7

Measuring and Monitoring Program Outcomes

Chapter Outline

The previous chapter discussed how a program's process and performance can be monitored. The ultimate goal of all programs, however, is not merely to function well, but to bring about change—to affect some problem or social condition in beneficial ways. The changed conditions are the intended outcomes or products of the programs. Assessing the degree to which a program produces these outcomes is a core function of evaluators.

A program's intended outcomes are ordinarily identified in the program's impact theory. Sensitive and valid measurement of those outcomes is technically challenging but essential to assessing a program's success. In addition, ongoing monitoring of outcomes can be critical to effective program management. Interpreting the results of outcome measurement and monitoring, however, presents a challenge to stakeholders because a given set of outcomes can be produced by factors other than program processes. This chapter describes how program outcomes can be identified, how they can be measured and monitored, and how the results can be properly interpreted.

Assessing a program's effects on the clients it serves and the social conditions it aims to improve is the most critical evaluation task because it deals with the "bottom line" issue for social programs. No matter how well a program addresses target needs, embodies a good plan of attack, reaches its target population and delivers apparently appropriate services, it cannot be judged successful unless it actually brings about some measure of beneficial change in its given social arena. Measuring that beneficial change, therefore, is not only a core evaluation function but also a high-stakes activity for the program. For these reasons, it is a function that evaluators must accomplish with great care to ensure that the findings are valid and properly interpreted. For these same reasons, it is one of the most difficult and, often, politically charged tasks the evaluator undertakes.

Beginning in this chapter and continuing through Chapter 10, we consider how best to identify the changes a program should be expected to produce, how to devise measures of these changes, and how to interpret such measures. Consideration of program effects begins with the concept of a program *outcome,* so we first discuss that pivotal concept.

Program Outcomes

An **outcome** is the state of the target population or the social conditions that a program is expected to have changed. For example, the amount of smoking among teenagers after

exposure to an antismoking campaign in their high school is an outcome. The attitudes toward smoking of those who had not yet started to smoke is also an outcome. Similarly, the "school readiness" of children after attending a preschool program would be an outcome, as would the body weight of people who completed a weight-loss program, the management skills of business personnel after a management training program, and the amount of pollutants in the local river after a crackdown by the local environmental protection agency.

Notice two things about these examples. First, outcomes are observed characteristics of the target population or social conditions, not of the program, and the definition of an outcome makes no direct reference to program actions. Although the services delivered to program participants are often described as program "outputs," *outcomes,* as defined here, must relate to the *benefits* those products or services might have for the participants, not simply their receipt. Thus, "receiving supportive family therapy" is not a program outcome in our terms but, rather, the delivery of a program service. Similarly, providing meals to 100 housebound elderly persons is not a program outcome; it is service delivery, an aspect of program process. The nutritional benefits of those meals for the health of the elderly, on the other hand, are outcomes, as are any improvements in their morale, perceived quality of life, and risk of injury from attempting to cook for themselves. Put another way, outcomes always refer to characteristics that, in principle, could be observed for individuals or situations that have not received program services. For instance, we could assess the amount of smoking, the school readiness, the body weight, the management skills, and the water pollution in relevant situations where there was no program intervention. Indeed, as we will discuss later, we might measure outcomes in these situations to compare with those where the program was delivered.

Second, the concept of an outcome, as we define it, does not necessarily mean that the program targets have actually changed or that the program has caused them to change in any way. The amount of smoking by the high school teenagers may not have changed since the antismoking campaign began, and nobody may have lost any weight during their participation in the weight-loss program. Alternatively, there may be change but in the opposite of the expected direction—the teenagers may have increased their smoking, and program participants may have gained weight. Furthermore, whatever happened may have resulted from something other than the influence of the program. Perhaps the weight-loss program ran during a holiday season when people were prone to overindulge in sweets. Or perhaps the teenagers decreased their smoking in reaction to news of the smoking-related death of a popular rock music celebrity. The challenge for evaluators, then, is to assess not only the outcomes that actually obtain but also the degree to which any change in outcomes is attributable to the program itself.

Outcome Level, Outcome Change, and Net Effect

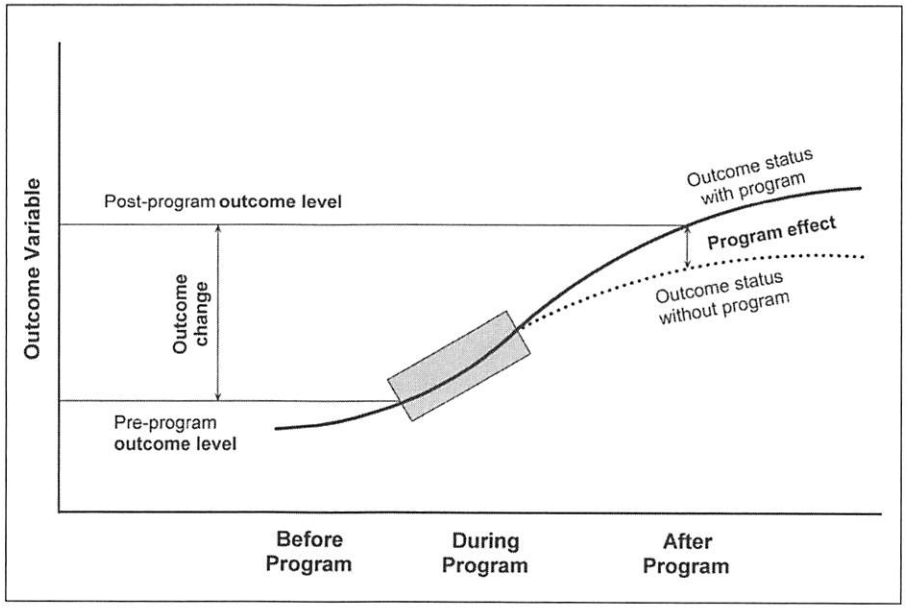
The foregoing considerations lead to important distinctions in the use of the term *outcome:*

- **Outcome level** is the status of an outcome at some point in time (e.g., the amount of smoking among teenagers).
- **Outcome change** is the difference between outcome levels at different points in time.
- **Program effect** is that portion of an outcome change that can be attributed uniquely to a program as opposed to the influence of some other factor.

Consider the graph in Exhibit 7-A, which plots the levels of an outcome measure over time. The vertical axis represents an *outcome variable* relevant to a program we wish to evaluate. An outcome variable is a measurable characteristic or condition of a program's target population that could be affected by the actions of the program. It might be amount of smoking, body weight, school readiness, extent of water pollution, or any other outcome falling under the definition above. The horizontal axis represents time, specifically, a period ranging from before the program was delivered to its target population until some time afterward. The solid line in the graph shows the average outcome level of a group of individuals who received program services. Note that their status over time is not depicted as a straight horizontal line but, rather, as a line that wiggles around. This is to indicate that smoking, school readiness, management skills, and other such outcome dimensions are not expected to stay constant—they change as a result of many natural causes and circumstances quite extraneous to the program. Smoking, for instance, tends to increase from the preteen to the teenage years. Water pollution levels may fluctuate according to the industrial activity in the region and weather conditions, for example, heavy rain that dilutes the concentrations.

If we measure the outcome variable (more on this shortly), we can determine how high or low the target group is with respect to that variable, for example, how much smoking or school readiness they display. This tells us the *outcome level,* often simply called the outcome. When measured after the target population has received program services, it tells us something about how that population is doing—how many teenagers are smoking, the average level of school readiness among the preschool children, how many pollutants there are in the water. If all the teenagers are smoking, we may be disappointed, and, conversely, if none are smoking, we may be pleased. All by themselves, however, these outcome levels do not tell us much about how effective the program was, though they may constrain the possibilities. If all the teens are smoking, for instance, we can be fairly sure that the antismoking program was not a great

EXHIBIT 7-A
Outcome Level, Outcome Change, and Program Effect



success and possibly was even counterproductive. If none of the teenagers are smoking, that finding is a strong hint that the program has worked because we would not expect them all to spontaneously stop on their own. Of course, such extreme outcomes are rarely found and, in most cases, outcome levels alone cannot be interpreted with any confidence as indicators of a program's success or failure.

If we measure outcomes on our target population before and after they participate in the program, we can describe more than the outcome level, we can also discern outcome *change*. If the graph in Exhibit 7-A plotted the school readiness of children in a preschool program, it would show that the children show less readiness before participating in the program and greater readiness afterward, a positive change. Even if their school readiness after the program was not as high as the preschool teachers hoped it would be, the direction of before-after change shows that there was improvement. Of course, from this information alone, we do not actually know that the preschool program had anything to do with the children's improvement in school readiness. Preschool-aged children are in a developmental period when their cognitive and motor skills increase rather rapidly through normal maturational processes. Other factors may also be at work; for example, their parents may be reading to them and otherwise supporting their intellectual development and preparation for entering school, and that may account for at least part of their gain.

The dashed line in Exhibit 7-A shows the trajectory on the outcome variable that would have been observed if the program participants had not received the program. For the preschool children, for example, the dashed line shows how their school readiness would have increased if they had not been in the preschool program. The solid line shows how school readiness developed when they were in the program. A comparison of the two lines indicates that school readiness would have improved even without exposure to the program, but not quite as much.

The difference between the outcome level attained with participation in the program and that which the same individuals would have attained had they not participated is the part of the change in outcome that the program produced. This is the value added or net gain part of the outcome that would not have occurred without the program. We refer to that increment as the program effect or, alternatively, the program impact. It is the only part of the outcome for which the program can honestly take credit.

Estimation of the program effect, or impact assessment, is the most demanding evaluation research task. The difficulties are highlighted in Exhibit 7-A, where the program effect is shown as the difference between the outcome that actually occurred and the outcome that would have occurred in the absence of the program. It is, of course, impossible to simultaneously observe outcomes for the same people (or other entities) under conditions when they both receive and do not receive a program. We must, therefore, observe the outcome after program participation and then somehow estimate what that outcome would have been without the program. Because the latter outcome is hypothetical for individuals who, in fact, did receive the program, it must be inferred rather than measured or observed. Developing valid inferences under these circumstances can be difficult and costly. Chapters 8 and 9 describe the methodological tools evaluators have available for this challenging task.

Although outcome levels and outcome changes have quite limited uses for determining program effects, they are of some value to managers and sponsors for monitoring program performance. This application will be discussed later in this chapter. For now we continue our exploration of the concept of an outcome by discussing how outcomes can be identified, defined, and measured for the purposes of evaluation.

Identifying Relevant Outcomes

The first step in developing measures of program outcomes is to identify very specifically what outcomes are relevant candidates for measurement. To do this, the evaluator must consider the perspectives of stakeholders on expected outcomes, the outcomes that are specified in the program's impact theory, and relevant prior research. The evaluator will also need to give attention to unintended outcomes that may be produced by the program.

Stakeholder Perspectives

Various program stakeholders have their own understanding of what the program is supposed to accomplish and, correspondingly, what outcomes they expect it to affect. The most direct sources of information about these expected outcomes usually are the stated objectives, goals, and mission of the program. Funding proposals and grants or contracts for services from outside sponsors also often identify outcomes that the program is expected to influence.

A common difficulty with information from these sources is a lack of the specificity and concreteness necessary to clearly identify specific outcome measures. It thus often falls to the evaluator to translate input from stakeholders into workable form and negotiate with the stakeholders to ensure that the resulting outcome measures capture their expectations.

For the evaluator's purposes, an outcome description must indicate the pertinent characteristic, behavior, or condition that the program is expected to change. However, as we discuss shortly, further specification and differentiation may be required as the evaluator moves from this description to selecting or developing measures of this outcome. Exhibit 7-B presents examples of outcome descriptions that would usually be serviceable for evaluation purposes.

Program Impact Theory

A full articulation of the program impact theory, as described in Chapter 5, is especially useful for identifying and organizing program outcomes. An impact theory expresses the outcomes of social programs as part of a logic model that connects the program's activities to proximal (immediate) outcomes that, in turn, are expected to lead to other, more distal outcomes. If correctly described, this series of linked relationships among outcomes represents the program's assumptions about the critical steps between program services and the ultimate social benefits the program is intended to produce. It is thus especially important for the evaluator to draw on this portion of program theory when identifying those outcomes that should be considered for measurement.

Exhibit 7-C shows several examples of the portion of program logic models that describes the impact theory (additional examples are found in Chapter 5). For the purposes of outcome assessment, it is useful to recognize the different character of the more proximal and more distal outcomes in these sequences. Proximal outcomes are those that the program services are expected to affect most directly and immediately. These can be thought of as the "take away" outcomes—those the program participants experience as a direct result of their participation and take with them out the door as they leave. For most social programs, these proximal outcomes are

EXHIBIT 7-B
Examples of Outcomes Described Specifically Enough to Be Measured

Juvenile delinquency
Behavior of youths under the age of 18 that constitute chargeable offenses under applicable laws irrespective of whether the offenses are detected by authorities or the youth is apprehended for the offense.

Contact with antisocial peers
Friendly interactions and spending time with one or more youths of about the same age who regularly engage in behavior that is illegal and/or harmful to others.

Constructive use of leisure time
Engaging in behavior that has educational, social, or personal value during discretionary time outside of school and work.

Water quality
The absence of substances in the water that are harmful to people and other living organisms that drink the water or have contact with it.

Toxic waste discharge
The release of substances known to be harmful into the environment from an industrial facility in a manner that is likely to expose people and other living organisms to those substances.

Cognitive ability
Performance on tasks that involve thinking, problem solving, information processing, language, mental imagery, memory, and overall intelligence.

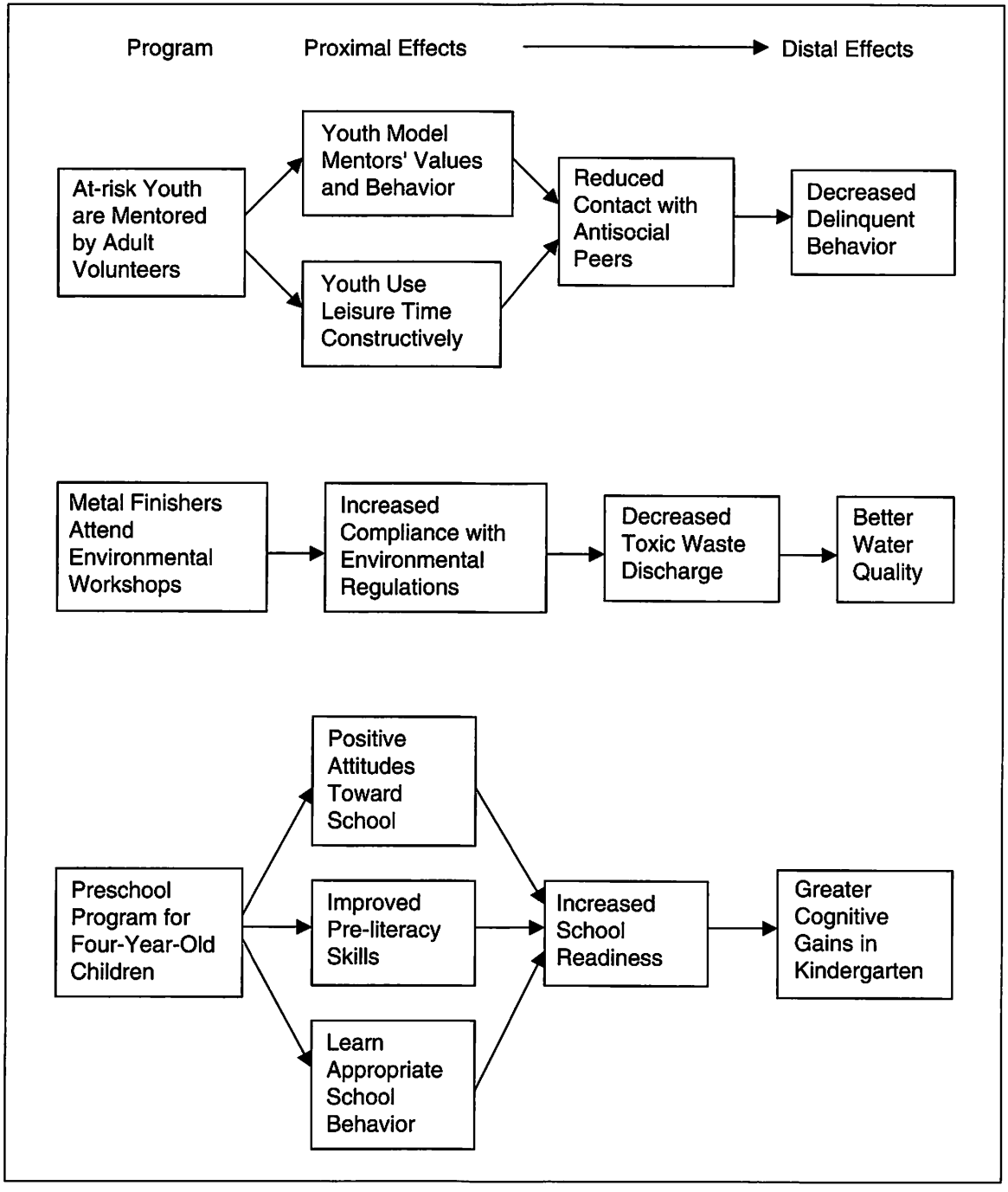
School readiness
Children's ability to learn at the time they enter school; specifically, the health and physical development, social and emotional development, language and communication skills, and cognitive skills and general knowledge that enable a child to benefit from participation in formal schooling.

Positive attitudes toward school
A child's liking for school, positive feelings about attending, and willingness to participate in school activities.

psychological—attitudes, knowledge, awareness, skills, motivation, behavioral intentions, and other such conditions that are susceptible to relatively direct influence by a program's processes and services.

Proximal outcomes are rarely the ultimate outcomes the program intends to generate, as can be seen in the examples in Exhibit 7-C. In this regard, they are not the most important outcomes from a social or policy perspective. This does not mean, however,

EXHIBIT 7-C
Examples of Program Impact Theories Showing Expected Program Effects on Proximal and Distal Outcomes



that they should be overlooked in the evaluation. These outcomes are the ones the program has the greatest capability to affect, so it can be very informative to know whether they are attained. If the program fails to produce these most immediate and direct outcomes, and the program theory is correct, then the more distal outcomes in the sequence are unlikely to occur. In addition, the proximal outcomes are generally the easiest to measure and to attribute to the program's efforts. If the program is successful at generating these outcomes, it is appropriate for it to receive credit for doing so. The more distal outcomes, which are more difficult to measure and attribute, may yield ambiguous results. Such results will be more balanced and interpretable if information is available about whether the proximal outcomes were attained.

Nonetheless, it is the more distal outcomes that are typically the ones of greatest practical and political importance. It is thus especially important to clearly identify and describe those that can reasonably be expected to result from the program activities. The value of careful development of the impact theory for these purposes is that it provides the basis for assessing what outcomes are actually reasonable, given the nature of the program.

Generally, however, a program has less direct influence on the distal outcomes in its impact theory. In addition, distal outcomes are also influenced by many other factors outside of the program's control. This circumstance makes it especially important to define the expected distal outcomes in a way that aligns as closely as possible with the aspects of the social conditions that the program activities can affect. Consider, for instance, a tutoring program for elementary school children that focuses mainly on reading, with the intent of increasing educational achievement. The educational achievement outcomes defined for an evaluation of this program should distinguish between those closely related to reading skills and those areas, such as mathematics, that are less likely to be influenced by what the program is actually doing.

Prior Research

In identifying and defining outcomes, the evaluator should thoroughly examine prior research on issues related to the program being evaluated, especially evaluation research on similar programs. Learning which outcomes have been examined in other studies may call attention to relevant outcomes that might otherwise have been overlooked. It will also be useful to determine how various outcomes have been defined and measured in prior research. In some cases, there are relatively standard definitions and measures that have an established policy significance. In other cases, there may be known problems with certain definitions or measures that the evaluator will need to know about.

Unintended Outcomes

So far, we have been considering how to identify and define those outcomes the stakeholders expect the program to produce and those that are evident in the program's impact theory. There may be significant unintended outcomes of a program, however, that will not be identified through these means. These outcomes may be positive or negative, but their distinctive character is that they emerge through some process that is not part of the program's design and direct intent. That feature, of course, makes them very difficult to anticipate. Accordingly, the evaluator must often make a special effort to identify any potential unintended outcomes that could be significant for assessing the program's effects on the social conditions it addresses.

Prior research can often be especially useful on this topic. There may be outcomes that other researchers have discovered in similar circumstances that can alert the evaluator to possible unanticipated program effects. In this regard, it is not only other evaluation research that is relevant but also any research on the dynamics of the social conditions in which the program intervenes. Research about the development of drug use and the lives of users, for instance, may provide clues about possible responses to a program intervention that the program plan has not taken into consideration.

Often, good information about possible unintended outcomes can be found in the firsthand accounts of persons in a position to observe those outcomes. For this reason, as well as others we have mentioned elsewhere in this text, it is important for the evaluator to have substantial contact with program personnel at all levels, program participants, and other key informants with a perspective on the program and its effects. If unintended outcomes are at all consequential, there should be someone in the system who is aware of them and who, if asked, can alert the evaluator to them. These individuals may not present this information in the language of unintended outcomes, but their descriptions of what they see and experience in relation to the program will be interpretable if the evaluator is alert to the possibility that there could be important program effects not articulated in the program logic or intended by the core stakeholders.

Measuring Program Outcomes

Not every outcome identified through the procedures we have described will be of equal importance or relevance, so the evaluator does not necessarily need to measure all of them in order to conduct an evaluation. Instead, some selection may be appropriate. In addition, some important outcomes—for example, very long-term ones—may be quite

difficult or expensive to measure and, consequently, may not be feasible to include in the evaluation.

Once the relevant outcomes have been chosen and a full and careful description of each is in hand, the evaluator must next face the issue of how to measure them. Outcome measurement is a matter of representing the circumstances defined as the outcome by means of observable indicators that vary systematically with changes or differences in those circumstances. Some program outcomes have to do with relatively simple and easily observed circumstances that are virtually one-dimensional. One intended outcome of an industrial safety program, for instance, might be that workers wear their safety goggles in the workplace. An evaluator can capture this outcome quite well for each worker at any given time with a simple observation and recording of whether or not the goggles are being worn—and, by making periodic observations, extend the observation to indicate how frequently they are worn.

Many important program outcomes, however, are not as simple as whether a worker is wearing safety goggles. To fully represent an outcome, it may be necessary to view it as multidimensional and differentiate multiple aspects of it that are relevant to the effects the program is attempting to produce. Exhibit 7-B, for instance, provides a description of juvenile delinquency in terms of legally chargeable offenses committed. The chargeable delinquent offenses committed by juveniles, however, have several distinct dimensions that could be affected by a program attempting to reduce delinquency. To begin with, both the frequency of offenses and the seriousness of those offenses are likely to be relevant. Program personnel would not be happy to discover that they had reduced the frequency of offenses but that those still committed were now much more serious. Similarly, the type of offense may require consideration. A program focusing on drug abuse, for example, may expect drug offenses to be the most relevant outcome, but it may also be sensible to examine property offenses, because drug abusers may commit these to support their drug purchases. Other offense categories may be relevant, but less so, and it would obscure important distinctions to lump all offense types together as a single outcome measure.

Most outcomes are multidimensional in this way; that is, they have various facets or components that the evaluator may need to take into account. The evaluator generally should think about outcomes as comprehensively as possible to ensure that no important dimensions are overlooked. This does not mean that all must receive equal attention or even that all must be included in the coverage of the outcome measures selected. The point is, rather, that the evaluator should consider the full range of potentially relevant dimensions before determining the final measures to be used. Exhibit 7-D presents several examples of outcomes with various aspects and dimensions broken out.

One implication of the multiple dimensions of program outcomes is that a single outcome measure may not be sufficient to represent their full character. In the case of

EXHIBIT 7-D
Examples of the Multiple Dimensions and Aspects That Constitute Outcomes

Juvenile delinquency

- Number of chargeable offenses committed during a given period
- Severity of offenses
- Type of offense: violent, property crime, drug offenses, other
- Time to first offense from an index date
- Official response to offense: police contact or arrest; court adjudication, conviction, or disposition

Toxic waste discharge

- Type of waste: chemical, biological; presence of specific toxins
- Toxicity, harmfulness of waste substances
- Amount of waste discharged during a given period
- Frequency of discharge
- Proximity of discharge to populated areas
- Rate of dispersion of toxins through aquifers, atmosphere, food chains, and the like

Positive attitudes toward school

- Liking for teacher
- Liking for classmates
- Liking for school activities
- Willingness to go to school
- Voluntary participation in school activities

juveniles' delinquent offenses, for instance, the evaluation might use measures of offense frequency, severity, time to first offense after intervention, and type of offense as a battery of outcome measures that would attempt to fully represent this outcome. Indeed, multiple measures of important program outcomes help the evaluator guard against missing an important program accomplishment because of a narrow measurement strategy that leaves out relevant outcome dimensions.

Diversifying measures can also safeguard against the possibility that poorly performing measures will underrepresent outcomes and, by not measuring the aspects of the outcome a program most affects, make the program look less effective than it

actually is. For outcomes that depend on observation, for instance, having more than one observer may be useful to avoid the biases associated with any one of them. For instance, an evaluator who was assessing children's aggressive behavior with their peers might want the parents' observations, the teacher's observations, and those of any other person in a position to see a significant portion of the child's behavior. An example of multiple measures is presented in Exhibit 7-E.

EXHIBIT 7-E
Multiple Measures of Outcomes

A community intervention to prevent adolescent tobacco use in Oregon included youth anti-tobacco activities (e.g., poster and T-shirt giveaways) and family communication activities (e.g., pamphlets to parents). In the impact assessment the outcomes were measured in a variety of ways:

Outcomes for youths

- Attitudes toward tobacco use
- Knowledge about tobacco
- Reports of conversations about tobacco with parents
- Rated intentions to smoke or chew tobacco
- Whether smoked or chewed tobacco in last month and, if so, how much

Outcomes for parents

- Knowledge about tobacco
- Attitudes toward community prevention of tobacco use
- Attitudes toward tobacco use
- Intentions to talk to children about not using tobacco
- Reports of talks with their children about not using tobacco

SOURCE: Adapted from A. Biglan, D. Ary, H. Yudelson, T. E. Duncan, D. Hood, L. James, V. Koehn, Z. Wright, C. Black, D. Levings, S. Smith, and E. Gaiser, "Experimental Evaluation of a Modular Approach to Mobilizing Antitobacco Influences of Peers and Parents," *American Journal of Community Psychology*, 1996, 24(3):311-339.

Multiple measurement of important outcomes thus can provide for broader coverage of the concept and allow the strengths of one measure to compensate for the weaknesses of another. It may also be possible to statistically combine multiple measures

into a single, more robust and valid composite measure that is better than any of the individual measures taken alone. In a program to reduce family fertility, for instance, changes in desired family size, adoption of contraceptive practices, and average desired number of children might all be measured and used in combination to assess the program outcome. Even when measures must be limited to a smaller number than comprehensive coverage might require, it is useful for the evaluator to elaborate all the dimensions and variations in order to make a thoughtful selection from the feasible alternatives.

Measurement Procedures and Properties

Data on program outcomes have relatively few basic sources—observations, records, responses to interviews and questionnaires, standardized tests, physical measurement apparatus, and the like. The information from such sources becomes measurement when it is operationalized, that is, generated through a set of specified, systematic operations or procedures. The measurement of many outcome variables in evaluation uses procedures and instruments that are already established and accepted for those purposes in the respective program areas. This is especially true for the more distal and policy- relevant outcomes. In health care, for instance, morbidity and mortality rates and the incidence of disease or health problems are measured in relatively standardized ways that differ mainly according the nature of the health problem at issue. Academic performance is conventionally measured with standardized achievement tests and grade point average. Occupations and employment status ordinarily are assessed by means of measures developed by the Bureau of the Census.

For other outcomes, various ready-made measurement instruments or procedures may be available, but with little consensus about which are most appropriate for evaluation purposes. This is especially true for psychological outcomes such as depression, self-esteem, attitudes, cognitive abilities, and anxiety. In these situations, the task for the evaluator is generally to make an appropriate selection from the options available. Practical considerations, such as how the instrument is administered and how long it takes, must be weighed in this decision. The most important consideration, however, is how well a ready-made measure matches what the evaluator wants to measure. Having a careful description of the outcome to be measured, as illustrated in Exhibit 7-B, will be helpful in making this determination. It will also be helpful if the evaluator has differentiated the distinct dimensions of the outcome that are relevant, as illustrated in Exhibit 7-D.

When ready-made measurement instruments are used, it is especially important to ensure that they are suitable for adequately representing the outcome of interest. A measure is not necessarily appropriate just because the name of the instrument, or the

label given for the construct it measures, is similar to the label given the outcome of interest. Different measurement instruments for the "same" construct (e.g., self-esteem, environmental attitudes) often have rather different content and theoretical orientations that give them a character that may or may not match the program outcome of interest once that outcome is carefully described.

For many of the outcomes of interest to evaluators, there are neither established measures nor a range of ready-made measures from which to choose. In these cases, the evaluator must develop the measures. Unfortunately, there is rarely sufficient time and resources to do this properly. Some ad hoc measurement procedures, such as extracting specific relevant information from official records of known quality, are sufficiently straightforward to qualify as acceptable measurement practice without further demonstration. Other measurement procedures, however, such as questionnaires, attitude scales, knowledge tests, and systematic observational coding schemes, are not so straightforward. Constructing such measures so that they measure what they are supposed to in a consistent fashion is often not easy. Because of this, there are well-established measurement development procedures for doing so that involve a number of technical considerations and generally require a significant amount of pilot testing, analysis, revision, and validation before a newly developed measure can be used with confidence (see, e.g., DeVellis, 2003; Nunnally and Bernstein, 1994). When an evaluator must develop a measure without going through these steps and checks, the resulting measure may be reasonable on the surface but will not necessarily perform well for purposes of accurately assessing program outcomes.

When ad hoc measures must be developed for an evaluation without the opportunity for that development to be done in a systematic and technically proper manner, it is especially important that their basic measurement properties be checked before weight is put on them in an evaluation. Indeed, even in the case of ready-made measures and accepted procedures for assessing certain outcomes, it is wise to confirm that the respective measures perform well for the specific situation to which they will be applied. There are three measurement properties of particular concern: reliability, validity, and sensitivity.

Reliability

The **reliability** of a measure is the extent to which the measure produces the same results when used repeatedly to measure the same thing. Variation in those results constitutes measurement error. So, for example, a postal scale is reliable to the extent that it reports the same "score" (weight) for the same envelope on different occasions. No measuring instrument, classification scheme, or counting procedure is perfectly reliable, but different types of measures have reliability problems to varying degrees.

Measurements of physical characteristics for which standard measurement devices are available, such as height and weight, will generally be more consistent than measurements of psychological characteristics, such as intelligence measured with an IQ test. Performance measures, such as standardized IQ tests, in turn, have been found to be more reliable than measures relying on recall, such as reports of household expenditures for consumer goods. For evaluators, a major source of unreliability lies in the nature of measurement instruments that are based on participants' responses to written or oral questions posed by researchers. Differences in the testing or measuring situation, observer or interviewer differences in the administration of the measure, and even respondents' mood swings contribute to unreliability.

The effect of unreliability in measures is to dilute and obscure real differences. A truly effective intervention, the outcome of which is measured unreliably, will appear to be less effective than it actually is. The most straightforward way for the evaluator to check the reliability of a candidate outcome measure is to administer it at least twice under circumstances when the outcome being measured should not change between administrations of the measure. Technically, the conventional index of this *test-retest* reliability is a statistic known as the product moment correlation between the two sets of scores, which varies between .00 and 1.00. For many outcomes, however, this check is difficult to make because the outcome may change between measurement applications that are not closely spaced. For example, questionnaire items asking students how well they like school may be answered differently a month later, not because the measurement is unreliable but because intervening events have made the students feel differently about school. When the measure involves responses from people, on the other hand, closely spaced measures are contaminated because respondents remember their prior response rather than generating it anew. When the measurement cannot be repeated before the outcome can change, reliability is usually checked by examining the consistency among similar items in a multi-item measure administered at the same time (referred to as internal consistency reliability).

For many of the ready-made measures that evaluators use, reliability information will already be available from other research or from reports of the original development of the measure. Reliability can vary according to the sample of respondents and the circumstances of measurement, however, so it is not always safe to assume that a measure that has been shown to be reliable in other applications will be reliable when used in the evaluation.

There are no hard-and-fast rules about acceptable levels of reliability. The extent to which measurement error can obscure a meaningful program outcome depends in large part on the magnitude of that outcome. We will discuss this issue further in Chapter 10. As a rule of thumb, however, researchers generally prefer that their measures have reliability coefficients of .90 or above, a range that keeps measurement error

small relative to all but the smallest outcomes. For many outcome measures applied under the circumstances characteristic of program evaluation, however, this is a relatively high standard.

Validity

The issue of measurement validity is more difficult than the problem of reliability. The **validity** of a measure is the extent to which it measures what it is intended to measure. For example, juvenile arrest records provide a valid measure of delinquency only to the extent that they accurately reflect how much the juveniles have engaged in chargeable offenses. To the extent that they also reflect police arrest practices, they are not valid measures of the delinquent behavior of the juveniles subject to arrest.

Although the concept of validity and its importance are easy to comprehend, it is usually difficult to test whether a particular measure is valid for the characteristic of interest. With outcome measures used for evaluation, validity turns out to depend very much on whether a measure is accepted as valid by the appropriate stakeholders. Confirming that it represents the outcome intended by the program when that outcome is fully and carefully described (as discussed earlier) can provide some assurance of validity for the purposes of the evaluation. Using multiple measures of the outcome in combination can also provide some protection against the possibility that any one of those measures does not tap into the actual outcome of interest.

Empirical demonstrations of the validity of a measure depend on some comparison that shows that the measure yields the results that would be expected if it were, indeed, valid. For instance, when the measure is applied along with alternative measures of the same outcome, such as those used by other evaluators, the results should be roughly the same. Similarly, when the measure is applied to situations recognized to differ on the outcome at issue, the results should differ. Thus, a measure of environmental attitudes should sharply differentiate members of the local Sierra Club from members of an off-road dirt bike association. Validity is also demonstrated by showing that results on the measure relate to or "predict" other characteristics expected to be related to the outcome. For example, a measure of environmental attitudes should be related to how favorably respondents feel toward political candidates with different positions on environmental issues.

Sensitivity

The principal function of outcome measures is to detect changes or differences in outcomes that represent program effects. To accomplish this well, outcome measures should be sensitive to such effects. The **sensitivity** of a measure is the extent to which the values on the measure change when there is a change or difference in the thing

being measured. Suppose, for instance, that we are measuring body weight as an outcome for a weight-loss program. A finely calibrated scale of the sort used in physicians' offices might measure weight to within a few ounces and, correspondingly, be able to detect weight loss in that range. In contrast, the scales used to weigh trucks on interstate highways are also valid and reliable measures of weight, but they are not sensitive to differences smaller than a few hundred pounds. A scale that was not sensitive to meaningful fluctuations in the weight of the dieters in the weight-loss program would be a poor choice to measure that outcome.

There are two main ways in which the kinds of outcome measures frequently used in program evaluation can be insensitive to changes or differences of the magnitude the program might produce. First, the measure may include elements that relate to something other than what the program could reasonably be expected to change. These dilute the concentration of elements that are responsive and mute the overall response of the measure. Consider, for example, a math tutoring program for elementary school children that has concentrated on fractions and long division problems for most of the school year. The evaluator might choose an off-the-shelf math achievement test as a reasonable outcome measure. Such a test, however, will include items that cover a wider range of math problems than fractions and long division. Large gains the children have made in these latter areas might be obscured by the items on other topics that are averaged into the final score. A more sensitive measure, clearly, would be one that covered only the math topics that the program actually taught.

Second, outcome measures may be insensitive to the kinds of changes or differences induced by programs when they have been developed largely for diagnostic purposes, that is, to detect individual differences. The objective of measures of this sort is to spread the scores in a way that differentiates individuals who have more or less of the characteristic being measured. Most standardized psychological measures are of this sort, including, for example, personality measures, measures of clinical symptoms (depression, anxiety, etc.), measures of cognitive abilities, and attitude scales. These measures are generally good for determining who is high or low on the characteristic measured, which is their purpose, and thus are helpful for, say, assessing needs or problem severity. However, when applied to a group of individuals who differ widely on the measured characteristic before participating in a program, they may yield such a wide variation in scores after participation that any increment of improvement experienced by each individual will be lost amid the differences between individuals. From a measurement standpoint, the individual differences to which these measures respond so well constitute irrelevant noise for purposes of detecting change or group differences and tend to obscure those effects. Chapter 10 discusses some ways the evaluator can compensate for the insensitivity of measures of this sort.

The best way to determine whether a candidate outcome measure is sufficiently sensitive for use in an evaluation is to find research in which it was used successfully to detect

change or difference on the order of magnitude the evaluator expects from the program being evaluated. The clearest form of this evidence, of course, comes from evaluations of very similar programs in which significant change or differences were found using the outcome measure. Appraising this evidence must also take the sample size of the prior evaluation studies into consideration, because the size of the sample affects the ability to detect effects.

An analogous approach to investigating the sensitivity of an outcome measure is to apply it to groups of known difference, or situations of known change, and determine how responsive it is. Consider the example of the math tutoring program mentioned earlier. The evaluator may want to know whether the standardized math achievement tests administered by the school system every year will be sufficiently sensitive to use as an outcome measure. This may be a matter of some doubt, given that the tutoring focuses on only a few math topics, while the achievement test covers a wide range. To check sensitivity before using this test to evaluate the program, the evaluator might first administer the test to a classroom of children before and after they study fractions and long division. If the test proves sufficiently sensitive to detect changes over the period when only these topics are taught, it provides some assurance that it will be responsive to the effects of the math tutoring program when used in the evaluation.

Choice of Outcome Measures

As the discussion so far has implied, selecting the best measures for assessing outcomes is a critical measurement problem in evaluations (Rossi, 1997). We recommend that evaluators invest the necessary time and resources to develop and test appropriate outcome measures (Exhibit 7-F provides an instructive example). A poorly conceptualized outcome measure may not properly represent the goals and objectives of the program being evaluated, leading to questions about the validity of the measure. An unreliable or insufficiently sensitive outcome measure is likely to underestimate the effectiveness of a program and could lead to incorrect inferences about the program's impact. In short, a measure that is poorly chosen or poorly conceived can completely undermine the worth of an impact assessment by producing misleading estimates. Only if outcome measures are valid, reliable, and appropriately sensitive can impact estimates be regarded as credible.

Monitoring Program Outcomes

With procedures for adequate measurement of significant program outcomes formulated, various approaches to learning something about those outcomes can be undertaken by the evaluator or program managers. The simplest approach is outcome monitoring, which we defined in Chapter 6 as the continual measurement and reporting

EXHIBIT 7-F
Reliability and Validity of Self-Report Measures With Homeless Mentally Ill Persons

Evaluations of programs for homeless mentally ill people typically rely heavily on self-report measures. But how reliable and valid are such measures, particularly with persons who have psychiatric problems? One group of evaluators built a measurement study into their evaluation of case management services for homeless mentally ill clients. They focused on self-report measures of psychiatric symptoms, substance abuse, and service utilization.

Psychiatric symptoms. Self-report on the Brief Symptom Inventory (BSI) was the primary measure used in the evaluation to assess psychiatric symptoms. Internal consistency reliability was examined for five waves of data collection and showed generally high reliabilities (.76-.86) on the scales for anxiety, depression, hostility, and somatization but lower reliability for psychoticism (.65-.67). To obtain evidence for the validity of these scales, correlations were obtained between them and comparable scales from the Brief Psychiatric Rating Schedule (BPRS), rated for clients by master's-level psychologists and social workers. Across the five waves of data collection, these correlations showed modest agreement (.40-.60) for anxiety, depression, hostility, and somatization. However, there was little agreement regarding psychotic symptoms (–.01 to .22).

Substance abuse. The evaluation measure was clients' estimation of how much they needed treatment for alcohol and other substance abuse using scales from the Addiction Severity Index (ASI). For validation, interviewers rated the clients' need for alcohol and other substance abuse treatment on the same ASI scales. The correlations over the five waves of measurement showed moderate agreement, ranging from .44 to .66 for alcohol and .47 to .63 for drugs. Clients generally reported less need for service than the interviewers.

Program contact and service utilization. Clients reported how often they had contact with their assigned program and whether they had received any of 14 specific services. The validity of these reports was tested by comparing them with case managers' reports at two of the waves of measurement. Agreement varied substantially with content area. The highest correlations (.40-.70) were found for contact with the program, supportive services, and specific resource areas (legal, housing, financial, employment, health care, medication). Agreement was considerably lower for mental health, substance abuse, and life skills training services. The majority of the disagreements involved a case manager reporting service and the client reporting none.

(Continued)

EXHIBIT 7-F
Reliability and Validity of Self-Report Measures With Homeless Mentally Ill Persons (continued)

The evaluators concluded that the use of self-report measures with homeless mentally ill persons was justified but with caveats: Evaluators should not rely solely on self-report measures for assessing psychotic symptoms, nor for information concerning the utilization of mental health and substance abuse services, since clients provide significant underestimates in these areas.

SOURCE: Adapted from Robert J. Calsyn, Gary A. Morse, W. Dean Klinkenberg, and Michael L. Trusty, "Reliability and Validity of Self-Report Data of Homeless Mentally Ill Individuals," *Evaluation and Program Planning*, 1997, 20(1): 47-54.

of indicators of the status of the social conditions the program is accountable for improving. It is similar to program monitoring, as described in Chapter 6, with the difference that the information that is regularly collected and reviewed relates to program outcomes rather than to only program process and performance. Outcome monitoring for a job training program, for instance, might involve routinely telephoning participants six months after completion of the program to ask whether they are employed and, if so, what job they have and what wages they are paid. Detailed discussions of outcome monitoring can be found in Affholter (1994) and Hatry (1999).

Outcome monitoring requires that indicators be identified for important program outcomes that are practical to collect routinely and that are informative with regard to the effectiveness of the program. The latter requirement is particularly difficult. As discussed earlier in this chapter, simple measurement of outcomes provides information only about the status or level of the outcome, such as the number of children in poverty, the prevalence of drug abuse, the unemployment rate, or the reading skills of elementary school students. The difficulty is in identifying *change* in that status and, especially, linking that change specifically with the efforts of the program in order to assess the program's effects or impact.

The source of this difficulty, as mentioned earlier, 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 not an easy matter. Isolating program effects in a convincing manner from other influences that might have similar effects requires the special techniques of impact evaluation discussed in Chapters 8 and 9.

All that said, outcome monitoring provides useful and relatively inexpensive information about program effects, usually in a reasonable time frame. Whereas an

impact assessment may take years to complete, the results of outcome monitoring may be available within months. Furthermore, impact assessments typically require expenditures that are magnitudes greater than those needed for outcome monitoring systems. Because of its limitations, however, outcome monitoring is mainly a technique for generating feedback to help program managers better administer and improve their programs, not one for assessing the program's effects on the social conditions it is intended to benefit. As an illustration, consider the outcome monitoring of a treatment program for alcoholism. A result showing that 80% of the program's clients no longer drink several months after the program ends would present evidence more consistent with effectiveness than one showing only 20% abstaining. Of course, neither result is sufficient to establish real program effects, because the measured level of abstinence will also be affected by the severity of the clients' cases and by other influences on drinking that may override that of the program itself. A good monitoring scheme, however, will also include indicators of the severity of the initial problem, exposure to other important influences, and other relevant factors. While falling short of formal impact assessment, reasonable interpretation and comparison of patterns of such indicators and, especially, of trends in those indicators as programs attempt to improve their effectiveness, can provide useful indications of a program's effectiveness.

Indicators for Outcome Monitoring

Indicators that are to be used for outcome monitoring should be as responsive as possible to program effects. For instance, the 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 geographic areas served by the program, such as census tracts, zip codes, or municipalities, are not good choices for outcome monitoring if they include an appreciable number of persons not actually served by the program. 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 dropout rates are unimportant as a measure of program performance, but only that they should be assessed as a service utilization issue, not as an outcome issue.

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. Photographs of the streets that independent observers rate for cleanliness would be informative for assessing the effectiveness of this program. 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, though not one that, in itself, necessarily improves participants' lives. The more pertinent information comes from participants' reports of whether very specific benefits resulted from the service delivered by the program (see Exhibit 7-G). The limitation of such indicators is that program participants may not always be in a position to recognize or acknowledge program benefits, as in the case of drug addicts who are 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 in the case of elderly persons who are asked about the visiting nurses who come to their homes.

EXHIBIT 7-G
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 that typical programs attempt to affect, the limitations of outcome indicators, and the pressures on program agencies, there are many pitfalls associated with program outcome monitoring. Thus, while outcome indicators can be a valuable source of information for program decisionmakers, they must be developed and used carefully.

One important consideration is that any outcome indicator to which program funders or other influential decision makers give serious attention will also inevitably receive emphasis from program staff and managers. If the outcome indicators are not appropriate or fail to cover all the important outcomes, 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 skills needed to work with these children. As a result, the indicator continued to move upward, but the actual placement of children in appropriate foster homes did not improve. In education, this response 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 postprogram employment among participants is a major outcome indicator, for instance, consider the pressure on the program staff assigned the task of telephoning participants after completion of the program to ascertain their job status. Even with a reasonable effort at honesty, ambiguous cases will more likely than not be recorded as employment. It is usually best for such information to be collected by persons independent from the program. If it is collected internal to the program, it is especially important that careful procedures be used and that the results be 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 misleading and, even with proper context, they 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 explanation of the results found on those indicators. We consider the kinds of information that can be helpful in the following discussion of the interpretation of outcome data.

Interpreting Outcome Data

Outcome data collected as part of routine outcome monitoring can be especially difficult to interpret if they are not accompanied by information about changes in client mix, relevant demographic and economic trends, and the like. Job placement rates, for instance, are more accurately interpreted as a program performance indicator in the light of information about the seriousness of the unemployment problems of the program participants and the extent of job vacancies in the local economy. A low placement rate may be no reflection on program performance when the program is working with clients with few job skills and long unemployment histories who are confronting an economy with few job vacancies.

Similarly, outcome data usually are more interpretable when accompanied by information about program process and service utilization. The job placement rate for clients completing training may look favorable but, nonetheless, be a matter for concern if, at the same time, the rate of training completion is low. The favorable placement rate may have resulted because all the clients with serious problems dropped out, leaving only the "cream of the crop" for the program to place. It is especially important to incorporate process and utilization information in the interpretation of outcome indicators when comparing different units, sites, or programs. 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 dropout rates, or coping with other extenuating factors.

Equally important for interpretation of outcome monitoring data is development of a framework that provides some standard for judging what constitutes better or worse outcomes within the inherent limitations of the data for which these judgments must be made. One useful framework, when it is applicable, is a comparison of outcome status with the preprogram status on the outcome measure to reveal the amount of change that has taken place. 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 simple pre-post (before and after) comparison of this sort need not be part of routine outcome monitoring. It can also be done by the evaluator as part of an

outcome assessment. As we have noted, the main drawback to this design is that the differences between before and after measures cannot be confidently ascribed to program effects because other processes at work in the intervening period may affect the pre-post differences. One of the main reasons people choose to enter job training programs, for instance, is that they are unemployed and experiencing difficulties obtaining employment. Hence, they are at a low point at the time of entry into the program and, from there, some are likely to locate jobs irrespective of their participation in the program. Pre-post comparisons of employment for such a program will thus always show some upward trend that has little to do with program effects.

Other trends between the two times can also influence pre-post change. A program to reduce crime may appear more effective if it coincides with, say, efforts to increase policing. Confounding factors can also skew a pre-post comparison in the other direction: An employment training program will appear ineffective if it is accompanied by a prolonged period of rising unemployment and depressed economic conditions. In general, then, pre-post comparisons may provide useful feedback to program administrators as part of outcome monitoring, but they do not usually provide credible findings about a program's impact. The rare exception is when there are virtually no intervening events or trends that might plausibly account for a pre-post difference. Exhibit 7-H provides an example of such a situation.

The information that results from measuring program outcome variables, or change in those variables, generally must be interpreted on the basis of the judgments of program administrators, stakeholders, or experts in relation to their expectations for good and poor performance. These judgments are easiest at the extremes—when outcomes are more positive than likely to occur for reasons unrelated to the program, or so negative that little but program failure can explain them.

For instance, suppose that, after a two-month vocational program to train tractor-trailer truck drivers, more than 90% of the participants (selected from among persons without such skills) qualified for the appropriate driver's license. Such a finding suggests that the program has been quite successful in imparting vocational skills—it seems rather unlikely that so large a proportion of previously unskilled persons who wanted to become tractor-trailer truck drivers would be able to qualify for licenses on their own in a two-month period. By the same token, we could draw a relatively firm judgment that the program was ineffective if all the participants failed the license examination.

In reality, of course, the observed outcome would probably be more ambiguous—say, only 30% passing the first time. This more typical finding is difficult to judge and raises the question of whether a comparable group receiving no training would have done as well. Expert judgments might be called on in such circumstances. For instance, persons familiar with adult vocational education and the typical outcomes of intervention programs in that field might be asked to draw on their background to

EXHIBIT 7-H
A Convincing Pre-Post Outcome Design for a Program to Reduce Residential Lead Levels in Low-Income Housing

The toxic effects of lead are especially harmful to children and can impede their behavioral development, reduce their intelligence, cause hearing loss, and interfere with important biological functions. Poor children are at disproportionate risk for lead poisoning because the homes available to low-income tenants are generally older homes, which are more likely to be painted with lead paint and to be located near other sources of lead contamination. Interior lead paint deteriorates to produce microscopic quantities of lead that children may ingest through hand-to-mouth activity. Moreover, blown or tracked-in dust may be contaminated by deteriorating exterior lead paint or roadside soil containing a cumulation of lead from the leaded gasoline used prior to 1980.

To reduce lead dust levels in low-income urban housing, the Community Lead Education and Reduction Corps (CLEARCorps) was initiated in Baltimore as a joint public-private effort. CLEARCorps members clean, repair, and make homes lead safe, educate residents on lead-poisoning prevention techniques, and encourage the residents to maintain low levels of lead dust through specialized cleaning efforts. To determine the extent to which CLEARCorps was successful in reducing the lead dust levels in treated urban housing units, CLEARCorps members collected lead dust wipe samples immediately before, immediately after, and six months following their lead hazard control efforts. In each of 43 treated houses, four samples were collected from each of four locations—floors, window sills, window wells, and carpets—and sent to laboratories for analysis.

Statistically significant differences were found between pre and post lead dust levels for floors, window sills, and window wells. At the six-month follow-up, further significant declines were found for floors and window wells, with a marginally significant decrease for window sills.

Since no control group was used, it is possible that factors other than the CLEARCorps program contributed to the decline in lead dust levels found in the evaluation. Other than relevant, but modest, seasonal effects relating to the follow-up period and the small possibility that another intervention program treated these same households, for which no evidence was available, there are few plausible alternative explanations for the decline. The evaluators concluded, therefore, that the CLEARCorps program was effective in reducing residential lead levels.

SOURCE: Adapted from Jonathan P. Duckart, "An Evaluation of the Baltimore Community Lead Education and Reduction Corps (CLEARCorps) Program," *Evaluation Review*, 1998, 22(3):373-402.

judge whether a 30% outcome represents a success given the nature of the targets. Clearly, the usefulness and validity of such judgments, and hence the worth of an evaluation using them, depend heavily on the judges' expertise and knowledge of the program area.

Where possible, outcome values such as these might be compared with those from similar programs. This process is often referred to as "benchmarking" (Keehley et al., 1996), particularly when program performance on a particular outcome is compared with that of an especially effective program. As in all such comparisons, of course, the results are meaningful for evaluation purposes only when all other things are equal between the programs being compared, a difficult standard to meet in most instances.

Summary

- Programs are designed to affect some problem or need in positive ways. Evaluators assess the extent to which a program produces a particular improvement by measuring the outcome, the state of the target population or social condition that the program is expected to have changed.

- Because outcomes are affected by events and experiences that are independent of a program, changes in the levels of outcomes cannot be directly interpreted as program effects.

- Identifying outcomes relevant to a program requires information from stakeholders, review of program documents, and articulation of the impact theory embodied in the program's logic. Evaluators should also consider relevant prior research and consider possible unintended outcomes.

- To produce credible results, outcome measures need to be reliable, valid, and sufficiently sensitive to detect changes in outcome level of the order of magnitude that the program might be expected to produce. In addition, it is often advisable to use multiple measures or outcome variables to reflect multidimensional outcomes and to correct for possible weaknesses in one or more of the measures.

- Outcome monitoring can serve program managers and other stakeholders by providing timely and relatively inexpensive findings that can guide the fine-tuning and improvement of programs. Effective outcome monitoring requires a careful choice of indicators as well as careful interpretation of the resulting data.

- The interpretation of outcome measures and changes in such measures is difficult. Responsible interpretation requires consideration of a program's environment, events taking place during a program, and the natural changes undergone by targets

over time. Interpretation generally must rely on expert judgments of what constitutes good performance, though comparisons with other programs (benchmarking) can also be useful.

Key Concepts

Impact

See *program effect.*

Outcome

The state of the target population or the social conditions that a program is expected to have changed.

Outcome change

The difference between outcome levels at different points in time. See also *outcome level.*

Outcome level

The status of an outcome at some point in time. See also *outcome.*

Program effect

That portion of an outcome change that can be attributed uniquely to a program, that is, with the influence of other sources controlled or removed; also termed the program's impact. See also *outcome change.*

Reliability

The extent to which a measure produces the same results when used repeatedly to measure the same thing.

Sensitivity

The extent to which the values on a measure change when there is a change or difference in the thing being measured.

Validity

The extent to which a measure actually measures what it is intended to measure.