# EVALUATION

## A Systematic Approach

## SEVENTH EDITION

PETER H. ROSSI
*University of Massachussets, Amherst*

MARK W. LIPSEY
*Vanderbilt University, Nashville, TN*

HOWARD E. FREEMAN

SAGE Publications
*International Educational and Professional Publisher*
Thousand Oaks ▪ London ▪ New Delhi

# 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*

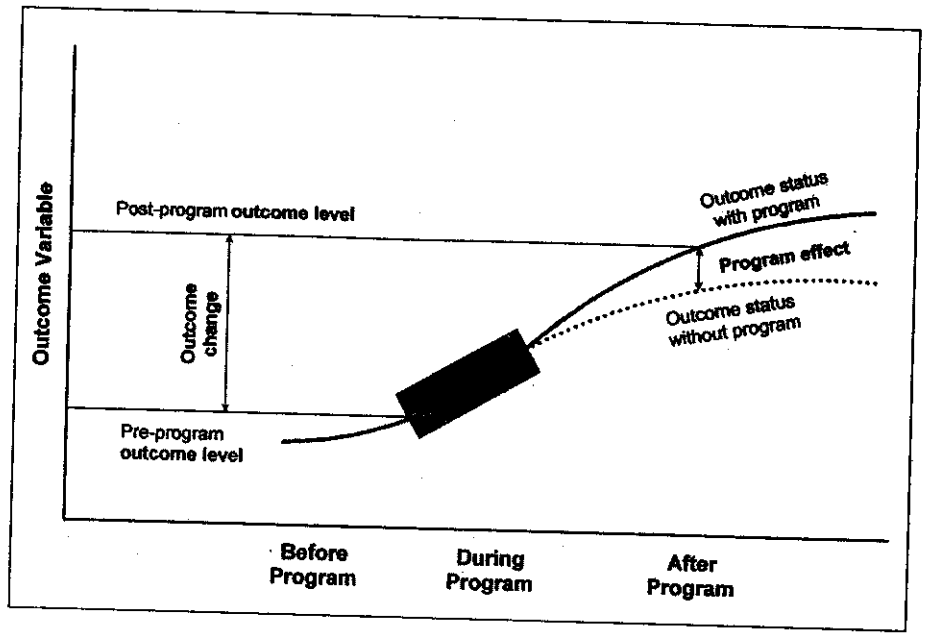
The foregoing considerations lead to important distinctions in the use of the term *outcome:*

- **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.

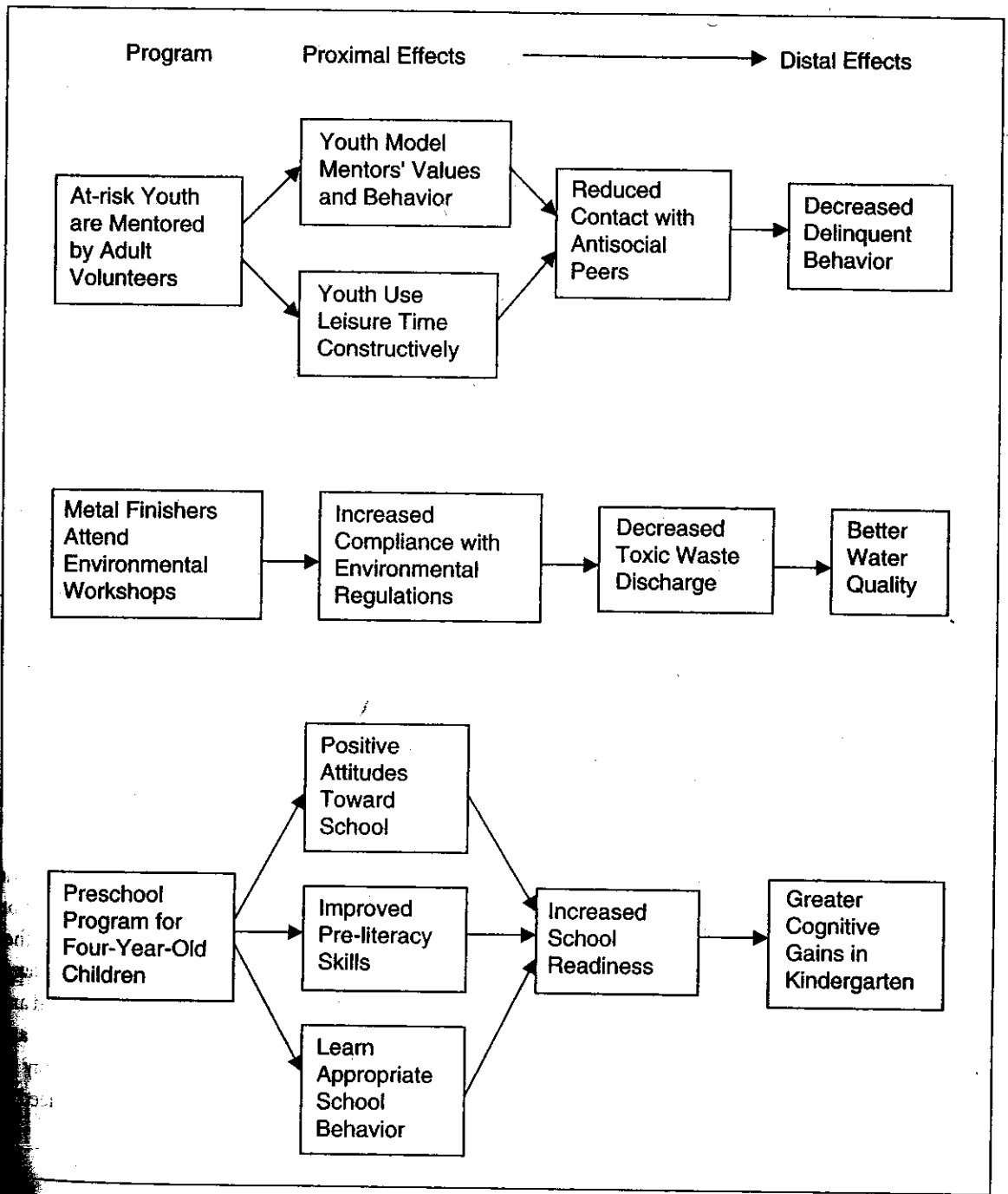
**School readiness**
Children's ability to learn at the time they enter school; specifically, the health and physical development, social and emotional development, language and communication skills, and cognitive skills and general knowledge that enable a child to benefit from participation in formal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

## easuring 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.