3

Defining Evaluation Purposes

Evaluation exists to improve the way that programs and policies function by providing information that can be used in democratic institutions to advance social betterment. If an evaluation is to aid in sensemaking about a program or policy, a series of decisions must be made about how the evaluation will be structured and carried out. These decisions will in turn affect the extent to which the evaluation provides useful information for improving, overseeing, selecting, or understanding public policies and programs. In the previous chapter, we discussed the rationale for using values inquiry as a guide for many key evaluation decisions, such as the selection of outcomes to measure. In this chapter we address another overarching concern, the selection of the general purpose of the evaluation. Social betterment is the ultimate purpose, but various types of evaluation have the potential to support this overall goal. Therefore evaluators should also consider more immediate reasons why people engage in sensemaking about policies and programs; we call these reasons evaluation *purposes*.

Four Purposes of Evaluation

We identify four primary purposes for which evaluation findings can be employed: assessment of merit and worth, program and organizational improvement, oversight and compliance, and knowledge

development. The *ultimate* purpose in carrying out most evaluations is to improve social conditions, but social betterment is a distal objective, mediated by democratic institutions and many interests and constituencies. More immediate, more proximal purposes need to be identified to drive the design and conduct of evaluation toward this ultimate goal. Identification of the immediate purpose can help evaluators make decisions about an evaluation's specific form.

An evaluation may be directed toward one of these purposes or (especially if it has a large budget and long time frame) multiple purposes. Indeed, of the skills that evaluators and evaluation program administrators can develop, among the most valuable is the ability to create an evaluation that can serve more than one purpose well (a process we discuss in some detail in Chapter Five). At the same time, evaluators need to clearly distinguish these four purposes. Most evaluation studies serve primarily one or two of them because attempting to do everything with limited resources can lead to doing nothing well, and because trade-offs among the purposes must often be considered if multipurpose evaluations are to be successful.

Although we do not emphasize it here, we recognize that motives other than those implied by the four purposes can also underlie an evaluation. For example, administrators may solicit an evaluation because they are required to do so by legislative mandate or by external funding agencies. Moreover, the motives underlying an evaluation are sometimes less than pure. Evaluations are sometimes commissioned to delay a decision, to duck responsibility for making a decision, or to improve public relations for a weak program (Weiss, 1998)—or even to try to torpedo a program (Suchman, 1967). Our focus is not on these uses of evaluation as a political tactic. Instead, we focus on four legitimate purposes to which the sensemaking of evaluation can contribute. Careful planning in light of these four purposes should maximize an evaluation's ability to contribute to sensemaking. This can hold true even when the evaluation originated as a political tactic.

Distinguishing the Purposes

Evaluators have long recognized that evaluation can serve different purposes. Scriven (1967) made the now classic distinction between formative and summative evaluations. *Formative* evaluations are those designed to facilitate program improvement, whereas *summative* evaluations are those intended to provide a definitive judgment of program or policy's merit and worth. Formative evaluation is exemplified by midterm course evaluations that provide feedback about things a teacher might do differently to try to improve a class. Summative evaluation is illustrated by *Consumer Reports*–type findings that offer a bottom-line judgment on the merits of some product.

Since Scriven introduced the distinction between formative and summative evaluation, evaluation scholars have identified a third possible purpose of evaluation, knowledge development (see especially Patton, 1997, on three uses of evaluation findings, and Chelimsky, 1997, on three perspectives on evaluation). When knowledge development is its primary purpose, an evaluation focuses on developing or testing (or both developing and testing) general propositions about such matters as the causes of social problems, the solutions to social problems, and the processes of policymaking, even though the knowledge may not directly improve or judge the specific program or policy being studied. Although some consensus may be developing among evaluation scholars in defining these three purposes of evaluation, there is enough divergence in previous definitions to lead us to suggest four purposes rather than three, adding oversight and compliance as a separate category.

The Four Purposes Illustrated

During the 1990s, so-called boot camps became a popular alternative to prison and other criminal sentences in many U.S. states, especially for juvenile offenders (Austin, Jones, & Boylard, 1993). Boot camps are modeled after military training centers. The underlying

rationale is that the hard work of the boot camp will decrease participants' subsequent motivation to commit crime and will also improve their personal skills, which will in turn lead to enhanced success in a life without crime. As a way of understanding the four evaluation purposes, consider the kinds of questions that interested parties might ask after the first states have implemented boot camps.

Some questions will deal with the *merit and worth* of boot camps. Legislators, who must decide whether to renew and expand the program, will want to know: What are the effects of the boot camps? In particular, do they seem to reduce recidivism more than traditional sentences? Considering their costs, are boot camps worth it? Answers to these questions are likely to be important also for legislators in states that are considering boot camps, for the public, and for judges who are sentencing eligible defendants. Most of these groups will also want to know: Are the effects larger for some groups than for others? Are there different types of boot camps, and if so, is one type better than the others? Defendants, their families, and their advocates (for example, the American Civil Liberties Union) may also want to know whether the rights of those assigned to boot camps are violated.

Other questions will relate to *program and organizational improvement*. These will probably be of primary concern to boot camp managers and staff as they consider ways to do their jobs better. But these questions will also, though perhaps less directly, be of interest to all others who want the program to be a success. What are the program's apparent strengths and weaknesses? Are there impediments that can be removed? Do some components, such as physical training, seem more important than others? Does implementation follow generally accepted practices or conform to best practices (if such standards exist)? Are any actions needed to increase participation (such as educating judges about this sentencing option), to improve staff skills (such as training staff on the developmental needs of adolescent offenders), to modify eligibility criteria for par-

ticipation (such as excluding violent offenders), or to internally reallocate program resources (such as increasing psychological services and eliminating costly Outward Bound sessions)? What additional information do program managers and staff need to improve operations?

The purpose of *oversight and compliance* will also generate questions. These will be of special interest to those with specific responsibility for ensuring that the program is operating according to its mandates, such as agency directors and legislators. The public and the press may also want to know about this. The general question for this purpose is: Does the program comply with statutes, rules, regulations, and other mandates for its operations? A variety of specific questions may follow. Do those referred to boot camps all meet the specified eligibility criteria? Have staff been screened and trained as they are supposed to be? Are safety and health requirements followed?

Still other questions fall under the rubric of *knowledge development*. Some of these will be related to boot camps only by convenience. For example, a criminal justice researcher may try to test labeling theory as it applies to recidivism. A methodologist might wonder if she can illustrate a new statistical procedure with data from the boot camp evaluation. Other knowledge development questions may be more directly related to boot camps: Can one develop a well-supported theory of boot camps as a treatment alternative? Does the boot camp experience contribute to a broader theory of the causes and treatment of criminal behavior? Can it lead to better ways to think about what works in criminal justice?

Many important questions will fall at the intersection of two or more evaluation purposes. In particular, the question of why the program works relates both to the assessment of merit and worth and to knowledge development and may also be important for program improvement. Before getting to more complicated questions such as this, it is important to better understand each of the four individual purposes.

The Four Purposes Defined

In this section we address the characteristics of the four purposes in greater detail.

Assessment of Merit and Worth. The evaluation purpose of assessment of merit and worth refers to the development of warranted judgments about the effects and other valued characteristics of a program or policy—and thus about a program or policy's value. By *effects* (or *outcomes*, a term we use synonymously), we mean the actual consequences of a program or policy, intended or unintended, positive or negative. For example, in an evaluation of a prekindergarten program, one might look for any effects on children's social skills, on their school readiness, and on parents' educational aspirations for their children, to mention but a few. Our definition also refers to *other valued characteristics*. The merit and worth of a program depends in part on whether it safeguards participants' (and others') rights and liberties. Programs that discriminate in hiring, violate privacy, or degrade participants are less meritorious and worthy than programs that do not. Although these other valued characteristics could conceivably be treated as effects, we prefer to highlight them separately because evaluators who focus primarily on effects often neglect to investigate practices that might indicate incursions into democratically established and protected rights.

We also distinguish between merit and worth. *Merit*, as we use the term, refers to the quality of a program or policy in terms of performance, and *worth* refers to the value this performance brings to the larger social good (Patton, 1997; Scriven, 1993). Consider an AIDS/HIV intervention. Its merit might consist of its success in reducing risky behaviors such as needle sharing or unprotected sex. The discovery of a vaccine for HIV would not necessarily alter the merit of the program—which might still function well at reducing the same behaviors—but it would presumably reduce the worth or value of the program to society.

As we mentioned, assessment of merit and worth corresponds to what Scriven (1967) called summative evaluation. Although it is now common to use the phrase *determination of merit and worth* in describing this evaluation purpose, we prefer the term assessment of merit and worth. The distinction is not trivial. *Determination* implies a role for evaluation that cannot be justified. It suggests coming to a fixed answer or settling a question. The very origins of the word suggest *coming to an end.* In evaluation, these connotations are unfortunate and undesirable. For one thing the limits of evaluation argue against using a term that implies finality. The history of evaluation practice teaches that the criteria for merit and worth can be slippery and subject to change—and that this can be a good thing (consider our example showing that the criteria used to judge preschool programs have expanded over time; Barnett, 1995). And even if the criteria for merit and worth were stable and well known, evaluation methods would remain fallible. In addition, evaluation information is not the end of the process of judging merit and worth. Although evaluation can provide assessment information that is useful to democratic institutions as they make sense about policies and programs, as they define merit and judge worth, this is very different from expecting evaluation to *determine* merit and worth (see also Stake, 1997).

Program and Organizational Improvement. When the evaluation purpose is program and organizational improvement, efforts are made to provide timely feedback designed to modify and enhance program operations. Formative evaluation (Scriven, 1967), as we mentioned, is the precursor term. When an evaluation is aimed at program improvement, it is likely to provide information about program effects and especially processes. Relative to an assessment of merit and worth, this evaluation is likely to have less concern with methodological rigor and validity and more concern with timeliness of information. Also, feedback is likely to be directed to program staff, the individuals who will make adjustments in program operations (see, for example, Wholey, 1983).

There are various models of program improvement. Following Scriven (1993), one approach to program improvement involves a comparatively casual and swift assessment of merit and worth, with the results presented to program staff who can then use this feedback to establish the need for program modifications. The metaphor Scriven suggests (1991, p. 169), which he credits to Bob Stake, is of the cook tasting the soup before the guests arrive and making timely adjustments. Another approach is analogous to the method of the auto mechanic, who does not judge the value of a car relative to alternative cars or other means of transportation but instead tinkers with the carburetor to try to enhance this car's performance. This type of improvement-oriented evaluation often focuses on identifying program elements that are not meeting expectations, examining alternatives, and choosing one of these alternatives. In a third approach to program improvement evaluation, program operations are compared to some supposed standard of best practices. In yet another approach, program staff are helped to a common understanding of the program and its desired outcomes through *evaluability assessment* (Rutman, 1980), the construction of program theory (Bickman, 1987), or a form of *developmental evaluation* (Patton, 1997). The evaluator's hope is that the new, shared view of the program will lead to better program services. In whatever form it takes, program improvement evaluation often provides information on current operations, outputs, and outcomes.

Chelimsky (1997) points out that some evaluators do not focus so much on improving a specific program as on improving organizational capacity to set policies, design and administer programs, and evaluate. We see this as similar to program improvement; the difference is that the objective of improvement is addressed from a broader, systems perspective. Thus program and organizational improvement evaluation includes much of what Patton (1997) refers to as developmental evaluation. It also addresses concerns about sustainability, an issue that arises with some frequency in so-called developing countries, where the question is whether service delivery organizations have the infrastruc-

ture to deliver services in the absence of continuing external support (Bamberger, 2000). (In further discussions of this purpose, we occasionally use *program improvement* as shorthand for program and organizational improvement.)

Oversight and Compliance. Evaluations with the purpose of evaluating oversight and compliance estimate the extent to which a program meets specified expectations such as the directives of statutes, regulations, or other mandates, including requirements to reach specified levels of performance. Traditionally, oversight and compliance evaluations focus on such issues as whether the program services being delivered are the services that have been authorized (McLaughlin, 1975), whether program clients meet established eligibility criteria (U.S. General Accounting Office, 1998b), or what percentage of the target population is being served (U.S. General Accounting Office, 1998a). Such evaluations can help meet program sponsors', funders', and the public's need to oversee the program and hold staff and administrators accountable. Recently, measurement of performance indicators has been more widely adopted as means of extending oversight from strictly procedural issues to outputs (such as the number of clients served) and outcomes (such as clients' performance) (Newcomer, 1997). For example, several U.S. states have set standards for public schools to meet, and hold the schools accountable for reaching them.

Oversight and compliance evaluations can indicate whether a program is meeting formal expectations and, if they have an outcome monitoring component, can also show what level participants are achieving on outcome measures. Still, such evaluations do not in and of themselves give a strong warranted assessment of merit and worth. They typically do not sort out the extent to which the program is responsible for the outcomes. Thus outcomes might be at a desirable level due to an improved external environment, such as a growing economy, rather than due to program effectiveness. Conversely, a program may operate in full accord with legislation and regulations and still not be effective.

Knowledge Development. Knowledge development refers to efforts to discover and test general theories and propositions about social processes and mechanisms as they occur in the context of social policies and programs. For some scholars, the world of social policies and programs is a valuable laboratory for developing and testing hypotheses and theories. The researcher interested in knowledge development may not be concerned with the specific program or policy per se but may be using it primarily as a venue that allows the investigation of some disciplinary research question. Alternatively, knowledge development can be a valuable adjunct to other evaluation purposes and can in some cases make major contributions to social betterment.

Knowledge development can focus on a wide variety of research questions involving large social science theories or on *small theories* of local programs (Lipsey, 1993), depending on the researcher's interests. For example, scholars of public administration might use evaluation work to develop general theoretical propositions about the implementation of social programs (Scheirer, 1987). Other scholars might develop a general classification system to describe the different types of services delivered in some area of human services, such as assisted living programs for the elderly (for example, Conrad & Buelow, 1990). Yet others might attempt to develop a theory of the treatment types that are effective for different types of clients in a program area, as Lipsey (1997) has in his meta-analysis of juvenile justice evaluations. Or an evaluation might allow a novel use of some research methodology, as in the use by Henry and Gordon (2000) of independent sample surveys in an interrupted time series design to evaluate the effectiveness of a public information campaign.

Evaluation Traditions

The four evaluation purposes have evolved in association with various general traditions in the field of program and policy evaluation. The assessment of merit and worth is associated with an

evaluation tradition that Shadish, Cook, and Leviton (1991) call the *manipulable solution theory of practice*. With prominent advocates such as Donald Campbell (1969b), this tradition has dominated much of the modern history of evaluation. It focuses on identifying programs and policies that work, typically by estimating the effect of the program or policy on outcomes considered to be important. One example of evaluation in this tradition comes from the early 1980s, when the city of Minneapolis was the site of an evaluation of a revision in police policy about domestic assaults, that is, fights between husbands and wives (or other domestic partners). Standard policy at the time was to end the fight and separate the spouses but not to arrest the assailant. A randomized experiment, a potentially powerful method for causal analysis, was carried out to see the effects of an alternative policy (Sherman & Berk, 1984). When a domestic assault meeting eligibility requirements occurred, the investigating police either followed the standard policy, or arrested the assailant, depending on a random schedule. Rates of reassault were estimated from police records and from interviews with victims. The results of the evaluation favored the new policy of arresting domestic assailants (however, this result has not replicated in follow-up evaluations at other sites) (Sherman, Smith, Schmidt, & Rogan, 1992). In terms of evaluation purposes, what is noteworthy is that this evaluation was carried out to try to judge the relative merit of competing policies.

The evaluation purpose of oversight and compliance is associated with a less visible but still important tradition, epitomized in much of the work carried out by the U.S. General Accounting Office (GAO) on behalf of Congress, which is interested in assessing the extent to which policies and programs follow the directives of statutes, regulations, or other mandates. One example is a study Congress asked GAO to conduct on the earned income credit (EIC), a tax credit available to employed persons who meet established low-income guidelines, with different eligibility cutoffs depending on family size. A stated objective of the EIC is to provide an incentive for work rather than welfare. Potential questions of

oversight and compliance involve the percentage of eligible taxpayers who apply for the EIC, the pattern of taxpayers' EIC use over time, and the changes in income and other outcomes that follow that use (U.S. General Accounting Office, 1997).

The purpose of program and organizational improvement is associated with several evaluation camps that fall within what Shadish, Cook, and Leviton (1991) call the *stakeholder service theory of practice*. One important version of this approach is captured in the work of Wholey and his colleagues (Wholey, 1979, 1983; Wholey, Scanlon, Duffy, Fukumoto, & Vogt, 1970). Wholey emphasizes the use of evaluation to guide program managers' efforts to administrate and revise programs. He focuses on program managers and staff as a stakeholder group that can deliver incremental improvements, and designs monitoring systems that can help them accomplish this. Another stakeholder service approach is reflected in the work of Patton (1997) (discussed in Chapter One). Patton's *utilization-focused evaluation* is well encapsulated in its motto, "intended use by intended users." Patton discusses how to identify the key stakeholders for an evaluation and determine their key information needs. Success in evaluation, from this perspective, lies in providing the information that satisfies those needs. Although in principle this could involve any of the four evaluation purposes, it appears that in practice stakeholders usually tell utilization-focused evaluators that they want program and organizational improvement. Similarly, the more recent trend of *empowerment evaluation* appears to consist mostly of efforts to teach program staff (and perhaps participants) to collect information for program and organizational improvement.

This purpose is exemplified by one component of an evaluation (Affholter, 1994) of the Intensive Crisis Counseling Program (ICCP) in Florida. In the ICCP, professional counselors under contract with the program offered intensive, short-term, home-based family counseling, with the expressed goal of reducing abuse, neglect, and placements in emergency shelters and foster care. During the evaluation, outcome data, such as the number of families remaining intact, were examined for each counselor. Staff then

investigated counselors who had poorer outcomes. They discovered, for example, that one contractor had stopped doing service delivery, beyond the initial assessment, in the families' homes. By leading to changes in the way services were delivered, the evaluation was used for the purpose of program improvement.

Finally, the knowledge development purpose is associated with what Shadish, Cook, and Leviton (1991) call the *generalizable explanation theory of practice*. This approach, advocated by Cronbach (1982), focuses on developing explanations about why programs and policies have their effects, under the assumption that such explanations are useful to program planners and staff. One general offshoot of this approach emphasizes program theory (for example, Bickman, 1990; Chen, 1990; also see Pawson & Tilley, 1997). The knowledge development purpose is also associated with the tradition present in a number of different academic disciplines of using program and policy evaluation as a vehicle for testing theoretically relevant hypotheses. The knowledge development tradition is illustrated by Lipsey and Wilson's summary (1993) of the results of 302 meta-analyses of the efficacy of psychological, educational, and behavioral treatments. This study was done largely to assess the accuracy of the common claim that when it comes to social interventions, "nothing works" (for example, Martinson, 1974). Each of the 302 meta-analyses had statistically combined a large number of individual evaluations in some treatment area, resulting in an *effect size* that would indicate in statistical terms how much better an intervention group did than a control or comparison group. Lipsey and Wilson found a striking tendency for positive average effects. Eighty-five percent of the meta-analyses had effect sizes of 0.20 or larger. Lipsey and Wilson's review leads more to the conclusion that "everything works" rather than that "nothing works." At least it appears to be the case that there is an average positive effect for most interventions that persist long enough to be the subject of a meta-analysis. Lipsey and Wilson also attempted to contribute to knowledge development in the area of evaluation methodology by examining whether different types of research designs were associated with smaller or larger

effects. In terms of purpose, the point is that although Lipsey and Wilson's study is based on evaluations, it was not designed to directly aid in overseeing, improving, or assessing the merit and worth of any particular policy or program. Instead, it was a general attempt at knowledge development.

Table 3.1 summarizes some key attributes of the four evaluation purposes, presenting the focus, a typical methodology, the usual audience, and associated evaluation traditions for each. The fact that for each purpose, different traditions have evolved and experienced long-term survival suggests that each purpose has value for assisting natural sensemaking, at least in certain conditions. From a commonsense realist perspective, which gives conceptual standing to lessons from practice (Putnam, 1990), this association between evaluation purposes and long-term practice traditions helps justify our distinction between the four purposes.

The fact that different traditions have evolved in association with the different purposes also suggests a possible problem. As Chelimsky (1997) argues, because they come from various traditions, evaluators differ sharply in their beliefs about the general purpose of evaluation. Some evaluators believe that generally the assessment of merit and worth is *the* purpose; others believe that generally program and organizational improvement is preferred; and so on. We contend, however, that the purpose(s) of an evaluation should be determined, not by an evaluator's predisposition, but by the evaluation's potential contribution to social betterment. This leads to the question of which of the four purposes best contributes to sensemaking support for democratic institutions and processes, and under what conditions.

Social Betterment and the Selection of Evaluation Purpose

As we have stated, the field of evaluation is premised on the belief that evaluations can contribute to social betterment, primarily by providing information that can support the deliberations, choices,

TABLE 3.1. Some Key Attributes of the Four Purposes of Evaluation.

	Assessing Merit and Worth	*Program and Organizational Improvement*	*Oversight and Compliance*	*Knowledge Development*
Focus	Support of judgment about value	Enhancement of program services	Compliance with formal expectations	Generation or testing of social science theory
Typical mode of inquiry	Causal analysis and values inquiry	Description, with timely observation and feedback	Description, including program activities and outcomes	Classification and causal analysis
Usual audience	Democratic institutions, the public	Administrators and program staff	Legislators, funders, the public	Social scientists, "conventional wisdom"
Evaluation tradition	Campbell's experimenting society (1969); the manipulable solution approach(see Shadish, Cook, & Leviton, 1991)	Wholey (1983) and feedback to managers; much of the stakeholder service approach (see Shadish, Cook, & Leviton, 1991)	GAO and state legislative oversight agencies	Cronbach (1982) and Chen (1990) program theory; the generalizable explanation approach (see Shadish, Cook, & Leviton, 1991)

and actions taken in democratic institutions. Recall also that evaluation's role is not limited to public deliberative processes such as debates in legislatures. Evaluation can be also influential in the administrative units that carry out policies and programs. Administration by public employees is an important part of the social betterment process, and in some ways it has the most proximate influence on social problems. With this in mind, in this section we consider the general conditions under which each of the four evaluation purposes might best aid sensemaking needs and capacities in a

democracy. In other words, we discuss the function that each evaluation purpose serves. (In Chapter Five, we discuss more nuanced, contextual factors that can influence the selection of an evaluation's purpose.)

Assessment of Merit and Worth

The fundamental rationale for assessing merit and worth is that doing so contributes to democratic deliberations about which major course of action to choose. Information about merit and worth helps when a decision must be made about adopting a new program or about selecting one of a set of alternative policies or about continuing an existing program or policy. How can democracies make informed choices about charter schools, welfare reform, and so on without information about merit? Of course decisions can be made without evaluation findings, and evaluation findings may not determine decisions. But the decisions are necessarily less well informed, less reasoned, without evidence of merit and worth. An assessment of merit and worth thus represents the highest contribution evaluation can make when democracies face a major fork in the road about which (if any) policy or program to adopt or drop. Given the centrality of democratic deliberations and choices to the social betterment process, and given the potential contribution of assessments of merit and worth to these deliberations, it is not surprising that many evaluators hold this purpose in high regard. Some even *define* evaluation solely in terms of assessing merit and worth (for example, Scriven, 1993).

In light of this key role for evaluations that assess merit and worth, why would one ever carry out an evaluation that emphasized another purpose? The primary reason is that the times when institutions confront major policy course changes are relatively rare. Democratic institutions have limits on their ability to deal with pressing public issues (Hilgartner & Bosk, 1988). Legislatures can consider only so many major policy initiatives at once. Administrators cannot drop their focus on service delivery to reconsider all

possible policy prescriptions. The media cannot keep the public informed if policies on transportation, education, the environment, health care, Social Security, employment, and training are all being considered for wholesale change at once. At best, a small number of major changes can be at the forefront of the public agenda at any time (McCombs & Zhu, 1995). But this does not mean that the remaining policy and program areas ought to go without evaluation. Instead, other evaluation purposes should rise to the fore, especially in the short run.

A second reason for giving priority to other evaluation purposes is that the timing may in other ways be bad for the assessment of merit and worth. Perhaps not enough is known yet about how to deal with a social problem or about how to get a program running smoothly. Or perhaps a good deal is known, but a number of compelling assessments of a program's merit and worth already exist. In either case the relative value of evaluations with other purposes would increase, at least for a while.

Program and Organizational Improvement

When program and organizational improvement is the primary evaluation purpose, the evaluation usually emphasizes timely feedback to modify and enhance program operations. This purpose of evaluation can serve social betterment by improving the options available for democratic consideration and selection. Under what conditions does program improvement best serve social betterment, and conversely, when can program improvement be a *dis*-service? On the one hand, program improvement efforts should be avoided when they will preclude the timely assessment of merit and worth needed for reasoned deliberative choices. For example, advocates of Project DARE, a drug abuse resistance education program, have tried to fight off evidence that the program is ineffective by contending that they have continued to improve it (Gorman, 1998). Of course, programs can and should be improved, and it is appropriate for democratic decisions to be revised

when evidence of merit and worth changes. But in the case of DARE, our sense is that claims of program improvement activity were used, perhaps disingenuously, in an attempt to fend off persuasive evidence that the program does not reduce substance abuse.

On the other hand, if program improvement work can be done *prior* to an assessment of merit and worth, social betterment is doubly served. Evaluation resources are often sufficient for this work because the front-end work for the assessment of merit and worth—planning, reviewing documents, constructing a program theory, testing instruments—can do double duty as the foundation for program improvement. Such sequencing of purposes, which has been built into some evaluation funding, can also help evaluators avoid the sin of premature assessment of merit and worth (see, for example, Sanders, 1999). Currently, unfortunately, evaluations conducted to stimulate program improvement are often not followed by assessments of merit and worth (Rog, 1985).

Program improvement also serves social betterment when, following an assessment of merit and worth, a program is adopted in another jurisdiction and steps are then taken to improve it. For example, a natural experiment (Rog, 1994) might be undertaken to investigate whether outcomes differ as a function of variations in services across sites of group delivery. If some service packages or types of service delivery are found to be more effective, they can be encouraged or mandated. Thus program effects may be enhanced at a time when it is unlikely that the program itself will be seriously scrutinized for termination or replacement.

Oversight and Compliance

When an evaluation focuses on the purpose of oversight and compliance, it assesses the extent to which a program meets formal expectations found in statutes, regulations, or other mandates. Oversight and compliance studies tend to focus on the fidelity with which a program is implemented. Fidelity here often includes both

adherence to specific program mandates and to externally established procedures, such as generally accepted accounting practices or requirements for equal opportunity in personnel decisions. Thus this purpose often motivates performance measurement and monitoring. In general, oversight is an important function especially for legislatures, courts, and higher administrative levels, which are supposed to ensure that policies are carried out properly and, increasingly, that outcomes are at expected levels. One-shot performance reviews and ongoing monitoring inform those with responsibility for oversight whether a policy is on track or needs some tweaking, without requiring the intense scrutiny of an assessment of merit and worth and without consideration of wholesale changes.

Under what conditions does evaluation best serve deliberative democracy by emphasizing oversight and compliance? First, evaluations with this purpose contribute well when some democratically selected program or policy has documented merit and worth and its effectiveness could suffer from poor compliance to established procedures. Consider programs to immunize poor children against childhood diseases. The effectiveness of vaccination against childhood diseases is well established. Program failure to comply with the rules and regulations could reduce program effectiveness, scope, and efficiency. Accordingly, oversight and compliance appears to be precisely the right evaluation purpose.

Second, there is social value in avoiding waste and fraud in social programs and in having the dictates of law and other democratically established mandates followed. Careful stewardship of financial resources is obviously important. Of course evaluations focusing on finances are often conducted by experts such as auditors who are in fields closely related to evaluation (Wisler, 1996). Other important mandates include, but are not limited to, protecting individual rights, providing equal opportunities to members of protected groups, and avoiding services that unnecessarily threaten clients' or others' dignity.

Finally, evidence from oversight and compliance evaluations may be valuable when combined with other evaluation purposes,

especially the assessment of merit and worth. A good assessment of merit and worth that produces negative findings may indicate poor implementation rather than a flawed program theory. Evidence that the program was implemented as directed may be collected as part of an assessment of merit and worth, but it may also come from a prior or concurrent oversight and compliance evaluation.

Knowledge Development

Knowledge development—the effort to construct and test theories and propositions about social processes and mechanisms as they occur in the context of social policies and programs—can aid in the long journey to social betterment. It can contribute to general deliberations about social problems and their proposed solutions—for example, by increasing people's understanding of the dynamics of the population in need and by explaining the underlying mechanisms by which programs and policies operate. Knowledge development is especially likely to move to the fore when very little is known about a social problem. And evaluators should be vigilant in looking for other opportunities to contribute to the knowledge base. Nevertheless, knowledge development should most often be a secondary rather than a primary evaluation purpose. First, in many cases, scholars in sociology, psychology, political science, education, health, and so on are in the best position to develop and test fundamental theories about human needs and social problems. Also, funds other than those budgeted for evaluation are often available for knowledge development.

In evaluation, there is special value to those forms of knowledge development that complement another evaluation purpose. In particular, the exploration of underlying mechanisms and testing for moderators of program effects is likely to add greatly to an assessment of merit and worth (Mark, Henry, & Julnes, 1998; Pawson & Tilley, 1997). Moreover, knowledge gained from experience with evaluation can lead to general propositions about obstacles to suc-

cessful program implementation (see, for example, St. Pierre & Kaltreider, 1997). At the same time, thoughtful judgment is needed to ensure that adding a knowledge development component does not seriously impede the primary evaluation purpose, which presumably has a more direct linkage to social betterment.

There is a potentially important instance in which knowledge development is likely to be the primary evaluation purpose. Some evaluations that look like assessments of merit and worth may in reality be better thought of as knowledge development. As noted previously, assessments of merit and worth serve democratic institutions that must make major program or policy choices, but sometimes evaluation is mandated when no such choice is forthcoming, perhaps for symbolic or political reasons. Some evaluators, especially those with academic research interests, may be inclined to do an evaluation that looks like an assessment of merit and worth even when no one is likely to be listening to the results in the immediate future. This choice is at best questionable when another purpose, such as program improvement, could be of benefit. But this approach can serve knowledge development, with the results enlightening discussions somewhere down the road or contributing to a subsequent meta-analysis designed to identify and explain the best program strategies for social betterment (Lipsey, 1997).

Exhibit 3.1 summarizes the generic ways in which social betterment drives the choice of evaluation purpose.

Relationships Among Purposes

So far this chapter has discussed the four purposes individually, but the reality of their application is far more complex. Metaphorically, evaluators might think of evaluation purposes as defining where they want to go in a dark room, thereby determining where they want to point a flashlight. If they happen to possess a powerful flashlight, which they point at a particular target, they are likely to illuminate some of the other objects in the room. Likewise, evaluations

EXHIBIT 3.1. How Social Betterment Drives Evaluation Purpose.

Social Betterment Consideration		*Evaluation Purpose*
• Democratic deliberations are needed about what course of action to take (for example, choosing a policy to meet a newly legitimized need or choosing between alternative programs).	→	Assessment of merit and worth
• A program of known worth could suffer from poor compliance. • A program might not be meeting established expectations and mandates or exercising sound financial stewardship. • It is necessary to rule out failure to implement as the explanation of failure to produce outcomes.	→	Oversight and compliance
• Efforts to enhance program operations will not inappropriately delay or invalidate assessment of merit and worth. • The program is generally judged of high worth and is not likely to undergo wholesale change.	→	Program and organizational improvement
• Mechanisms or moderators can be studied and will add to an assessment of merit and worth. • Learning that will contribute to the broader field can be added at low marginal cost to achieving other purposes.	→	Knowledge development

that emphasize one purpose may indirectly contribute to another. In addition, a single evaluation can combine two or more purposes, and a series of evaluations can be planned to address different purposes.

Contributions of One Evaluation Purpose to Another

An evaluation that directly emphasizes one purpose may also indirectly serve other purposes. For example, a credible assessment of merit and worth will often also indirectly stimulate program improvement. If program staff are simply informed about how well

or (especially) how poorly a program is doing, this often leads them to seek ways of doing better. Although probably less frequent in practice, an assessment of merit and worth can also have spin-off benefits for the two other purposes. It can indirectly contribute to knowledge development, as illustrated by Lipsey and Wilson's synthesis (1993) of meta-evaluations. And it can sometimes help satisfy the concerns that motivate an oversight and compliance evaluation. If it is clear that a program effectively reduces the social problem it was designed to address, the need to question its compliance with service delivery guidelines may be reduced.

Evaluations focused on oversight and compliance may also partially illuminate the other purposes. When a right way to carry out program activities is known, the findings of an oversight and compliance evaluation, showing that a program is being implemented incorrectly, may contribute to program improvement. In addition, when an oversight and compliance evaluation examines whether a program violates individual rights or involves fraud, it provides information relevant to merit and worth. Moreover, some evidence indicates that institutions are more likely to get the behaviors and outcomes that they monitor (Wood & Waterman, 1991, 1994). Evaluations that monitor behaviors that lead to reduced social problems can thus sometimes help produce the behavioral changes that will lead to better outcomes. Of course, "goal displacement" can also occur, whereby more important behaviors are reduced while more easily measured ones increase.

Multipurpose Evaluations

Beyond these incidental spin-offs from one purpose to another, a single evaluation can be explicitly designed to directly address more than one purpose. Consider efforts to identify and to test the underlying mechanisms responsible for a program or policy's effects. When successful, they answer questions about whether or not the policy works and why. Some writers on evaluation purposes (Chelimsky, 1997; Patton, 1997) have stated that the study of

mechanisms is a form of knowledge development. Evaluations that probe underlying mechanisms are likely to be relevant to social science theories about the causes of and treatments for social problems (Chen, 1990; Pawson & Tilley, 1997). But they also contribute to the assessment of merit and worth. Evaluation cannot adequately test underlying mechanisms without also assessing a program's effects on outcomes. (It is, however, possible to omit some valued outcomes because they have not been identified by prior theory, and this is a potential danger of being too theory driven.)

Series of Evaluation Studies

In some cases, evaluation purposes are combined in a series of evaluation studies. One common and often desirable sequence involves a series of evaluation studies that address, in turn, the purposes of program improvement, assessment of merit and worth, and oversight and compliance. Once ways of improving the program have been identified and implemented, the merit and worth of the presumably improved and well-implemented program are assessed, and finally, assuming the program is meritorious and worthy, program activities are monitored to verify that the program is still being carried out in the way found to be effective.

Conclusion

We have identified four possible purposes of evaluation: assessment of merit and worth, program and organizational improvement, oversight and compliance, and knowledge development. Although other pragmatic concerns, including some less than glorious, can sometimes motivate an evaluation, this fourfold set of purposes functions well as part of a vocabulary for planning evaluations in the service of social betterment. Different purposes and combinations of purposes are appropriate under different circumstances, and in all cases, considerations related to social betterment should drive purpose choice.

Assessments of merit and worth have a special role because they provide the information most useful for democratic deliberations intended to choose the best means to use in seeking desired ends. Oversight and compliance evaluation serves well when some democratically selected program or policy has documented merit and worth and its effectiveness could suffer from poor compliance to established procedures or when inadequate attention may be given to human rights and expenditure controls. Program and organizational improvement evaluation serves social betterment well—as long as it is not a ruse for avoiding consideration of merit and worth. It is often most useful when a social programs enjoys solid support and is not likely to be considered for wholesale change. Knowledge development is best seen as a purpose to be used in conjunction with other purposes. In particular, the study of underlying mechanisms can be a valuable adjunct to an assessment of merit and worth. Knowledge development can also motivate rigorous estimation of program effects during the interstices between the active attention to a program or policy within democratic institutions.

This framework of evaluation purposes contributes to a sensible way of thinking about evaluation. Together with the framework of inquiry presented in the next chapter, it provides a language evaluators from different traditions can use to communicate better and to understand each other's stance in the broader whole of evaluation. Even more important is its use in the planning of an evaluation. Although subsequent chapters present more nuanced and more complex guidelines for the choice of an evaluation purpose, the schematic summary in Exhibit 3.1 is an effective starting point for making judgments about the relative priority of the evaluation purposes in a given case. This framework is also applicable to the private sector. For example, a private company evaluating a training program might be concerned with assessing its merit and worth, with improving it, or with ensuring its compliance with regulations, and certainly there are researchers who have tested some pet theory in the context of evaluating a company's training program.

Selecting an evaluation purpose is of course not the end of evaluation planning. Once the evaluator has an evaluation purpose or combination of purposes in mind, he or she must turn to inquiry methods.