Handbook of Practical Program Evaluation*, eds. J. S. Wholey, H. P. Hatry, and K. E. Newcomer (San Francisco: Jossey-Bass, 1994), pp. 96-118.

such things as the number of clients served, the number of service units provided, cost per service unit, the quality of services provided, the nature and volume of advocacy or promotional efforts made by the program, and so forth. In contrast, program outcomes represent the results of those activities, such as improved health for the individuals served, increased community awareness of AIDS risk, better reading skills, and other such social benefits. Performance measurement is generally intended to encompass the quantity and quality of both outputs and outcomes.

An important distinction must be made here between *measuring or monitoring* the social conditions that programs aim to affect and assessing the *impact* of programs on those conditions. Measuring program outcomes means describing social conditions on some set of indicators that represent the nature or extent of those conditions. The number of homeless families who obtain housing, the unemployment rate, the reported quality of life among frail elderly persons, and the average math achievement scores of sixth-grade students are all measures of conditions that some program may strive to change. An effective program might hope to find that regular measurement reveals improved conditions or, at worst, no deterioration in them.

Measuring and monitoring the target social conditions, however, are not sufficient to show that the program activities have actually been the source of any changes observed. To demonstrate program impact on the conditions, the effects of the program must be distinguished from the effects of other influences on those conditions, such as outside social forces, natural trends, and ameliorative actions taken by other social programs or policies or by members of the target population themselves. Chapters 7-10 of this volume discuss the demanding nature of impact evaluation and the special methods required to isolate the cause-and-effect relationship between program action and the resulting outcomes. These methods are typically beyond the scope of performance measurement schemes. The monitoring of outcome conditions in such schemes is aimed at providing feedback about how bad those conditions are and whether they are changing in favorable directions, not at assessing the distinct impact of the program on those conditions.

This is not to say that outcome monitoring provides no useful information about program effects. Outcome measures that focus specifically on the recipients of program service, and are collected periodically so that the status of those recipients prior to service and after service can be ascertained, can be very revealing. A treatment program for alcoholism that shows that 80% of its clients no longer drink six months after the program ends presents evidence more consistent with effectiveness than one showing only 25% abstaining. Of course, neither may, in fact, have real effects because the severity of their cases may differ and other independent influences on drinking may override any program effects. A good monitoring scheme, however, will also include indicators of the severity of the initial problem, exposure to other important influences, and the like. Although falling short of formal impact assessment, reasonable interpretation and comparison of patterns of such indicators and, especially, trends in those indicators as programs attempt to improve their effectiveness can provide useful indications of program performance.

PERSPECTIVES ON PROGRAM MONITORING

There is and should be considerable overlap in the purposes of program monitoring whether they are driven by the information needs of evaluators, program managers and staff, or policymakers, sponsors, and stakeholders. Ideally, the monitoring activities undertaken as part of evaluation should meet the information needs of all these groups. In practice, however, limitations on time and resources may require giving priority to one set of information needs over another. At the risk of overemphasizing the differences in outlook, for didactic purposes it is useful to delineate the perspectives of the three key "consumer groups" on the purposes of program monitoring.

Monitoring From the Evaluator's Perspective

A number of practical considerations underlie the need for evaluation researchers to monitor programs. All too often a program's impact is sharply diminished and, indeed, sometimes reduced to zero because the appropriate intervention was not delivered, was not delivered to the right targets, or both. In our estimation, more program failures are due to such implementation problems than to lack of potentially effective services. Monitoring studies, therefore, are essential to understanding and interpreting impact findings. Knowing what took place is a prerequisite for explaining or hypothesizing why a program did or did not work. Without monitoring, the evaluator is engaged in "black box" research with no basis for speculating whether a larger dose of the program or a different means of delivering the intervention would have changed the impact results.

Also, for program staff to improve a program, secure support (particularly for expanding a program), and counter critics, evaluations that demonstrate effective performance are often required. Many program evaluations, therefore, will be process evaluations that focus on service utilization and organizational issues. For evaluators who work within a social program or agency, developing and maintaining program monitoring systems are likely to be their major responsibility.

Finally, monitoring provides information necessary for program dissemination. The essential features of an effective intervention can be reproduced elsewhere only if the evaluation documentation can describe the program in operational detail. The critical points in implementation need to be identified, solutions to managerial problems outlined, qualifications of successful program personnel documented, and so on. Sound program development and evaluation include communicating these features in detail. The results of program monitoring at the development stage can be profitably used in the diffusion of effective and efficient programs.

Monitoring From an Accountability Perspective

Monitoring information is also critical for those who sponsor and fund programs. Program managers have a responsibility to inform their sponsors and funders of the activities undertaken, the degree of implementation of programs, the problems encountered, and what the future holds (see Exhibit 6-E for one perspective on this matter). However, evaluators frequently are mandated to provide the same or similar information. Indeed, in some cases the sponsors and funders of programs perceive program evaluators as "their eyes and ears," as a second line of information on what is going on in a particular program.

Government sponsors and funding groups, including Congress, operate in the glare of the mass media. Their actions are also visible to the legislative groups who authorize programs and to governmental "watchdog" organizations. For example, at the federal level, the Office of Management and Budget, part of the executive branch, wields considerable authority over program development, funding, and expenditures. The U.S. General Accounting Office, an arm of Congress, advises members of the House and Senate on the utility of programs and in some cases conducts its own evaluations. Both state governments and those of large cities have analogous oversight groups. No social program that receives outside funding, whether public or private, can expect to avoid scrutiny and escape demand for accountability.

In addition to funders and sponsors, other stakeholders may press for program accountability. In the face of taxpayers' reservations about spending for social programs, together with the increased competition for resources resulting from cuts in available funding, all stakeholders are scrutinizing both the programs they support and those they do not. Concerned parties use monitoring information to lobby for the expansion of programs they advocate or find congenial with their self-interests and the curtailment or abandonment of those programs they disdain. Stakeholders, it should be noted, include the targets themselves. A dramatic illustration of their perspective occurred when President Reagan telephoned an artificial heart transplant patient to wish him well and, with all of the country listening, the patient complained about not receiving his Social Security check.

EXHIBIT 6-E Program and Service Utilization Studies

Any service organization, especially in an era of shrinking resources, needs to evaluate its services and activities. Through these evaluative activities, an organization can develop and maintain the flexibility needed to respond to an ever-changing environment. It has been suggested that, even in an ideal world, an organization needs to be self-evaluating. Self-evaluation requires an organization continually to review its own activities and goals and to use the results to modify, if necessary, its programs, goals, and directions.

Within the agency, the essential function of evaluation is to provide data on goal achievement and program effectiveness to a primary audience consisting of administration, middle management, and governing board. This primary audience, especially the administration and board, is frequently confronted with inquiries from important sources in the external environment, such as legislators and funding agencies. These inquiries often focus on issues of client utilization, accessibility, continuity, comprehension, outcome or effectiveness, and cost.

The building block of this information is the *patterns of use* or *client utilization* study. The patterns of use study, whether it consists of simple inquiries or highly detailed, sophisticated investigations, is basically a description. It describes who uses services and how, and it becomes evaluative when it is related to the requirements or purposes of the organization.

SOURCE: Adapted from G. Landsberg, *Program Utilization and Service Utilization Studies: A Key Tool for Evaluation*, New Directions for Program Evaluation, no. 20 (San Francisco: Jossey-Bass, December 1983), pp. 93-103.

Clearly, social programs operate in a political world. It could hardly be otherwise, given the stakes involved. The human and social service industry is not only huge in dollar volume and number of persons employed but is also laden with ideological and emotional baggage. Programs are often supported or opposed by armies of vocal community members; indeed, the social program sector is comparable only to the defense industry in its lobbying efforts, and the stands that politicians take with respect to particular programs often determine their fates in elections. Accountability information is the major weapon that stakeholders use in their battles as advocates and antagonists.

Monitoring From a Management Perspective

Management-oriented monitoring (including use of MISs) often is concerned with the same questions as evaluation and program accountability studies; the differences lie in the purposes to which the findings are to be put. Evaluators' interest in monitoring data generally centers on determining how a program's impact is related to its implementation. Accountability studies primarily provide information that decisionmakers, sponsors, and other stakeholders need to judge the appropriateness of program activities and to decide whether a program should be continued, expanded, or contracted. Such studies may use the same information base employed by program management staff, but they are usually conducted in a critical spirit. In contrast, management-oriented monitoring activities are concerned less with making decisive judgments and more with incorporating corrective measures as a regular part of program operations.

Monitoring from a management perspective is particularly vital during the implementation and pilot testing of new programs, especially innovative ones. No matter how well planned such programs may be, unexpected results and unwanted side effects often surface early in the course of implementation. Program designers and managers need to know rapidly and fully about these problems so that changes can be made as soon as possible in the program design. Suppose, for example, a medical clinic intended to help working mothers is open only during daylight hours. Monitoring may disclose that however great the demand is for clinic services, the clinic's hours of operation effectively screen out most of the target population. Or suppose that a program is predicated on the assumption that severe psychological problems are prevalent among children who act out in school. If it is found early on that most such children do not in fact have serious disorders, the program can be modified accordingly.

For programs that have moved beyond the development stage to actual operation, program monitoring serves management needs by providing information on coverage and process, and hence feedback on whether the program is meeting specifications. Fine-tuning of the program may be necessary when monitoring information indicates that targets are not being reached, that the implementation of the program costs more than initially projected, or that staff workloads are either too heavy or too light. Managers who neglect to monitor a program fully and systematically risk the danger of administering a program that is markedly different from its mandate.

Where monitoring information is to be used for both managerial and evaluation purposes, some problems must be anticipated. How much information is sensible to collect and report, in what forms, at what frequency, with what reliability, and with what degree of confidentiality are among the major issues on which evaluators and managers may disagree. For example, the experienced manager of a nonprofit children's recreational program may feel the highest priority is weekly information on attendance, which is added to graphs for the program's governing board. The evaluator, however, may be comfortable with aggregating the data monthly or even quarterly, but may believe that before being reported they should be adjusted to take into account variations in the weather, occurrence of holidays, and so on—even though the necessary adjustments require the use of sophisticated statistical procedures.

A second concern is the matter of proprietary claims on the data. For the manager, monitoring data on, say, the results of a program innovation should be kept confidential until discussed with the research committee of the board of directors and presented at the board meeting. The evaluator may wish immediately to write a paper for publication in the *American 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 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.

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.

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

EXHIBIT 6-I 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.

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 experi-

ence 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 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.

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.

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 observations made (see Miles and Huberman, 1994, and Patton, 1990, for guidance on observational methods). Observers must be trained in how to make observations and how to record them uniformly. There are three common ways of making systematic observations. The first approach, known as the narrative method, involves the least structuring: The observer is simply asked to record events in as much detail as possible and in the order in which they occur. Typically, observers are provided with a list of important types of activities to which their attention should be directed.

A more structured approach is to provide observers with a data guide: a set of questions they are required to answer from their observations or a checklist on which to record the different activities observed. A data guide physically can resemble a survey instrument. For example, a data guide for observers attending technical training classes may have questions such as "How did the instructor make use of available training aids?" Or it may call for ratings, for instance, regarding the clarity of the instructor's presentation. Some ratings may be purely descriptive. Others may call for expert judgments, such as a scale to assess the way a police officer handles an encounter with a suspect. Structured recording instruments simplify analysis considerably and increase the likelihood that observers will provide consistent information.

Evaluators may also arrange for program staff to generate monitoring data according to the evaluator's specifications. Sometimes staff are asked to provide narrative reports in the form of diaries; sometimes they are asked to complete data forms or questionnaires for the evaluator. The evaluator may also survey or interview staff about certain aspects of their experiences, observations, or activities. The most efficient approach is to use a structured questionnaire or data form that can be completed by interview or by the staff person alone. As with observational data, structured instruments lend themselves readily to tabulation. It is also generally wise to organize the data collection so as to minimize the work demanded by program staff and, correspondingly, their resistance to cooperating with the data collection.

In some circumstances, it is possible to reduce the data collection burden by developing adequate sampling approaches. This may allow one or a few observers to record project activities in an economical fashion or an efficient number of staff persons to provide data. Sometimes sampling is done by randomly selecting time periods for observation; in other instances, it is more appropriate to sample persons or events. In doing so, it is important to ensure that a representative sample is employed to avoid intentional or unintentional bias in the information obtained.

Service Record Data

We have already seen how program records can be used to assess the participation of targets in a program. Often the delivery of project services can also be monitored from service records. Exhibit 6-L, for example, describes the use of medical charts to track service delivery in a program providing primary medical care to homeless individuals; the information gathered was also useful for several other purposes.

Service records vary; they can be the equivalent of narrative reports or highly structured data forms on which project personnel check which services were given, how they were received, and the observable results. Their level of detail is related to the complexity of the project and to the number of alternatives that can be specified in advance. Often service rec-

EXHIBIT 6-L Using Clinic Encounter Records to Track Delivery of Medical Services to the Homeless

In the mid-1980s the Robert Wood Johnson Foundation and the Pew Memorial Trust funded 19 medical clinical programs serving homeless persons in as many cities throughout the country. To keep track of the medical problems encountered and the services delivered, each time a client was served, the attending medical person filled out a standard "encounter" record, with information about the person served, the medical condition, and the treatment prescribed. The encounter sheets also contained identifying information enabling the tracking of specific individuals throughout any number of subsequent clinic visits.

The records were sent each month to the Social and Demographic Institute at the University of Massachusetts where they were entered into a database. At the end of three years of clinic operation, more than 290,000 encounters had been entered into the database, pertaining to almost 94,000 individual clients served.

The database was used to document the incidence of various medical conditions as well as variations among sites and to identify the medical care needs of the homeless. In addition, the records were also used by each of the sites to monitor clinic activities.

SOURCE: Adapted from James D. Wright and Eleanor Weber, *Homelessness and Health* (New York: McGraw-Hill, 1987).

ord systems have serious limitations for monitoring purposes. In record systems designed primarily to serve administrative needs, the records are often not filled in completely if parts are viewed as irrelevant to staff for their purposes. If monitoring components are added, they may seem overly burdensome to program staff and they may not cooperate.

On the other hand, record information is inexpensive and efficient to obtain and analyze. Its use for monitoring depends on adequately training program staff, on providing the staff with motivation to complete records properly, and on incorporating quality-control checks to ensure that they follow through. A few items of data gathered consistently and reliably are generally much better for monitoring purposes than a more comprehensive set of information of doubtful reliability collected inconsistently.

It is also usually helpful if the record forms are structured as checklists whenever possible so that program staff can simply check off various items rather than provide narrative information. This procedure not only minimizes the time required of project staff but also yields data in a form convenient for analysis. In addition, if completed records are reviewed carefully as soon as possible for consistency and accuracy, omissions and inconsistencies can be caught in a timely way.

Again, it is important to recognize the risks involved in using service records as the only data source. Program staff, intentionally or unintentionally, may exaggerate the extent to which program elements are being delivered to targets because they are overly zealous about maintaining appearances of effectiveness or are simply displaying ritual compliance in what

they record. There are also occasions when the project staff's interpretation of a particular service differs from that of the program's designers or evaluators.

Management Information Systems

The introduction of MISs into the social program arena provides new opportunities for effective monitoring. In a sense, all record systems are MISs. However, the concept is usually reserved for those systems that organize information using computers and allow it to be accumulated and displayed in a variety of ways at specified periods or on demand. An MIS thus provides information on an ongoing basis that can be used for program managers' decision making, for reports produced for stakeholders, and for evaluation purposes.

For example, a community mental health center may have 5,000 clients, see 600 patients a week, provide 15 services, refer patients to 12 providers outside the center, and have 22 professionals treating patients within the center, including psychiatrists, social workers, psychologists, psychiatric nurses, and vocational counselors. The patients may vary in age, sex, ethnicity, length and outcome of treatment, diagnoses, and a host of other characteristics. Program managers might well be interested in ascertaining any or all of these features and their interrelationships. They may want to know on a monthly basis the average number of patient visits by diagnosis, ethnicity, sex, or age. They might want to know what types of patients are being treated by which types of personnel or what types are receiving which kinds of services. The number of permutations of even these few measures is huge. Thus, MISs require computers for storing data and retrieving information in a variety of combinations.

Typically, computer-based MISs produce tables periodically (e.g., monthly) containing information regularly used by staff and management. They may produce other tables on different schedules for sponsors and stakeholders. For example, a mental health center's MIS may produce a second set of tables quarterly to send to the county agency that provides its support and an annual set of tables to send to the National Institute of Mental Health in Washington. These tables may differ in the ways the data are accumulated, the summary statistics provided, and so on.

In addition, an MIS can be used to answer specific management and research questions. For example, the mental health center's director may become uneasy about the proportion of patients who drop out of treatment. She might want to see whether the dropouts cluster by ethnicity, by the provider treating them, or some other characteristic. Or suppose a university-based clinical psychologist has secured funds to undertake an innovative demonstration program with depressive young adults and wants to include the center as one of the sites if it has a large enough target population. The MIS could provide the needed information by reporting diagnoses by age. Similarly, the MIS can supply information requested by stakeholders; for example, the local mental health association may want to know whether elderly patients are provided with psychotherapy and rehabilitation services rather than drug therapy.

From the evaluator's perspective, two aspects of MISs can present problems. First, it is essential that the system include the information that will be required by the various users, including researchers and evaluators. For example, information about the full mix and amount of services for each patient will be important for evaluation and obviously must be entered into the system to be available. The task involves more than simply identifying the information components, however; each must be operationally defined and rules must be developed for entering and accessing information. If 97% of a center's patients receive three or fewer distinct services, the system may store only three service codes per patient. In this case, a rule must be stated that specifies which three are entered for those few patients who have four or more.

The second and perhaps more critical consideration is that all the persons who provide and enter data must understand the utility of the system, its rules and definitions, and their responsibility to collaborate in its implementation. The finest hardware and software and the most sophisticated, well-conceptualized system will be useless if service providers do not take the time to enter the required data after seeing each patient. If providers wait until the end of the day to put in what they remember, for instance, the result will be what those in the business refer to as a "gigo" system ("garbage in, garbage out"). A combination of lack of training, apathy, fear of the system's revealing negative information, and occasional sheer malice need to be overcome if the organization is to reap the benefits of a properly functioning MIS. Thus, realizing the potential of the system requires training, oversight, regular quality control procedures, sanctions, and tender loving care.

Program Participant Data

The final approach to collecting monitoring information is to obtain data about program performance directly from participants themselves. Such information is valuable for a number of reasons, among them the distinctive

perspective participants often have on a program. Securing participant data may be necessary for providers to know what is important to clients, including their satisfaction with and understanding of the intervention. Moreover, it may be the only way of finding out what was actually delivered.

In many programs, the services provided are not identical with those actually received and used by the participants. The literature on family planning has shown, for example, that participants may receive study guides, exercises, manuals, and equipment for use outside the classroom that are not used as intended. Thus, it may be critical to query participants to find out whether specific services were used or even received. For interventions involving complex interventions, it may also be important to ascertain participants' understanding of the interventions, the program's operating rules, and so on. In short, it is necessary to establish not only that designated services have been delivered but also that they were received, used, and understood as intended.

In our discussion of access earlier in this chapter, we pointed out that there are times when participants' satisfaction with a program is a key indicator in monitoring program implementation. Clearly, in this case the participant is the appropriate and sole information source. Information from participants must necessarily be obtained by interviews or self-administered questionnaires. Participants may be sampled in some systematic way, or an entire census may be conducted.

ANALYSIS OF MONITORING DATA

Data, of course, are useful only when they have been appropriately analyzed. In general, the analysis of monitoring data addresses the following three issues: description of the project, comparison between sites, and conformity of the program to its design.

Description of the Program Performance

Assessing the extent to which a program as implemented resembles the program as designed depends on having a full and accurate description of how the program actually operated. A description derived from monitoring data would cover the following topics: estimates of coverage and bias in participation, the types of services delivered, the intensity of services given to participants of significant kinds, and the reactions of participants to the services delivered. Descriptive statements might take the form of narrative accounts, especially when monitoring data are derived from qualitative sources, or quantitative summaries in the form of tables, graphs, and the like.

Comparison Between Sites

When a program includes more than one site, a second question concerns differences in program implementation between the sites. Comparison of sites permits an understanding of the sources of diversity in program implementation and outcomes, such as differences in staff, administration, targets, or surrounding environments, and it also can facilitate efforts to achieve standardization. In addition, between-site differences may provide clues as to why programs at some sites are more effective than those at others.

Conformity of the Program to Its Design

The third issue is the one with which we began: the degree of conformity between a program's design and its implementation. Shortfalls may occur because the program is not performing functions it is expected to or because it is not performing them as well as expected. Such discrepancies may lead to efforts to move the implementation of a project closer to the original design or to a respecification of the design itself. Such analysis also provides an opportunity to judge the appropriateness of impact evaluation and, if necessary, to opt for more formative evaluation to develop the desired convergence of design and implementation.

SUMMARY

- Program monitoring is a form of evaluation designed to describe how a program is operating and assess how well it performs its intended functions. It builds on program theory, which identifies the critical components, functions, and relationships assumed necessary for the program to be effective.
- The results of program monitoring allow performance to be assessed against the stipulations of program theory, administrative standards, applicable legal, ethical, or professional standards, and after-the-fact judgment calls.
- The common forms of program monitoring include process (or implementation) evaluation, management information systems (MISs), and performance measurement.
- Process evaluation assesses whether the program is delivered as intended to the targeted recipients and is typically conducted as a separate project by evaluation specialists. It may constitute a stand-alone evaluation when the only questions are about implementation of program operations, service delivery, and other such matters. Process evaluation is also often carried out in conjunction with an impact evaluation to determine what services the program provides to complement findings about what impact those services have.
- When program monitoring is integrated into a program's routine information collection and reporting, it constitutes an MIS. In such systems, data relating to program process and service utilization is obtained, compiled, and periodically summarized for review.
- Performance measurement refers to various program monitoring schemes developed in response to demands for accountability from public and nonprofit agencies. The most far-reaching of these is the Government Performance and Results Act of 1993 (GPRA), which requires federal agencies to identify their program goals and report on their results in attaining those goals.
- Performance measurement distinguishes program *outputs*, the products or services delivered to program participants, from program *outcomes*, the results of those activities, such as improved health for the individuals served. Performance measurement is designed to periodically report results on indicators of the quantity and quality of both outputs and outcomes.
- Program monitoring takes somewhat different forms and serves different purposes when undertaken from the perspectives of evaluation, accountability, and program management, but the types of data required and the data collection procedures used generally are the same or overlap considerably. In particular, it generally involves one or more of three relatively distinct domains of program performance: service utilization, organizational functions, or program outcomes.

- Service utilization issues typically break down into questions about coverage and bias. Coverage relates to how fully the target population participates in the program and bias relates to differential participation among those with different characteristics, for example, resistance to service, sociodemographic attributes, diagnosis, or location. The sources of data useful for assessing coverage are program records, surveys of program participants, and community surveys. Bias in program coverage can be revealed through comparisons of program users, eligible nonparticipants, and dropouts.
- Monitoring a program's organizational functions focuses on how well the program is organizing its efforts and using its resources to accomplish its essential tasks. Particular attention is given to identifying shortcomings in program implementation that prevent a program from delivering the intended services to the target population. Three sources of such implementation failures are incomplete interventions, delivery of the wrong intervention, and unstandardized or uncontrolled interventions.
- Program 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. Outcome monitoring requires that indicators be identified that are practical to collect routinely and informative with regard to program results. Because there are usually many influences on a social condition that are not under the program's control, finding outcome indicators that isolate results attributable to the program is often difficult without the special techniques of impact evaluation.
- Because of the dynamic nature of the social conditions typical programs attempt to affect, the limitations of outcome indicators, and the pressures on program agencies, there are many pitfalls associated with program monitoring. These include program distortions resulting from attempts to look good on inappropriate indicators, corruption of the indicators so they overstate performance, and misinterpretation of what indicators reveal about actual program performance.
- The data used for monitoring purposes are generally collected from three sources: data collected directly by the evaluator, service records, and program participants. In recent years, MISs have become an essential tool for organizing, storing, and retrieving data from program records in ways that serve the needs of multiple users.
- The analysis of monitoring data typically addresses such issues as description of program operations, comparison of sites, conformity of a program to its design, and program performance relative to standards or expectations.