From Utilization-Focused Evaluation, 4th ed
by Michael Quinn Patton (2008, S)age

4

Intended Uses of Findings

If you don't know where you're going, you'll end up somewhere else.

—Yogi Berra

Evaluation Wonderland

When Alice encounters the Cheshire Cat in Wonderland, she asks, "Would you tell me, please, which way I ought to walk from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you walk," said the Cat.
"—so long as I get somewhere," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

—Lewis Carroll

This story carries a classic evaluation message: To evaluate how well you're doing, you must have some place you're trying to get to. For programs, this has meant having goals and evaluating goal attainment. For evaluators, this means clarifying the intended uses of a particular evaluation.

In utilization-focused evaluation, the primary criterion by which an evaluation is judged is *intended use by intended users*. The previous chapter discussed identifying primary intended users. This chapter will offer a menu of intended uses.

Identifying Intended Uses from the Beginning

The last chapter described a follow-up study of 20 federal health evaluations that assessed use and identified factors related to varying degrees of use. A major finding from that study was that *none of our interviewees had carefully considered intended use prior to getting the evaluation's findings*. We found that decision makers, program officers, *and* evaluators typically devoted little or no attention to intended uses prior to data collection. The goal of those evaluations was to produce findings; then they'd worry about how to use whatever was found. Findings would determine use, so until findings were generated, no real attention was paid to use.

Utilization-focused evaluators, in contrast, work with intended users to determine priority uses early in the evaluation process. The agreed-on, intended uses then become the basis for subsequent design decisions. This increases the likelihood that an evaluation will have the desired impact. Specifying intended uses is evaluation's equivalent of program goal setting.

Let me emphasize this point with an analogy. Once a year, I hike the Grand Canyon and have written a book about my experiences there (Patton 1999a). Once I was unloading my backpack at the Lodge on the North Rim when a young couple approached and said, "We want to hike down into the Canyon for a couple of days. Can you tell us what we have to do and where to get equipment?" I explained that the Grand Canyon is 217 miles long; you have to apply for overnight backcountry permits for specific areas and trails 4 months in advance; the terrain is steep and rugged, so it's wise, indeed essential, to have trained for a hike to the bottom, including carrying a heavy pack with enough water because it's a desert environment. And you have to bring your own equipment. It turned out that they had long dreamed of coming to the Grand Canyon, but had never thought about what they'd do once they got there. Now their options were quite limited—a short day hike, visiting some vistas, taking some photos. But it was too late to undertake a significant inner Canyon hike or join a rafting expedition on the Colorado River, which takes a minimum of a week to complete and a reservation months in advance. Their lack of advance planning is like deciding to do an evaluation but having given no real thought about what you really want to do with it until you get to the end. Then, you'll find, your options are quite limited. Using an evaluation to inform a specific decision, for example, requires advance planning and preparation so that the evaluation provides the needed information in time to be useful. You have to know the terrain of that decision, what the decision environment is like, and what the challenges are likely to be.

Baseline Data on Evaluation Use

In the 1970s, as the profession of evaluation was just emerging, those of us interested in use began by trying to sort out the influence of evaluations on decisions about programs. At the time, that seemed a reasonable place to begin. Much of the early literature on program evaluation defined use as immediate, concrete, and observable influence on specific decisions and program activities resulting *directly* from evaluation findings. For example, Carol Weiss (1972c), one of the pioneers in studying use, stated, "Evaluation research is meant for immediate and direct use in improving the quality of social programming" (p. 10). It was with reference to immediate and direct use that Weiss (1972c) was speaking when she concluded that "a review of evaluation experience suggests that evaluation results have generally not exerted significant influence on program decisions" (p. 11). Weiss (1990) reaffirmed this conclusion in her 1987 keynote address at the American Evaluation Association: "The influence of evaluation on program decisions has not noticeably increased" (p. 7). The evaluation literature reviewed in the first chapter was likewise overwhelming in concluding that evaluation studies exert little influence in decision making.

King and Pechman (1982, 1984) defined use as "intentional and serious consideration of evaluation information by an individual with the potential to act on it." This definition lowers the stakes for use—the evaluation has to be taken seriously—but doesn't necessarily have to lead to action. But even evidence of evaluations being taken seriously seemed hard to come by at the time.

It was in this gloomy context that I set out with a group of students in search of evaluations that had actually been used to help us identify factors that might enhance use in the future. (Details about this follow-up study of the use of federal health evaluations were presented in Chapter 3 and in Patton et al. 1977.) Given the pessimistic picture of most writings on use, we began our study fully expecting our major problem would be finding even one evaluation that had had a significant impact on program decisions. What we found was considerably more complex and less dismal than our original impressions had led us to expect. Our results provide guidance in how to work with intended users to set *realistic* expectations about how much influence an evaluation will have. After reviewing these baseline results on use, we'll look at developments in studying and conceptualizing utilization in recent years.

Views from the Field on Evaluation Impact

Our major question on use to project managers, program directors, and evaluators was this:

> *We'd like to focus on the actual impact of this evaluation study . . . , to get at any ways in which the study may have had an impact—an impact on program operations, on planning, on funding, on policy, on decisions, on thinking about the program, and so forth. From your point of view, what was the impact of this evaluation study on the program we've been discussing?*

After coding responses for the nature and degree of impact (Patton 1986:33), we found that 78 percent of responding decision makers and 90 percent of responding evaluators felt that *the evaluation had had an impact on the program.*

We asked a follow-up question about the nonprogram impacts of the evaluations:

> *We've been focusing mainly on the study's impact on the program itself. Sometimes studies have a broader impact on things beyond an immediate program, things like general thinking on issues that arise from a study, or position papers, or legislation. To what extent and in what ways did this evaluation have an impact on any of these kinds of things?*

We found that 80 percent of responding decision makers and 70 percent of responding evaluators felt these specific evaluation studies had had identifiable nonprogram impacts.

The positive responses to the questions on impact were quite striking considering the predominance of the impression of nonuse in the evaluation literature. The main difference here, however, was that *the actual participants in each specific evaluation process were asked to define impact in terms that were meaningful to them and their situations.* None of the evaluations we studied led directly and immediately to the making of a major, concrete program decision. The more typical impact was one in which the evaluation provided additional pieces of information in the difficult puzzle of program action, permitting some reduction in the uncertainty within which any decision maker inevitably operates. In most such cases, though the use was modest, those involved considered the evaluation worthwhile.

The most dramatic example of use reported in our sample was evaluation of a pilot program. The program administrator had been favorable to the program in principle, was uncertain what the evaluation results would be, but was "hoping the results would be positive." The evaluation proved to be negative. The administrator was "surprised, but not alarmingly so. . . . We had expected a more positive finding or we would not have engaged in the pilot studies" [DM367:13]. The program was subsequently ended, with the evaluation carrying "about a third of the weight of the total decision" [DM367:8]. Thus, the evaluation served the purpose of contributing to a final decision, but was one of only several factors (politics, impressions already held, competing priorities and commitments) that influenced the decision.

Contrast such use with the experiences of a different decision maker we interviewed, one who had 29 years' experience in the federal government, much of that time directing research. He reported the impact of the evaluation about which he was interviewed as follows:

> It served two purposes. One is that it resolved a lot of doubts and confusions and misunderstandings that the advisory committee had . . . and the second was that it gave me additional knowledge to support facts that I already knew, and, as I say, broadened the scope more than I realized. In other words, the perceptions of where the organization was going and what it was accomplishing were a lot worse than I had anticipated . . . but I was somewhat startled to find out that they were worse, yet it wasn't very hard because it partly confirmed things that I was observing. [DM232:17]

He went on to say that, following the evaluation,

> we changed our whole functional approach to looking at the identification of what we should be working on. But again I have a hard time because these things, *none of these things occurred overnight, and in an evolutionary process it's hard to say, you know, at what point it made a significant difference or did it merely verify and strengthen the resolve that you already had.* [DM232:17]

As in this example, respondents frequently had difficulty assessing the degree to which an evaluation actually affected decisions made after completion of the evaluation. This was true, for example, in the case of a large-scale evaluation conducted over several years' at considerable cost. The findings revealed some deficiencies in the program, but, overall, were quite positive. Changes corresponding to those recommended in the study occurred when the report was published, but those changes could not be directly and simply attributed to the evaluation:

> A lot of studies like this confirmed what close-by people knew and they were already taking actions before the findings. *So you can't link the finding to the action, that's just confirmation. . . . The direct link between the finding and the program decision is very diffuse.* [DM361:12, 13]

In essence, we found that evaluations provided some additional information that was judged and used in the context of other available information to help reduce the unknowns in the making of incremental program changes. The impact ranged from "it sort of confirmed our impressions . . . , confirming some other anecdotal information or impression that we had" [DM209:7, 1] to providing a new awareness that carried over to other programs:

This kind of use to stimulate thinking about what's going on and reduce uncertainty emerged as highly important to decision makers. In some cases, it simply made them more confident and determined. On the other hand, where a need for change was indicated, an evaluation study could help speed up the process of change or provide a new impetus for finally getting things rolling. Reducing uncertainty, speeding things up, and getting things finally started are real impacts—not revolutionary—but real, important impacts in the opinion of the people we interviewed. We found few major, direction-changing decisions in most programs. Rather, evaluation findings were used as one piece of information that fed into a slow, evolutionary process of program development. Program development is, typically, a process of "muddling through" (Allison 1971; Lindblom 1965), and program evaluation is part of that muddling. Or, as Weiss (1980) has observed, even major decisions typically accrete gradually over time through small steps and minor adjustments rather than getting decided all at once at some single moment at the end of a careful, deliberative, and rational process.

The impacts of evaluation have most often been felt as ripples, not waves. The question is whether such limited impact is sufficient to justify the costs of evaluation. The decision makers and evaluators we interviewed were largely satisfied with the type and degree of use they experienced. But times have changed. The stakes are higher. There's more sophistication about evaluation and higher expectations for accountability. However, the point of a utilization-focused approach is not to assume either high or low expectations. The point is to find out what the expectations of intended users are and negotiate a shared understanding of realistic, intended use—a mutual commitment that can be met. In negotiating the nature and degree of evaluation use, that is, setting goals for the evaluation, it is important to challenge intended users to be both optimistic and realistic—the twin tensions in any goal-setting exercise. Whether the expected type and degree of use hoped for actually occurs can then be followed up as a way of evaluating the evaluation. The question utilization-focused evaluation asks is, "What are the expected uses by intended users before and during the evaluation?" To work with intended users in clarifying intended uses, the evaluator needs to offer a menu of options and possibilities. The options have grown considerably based on considerable research on use and theoretical work in recent years. After looking at these developments, I'll offer a framework that distinguishes six primary purposes evaluations can serve. First, however, the results from research and theory.

Conceptualizing Use Options: Distinctions from Research

Inquiries into utilization show that intended uses vary from evaluation to evaluation, greatly affected y the context within which the evaluation occurs. There can be no generic or absolute ideal of evaluation use because "use" depends in part on the values and goals of primary users. As Eleanor Chelimsky (1983) observed, "The concept of usefulness . . . depends upon the perspective and values of the observer. This means that one person's usefulness may be another person's waste" (p. 155). To help intended users deliberate on and commit to intended uses, evaluators need a menu of potential uses to offer. Utilization-focused evaluation is a menu-oriented approach. *It's a process for matching intended uses and intended users.*

Let's begin this consideration of options by looking at classic distinctions. Early on, three types emerged as important: instrumental use, conceptual use, and symbolic use (Leviton and Hughes 1981)—and these remain the major distinctions informing discussions of use (Cousins and Shula 2006; Alkin 2005; Weiss, Murphy-Graham, and Birkeland 2005). *Instrumental use* refers to evaluation findings directly informing a decision or contributing to solving a problem; the findings are linked to some subsequent action and in that sense become an *instrument* of action. An example of instrumental use would be an evaluation of the Drug Abuse Resistance Education (D.A.R.E.) program in a school district that showed no effects on student drug use so the School Board decides to no longer fund the program (Weiss et al. 2005; Government Accountability Office [GAO] 2003). In the international arena, an evaluation finds that broken solar water pumps in African villages go without needed repairs because, after initial installation, no follow-up maintenance program was put in place. Based on the evaluation findings, the international agency that funded the installation decides to establish a maintenance program. That is instrumental use.

An Exemplar of Instrumental Use by the U.S. Congress

Laura Leviton, a former president of the American Evaluation Association and long-time contributor to research and theory on evaluation use, reviewed the state of our knowledge about evaluation use in the *American Journal of Evaluation*. She concluded that article by citing an outstanding example of instrumental evaluation use and the characteristics of the evaluator and evaluation that contributed to such a high degree of utilization. She wrote,

> For me the most consummate evaluation practitioner in terms of identifiable policy impact is still Paul Hill, who conducted a major evaluation mandated by the U.S. Congress on behalf of the National Institute of Education (NIE) in the late 1970s. As Boruch and I documented (Leviton and Boruch 1984), this work led to a great many specific changes in amendments to federal education law. In retrospect I believe Hill employed some of the [following] principles.
>
> - He was expert in the ways of Congress, having been on the Congressional staff.
> - Hill had the substantive education policy expertise as well.
> - The NIE study provided, not a stand-alone data collection effort, but a body of evidence . . . : the study was a collection of syntheses, pre-existing material, and some new, highly targeted primary data collection.

- The evaluation questions already had been sharply framed by years of Congressional debate on the relevant issues.
- Some debates had long ago turned into hardened positions. Hill sought findings in areas where there was still room for cross-party negotiation.
- Congressional stakeholders were heavily consulted in planning the study, during the course of the study, and in interpretation. Hill therefore understood the mental models of his stakeholders and was effective in translating findings into action, most notably when his team provided the legislative language needed for the amendments. (Leviton 2003:533)

Conceptual use occurs when an evaluation influences how key people think about a program or policy; they understand it better in some significant way, but no action or decision flows from the findings. We found conceptual use to be widespread in our follow-up study of federal health evaluations. As one project manager reported,

> The evaluation led us to redefine some target populations and rethink the ways we connected various services. This rethinking happened over a period of months as we got a better perspective on what the findings meant. But we didn't so much change what we were doing as we changed how we thought about what we were doing. That has had big pay-offs over time. We're just a lot clearer now. [DM248:19]

An international example of conceptual use is the Inter-American Development Bank evaluation of initiatives in six Latin American countries aimed at decentralization of government services to increase effective citizen participation. The evaluation revealed complex and diverse understandings of and experiences with decentralization. What seemed on the surface to be a straightforward administrative process of decentralizing government services turned out to be deeply intertwined with political, cultural, social, and economic conditions and factors. The findings conceptually distinguished "deconcentration" from decentralization, a situation in which "citizens are told that they have new decision-making power to help gain their support for a program" but the central government retains actual responsibility for the service and control of the financial resources. Deconcentration describes "cases where a certain obeisance is shown to decentralization and popular participation, but where the power structure retains control" (Inter-American Development Bank 2001:9–10). The report also reviewed privatization as a popular approach to decentralization and concluded,

> Privatization does not necessarily mean decentralization. It means, rather, that more actors are participating in the economic life of the country. Whether they are participating in the political life is more a matter of political parties, organizations for representation, and the enabling environment. (Inter-American Development Bank 2001:5)

Such findings provide important conceptual insights for future planning but are not directed at a particular decision for a specific program at a concrete point in time (instrumental use).

In one of the first studies comparing instrumental use with conceptual use, Shea (1991) did a follow up of 332 Canadian program evaluations and found that 55 percent reported instrumental use while 65 percent reported conceptual use. He also found an inverse relationship between the two: the greater the instrumental use, the less the conceptual use, and vice versa. In addition, he found that (1) evaluators who identified specific decision makers who would take responsibility for utilization reported significantly more instrumental use and (2) he found a significant relationship between the extent of instrumental use and the number of contact hours that the evaluator spend in working with program personnel during the planning, implementation, and dissemination stages of the evaluation.

Weiss (2004) has added a time dimension to conceptual use in what she has called "enlightenment" use and defines as

> the longer term percolation of ideas from evaluation into organizational discourse. . . . Evaluations not infrequently change decision makers' perceptions about what is important, they cast doubt on assumptions that had long been taken for granted, they evoke new ideas, and they alter priorities. (P. 161)

> Generalizations from evaluation can percolate into the stock of knowledge that participants draw on. Empirical research has confirmed this. . . . [D]ecision makers indicate a strong belief that they are influenced by the ideas and arguments that have their origins in research and evaluation. Case studies of evaluations and decisions tend to show that generalizations and ideas that come from research and evaluation help shape the development of policy. The phenomenon has come to be known as "enlightenment" . . . , an engaging idea. The image of evaluation as increasing the wattage of light in the policy arena brings joy to the hearts of evaluators (Weiss 1990:176–77).

Owen and Rogers (1999:110) link instrumental use with enlightenment in a model that conceives of enlightenment as sometimes an end in itself, but also as the first stage leading to more instrumental use. First, enlightenment and understanding, then application and decision making.

Symbolic use refers to token or rhetorical support for an evaluation process or findings but with no real intent to take either the process or findings seriously. Symbolic use has become more prevalent as research and evaluation findings have become increasingly prominent in political dialogue. In the knowledge age, politicians and decision makers have to at least appear to be basing their views on data. This distinction carries a warning to evaluators not to believe naively easily expressed rhetoric about interest in evaluation. Look for evidence of and specific actions in support of evaluation processes and findings; a reasonable evaluation budget and time devoted to the evaluation are prime types of such evidence.

Symbolic use constitutes a shrewd political use of evaluation to give the appearance of being an evidence-based decision maker. Other political uses distinguish specific intents. *Persuasive use* refers to using evaluation findings, often quite selectively, to support one's position in political debates. So, for example, a police chief testifying before a School Board in support of funding for D.A.R.E. would emphasize findings that students feel more trusting of police after classes about the dangers of drug use taught by police and ignore the findings that the program has no effect on students' subsequent drug use (GAO 2003). Weiss et al. (2005) caution against judging such persuasive use as necessarily inappropriate. "When evaluation supports a course of action that already has advocates, there does not seem to be anything wrong with using evaluation evidence to strengthen the case" (pp. 13–14).

Another type of politically oriented use is "legitimative utilization" (Alkin 2005:435; Leviton 2003:533; Owen and

Rogers 1999) in which evaluation findings are used to support a decision that was actually made before the evaluation was ever conducted or was made without regard to evaluative evidence. This is what the critics of the Iraq War argue happened, namely, that President Bush and his neoconservative advisors had already decided immediately after the 9/11 terrorist attack on the World Trade Center that they would use the attack as justification for invading Iraq and deposing Saddam Hussein. They then set about gathering and presenting selective "evidence" to legitimate that predetermined decision (U.S. Senate Select Committee on Intelligence 2004; Hersh 2003). This happens in a program context when a decision is made to terminate a program and then an evaluation is commissioned for the purpose of legitimating the decision after the fact. Program staff is often fearful of just such an agenda when internal evaluations are commissioned in a time when resources are known to be constrained and some cuts somewhere will have to be made. To the extent that legitimative use is intentionally manipulative and deceptive, it becomes misuse.

Misuse of Evaluations

Studies of evaluation use have generated examples of and raised concerns about misuse. Evaluation processes and findings can be misrepresented and abused. The profession recognizes a critical distinction between *misevaluation*, in which an evaluator performs poorly or fails to adhere to standards and principles, and *misuse*, in which users manipulate the evaluation in ways that distort the findings or corrupt the inquiry.

Sources of misuse include hard-core politics, asking the wrong questions, pressures on internal evaluators to present only positive findings, petty self-interest, and ideology (Stevens and Dial 1994; Dial 1994; Duffy 1994; Mowbray 1994; Posavac 1994; Vroom, Columbo, and Nahan 1994; Alkin and Coyle 1988). Misuse, like use, is ultimately situational. Consider, for example, the case of an administrator who blatantly squashes several negative evaluation reports to prevent the results from reaching the general public. On the surface, such an action appears to be a prime case of misuse. Now consider the same action (i.e., suppressing negative findings) in a situation where the reports were invalid due to poor data collection. Thus, misuse in one situation may be conceived of as appropriate nonuse in another. Intentional nonuse of poorly conducted studies can be viewed as appropriate and responsible. Here are some premises with regard to misuse.

1. Misuse is *not* at the opposite end of a continuum from use. Two dimensions are needed to capture the complexities of real-world practice. One dimension is a continuum from appropriate nonuse to appropriate use. A second is a continuum from inappropriate nonuse to intentional misuse. Studying or avoiding misuse is quite different from studying or facilitating use.

2. Having conceptualized two separate dimensions, it is possible to explore the relationship between them. Consider the following proposition: *As use increases, misuse will also increase.* When people ignore evaluations, they ignore their potential uses as well as abuses. As evaluators successfully focus greater attention on evaluation data and increase actual use, there may be a corresponding increase in abuse, often within the same evaluation experience. Donald T. Campbell (1988:306) formulated a discouraging law along these lines that the more any social indicator is used for important societal decision

making, the more likely is that indicator to be corrupted.

3. Misuse can be either intentional or unintentional. Unintentional misuse can be corrected through the processes aimed at increasing appropriate and proper use. Intentional misuse is an entirely different matter that invites active intervention to correct whatever has been abused, either the evaluation process or findings. As with most problems, correcting misuse is more expensive and time-consuming than preventing it in the first place.

4. Working with multiple users who understand and value an evaluation is one of the best preventatives against misuse. Allies in use are allies against misuse. Indeed, misuse can be mitigated by working to have intended users take so much ownership of the evaluation that they become the champions of appropriate use, the guardians against misuse, and the defenders of the evaluation's credibility when misuse occurs.

5. Policing misuse is sometimes beyond the evaluator's control, but to the extent possible and realistic, professional evaluators have a responsibility to monitor, expose, and prevent misuse (Patton 2005a).

Evaluators' Perceptions of Nonuse and Misuse

Rated by evaluators as "a great problem"

Nonuse of evaluation results	68 percent
Intentional misuse of evaluation results	21 percent
Unintentional misuse of evaluation results	22 percent

SOURCE: Results of a 2006 online survey of members of the American Evaluation Association with 1,014 respondents (Fleischer 2007).

Appropriate versus Inappropriate Nonuse

The utility standards of the profession make it clear that a good evaluation is one that is used. Some use is good; more use is better. Appropriate and intended use by intended users is best. Misuse is bad. And nonuse? From a utilization-focused evaluation perspective, nonuse represents some kind of failure in the evaluation process. We often lay that failure at the feet of resistant or unappreciative stakeholders, but it can also be the evaluator's fault. *Nonuse due to misevaluation* (Patton 2005b:254), or justified nonuse (Cousins and Shula 2006:282) refers to appropriate nonuse because of weak evidence, a late report, poor evaluator performance, or other failures of the evaluator to adhere to the profession's standards and principles (see Chapter 1). In contrast, *political nonuse* occurs when the findings are ignored because they conflict with a potential user's values, prejudices, preferences, and predisposition—so the evaluation is just simply ignored. Utilization-focused evaluation attempts to reduce political nonuse by creating a climate and process in which those involved are willing and prepared to examine their basic assumptions and incorporate evidence into their understandings, even when they had hoped for, or would have preferred, different results.

Aggressive nonuse, or calculated resistance, refers to situations where an evaluation or evaluator is attacked and use undermined because the results conflict with or raise questions about a preferred position. Resistance to evaluation findings can be a specific example of the more general phenomenon of resistance to change. A major reason for identifying and involving primary intended users in the evaluation is to anticipate and short-circuit inappropriate and specious attacks, or at least to have allies among informed and credible intended users in fending off such politically motivated attacks.

Most resistance to evaluations is behind the scenes, but occasionally political reports grab media attention and the whole world gets to watch the circus of attacks and counterattacks. A prominent example was the May, 2005 release of a report by the human rights organization Amnesty International on conditions in the U.S. military prison at Guantanamo Bay in Cuba where alleged terrorists were being held. The report, citing interviews with prisoners and people who had been inside the prison, concluded that prisoners had been mistreated and called for the prison to be shut down. The report got considerable international media attention. Amnesty International has an explicit agenda and its recommendation to close the Guantanamo facility could be expected, but the cases cited and interview results were viewed as credible by some reporters, so the Bush Administration needed to make a response. The tone of the response gives a flavor of the rhetoric that can accompany an aggressive attack on disputed and unwelcome evaluation conclusions. President Bush, addressing a news conference at the White House on May 31, 2005, said the Amnesty document was an "absurd report. It's absurd. It's an absurd allegation. The United States is a country that promotes freedom around the world." He went on to attack the investigation's methods and resulting data asserting that the Amnesty allegations were based on interviews with detainees who hated America and were trained to lie. President Bush's remarks were echoed by Vice President Dick Cheney, who said that same day in a videotaped interview with CNN's Larry King, "Frankly, I was offended by it. For Amnesty International to suggest that somehow the United States is a violator of human rights, I frankly just don't take them seriously."

In the early 1970s, I was involved in an independent survey of teachers in Kalamazoo, Michigan with funds from the local and national education associations. The School District refused to cooperate with the study and when the results came in showing very low morale, widespread complaints about working conditions, a dysfunctional accountability system, and allegations of administrative abuses, the Superintendent publicly attacked the findings, calling them "absurd." He attacked my integrity, saying I was an out-of-state paid-gun-for-hire, and further asserted that the teachers association instructed teachers how to respond. He dismissed the results out of hand. Fortunately, the school board members actually read the report, including pages of in-depth quotations from teachers and documented cases of problems. The school board made instrumental use of the report by requiring major administrative changes in the District and, subsequently, the superintendent "resigned." (For details, see Patton 2002a:17–20.)

The point: Evaluation is a political activity and as the varieties of use, nonuse, and misuse illustrate, utilization is also a political activity—and sometimes the politics gets rough. This work is not for the feint of heart; it's not just an academic exercise. The stakes can get very high, very fast. Some more recent use distinctions further reinforce this caution.

More Recent Use Distinctions

The classic three types of use—instrumental, conceptual, and symbolic—have long framed inquiries into evaluation use and led to concerns about misuse. Over time, as the field has matured and inquiries into utilization have broadened and deepened, additional distinctions have emerged from research and theory. Weiss et al. (2005), based on case studies of the use of D.A.R.E. evaluations, have identified *imposed use* that occurs when those with the power to do so mandate an action based on evaluative judgments; in essence, those at a higher level of authority require a prescribed use by those at a lower level. For example, a federal requirement that to receive funding a school district curriculum must be on an approved list of "evidence-based" or evaluated programs. In the case of D.A.R.E., administrators in some districts felt forced to drop the program, despite local support, because it did not qualify as a preapproved, evidence-based program by the federal authorities.

I have become concerned about o*veruse*, which occurs when too much emphasis is placed on evaluation findings. For example, weak evaluation results are overused when treated as if they are definitive, or imposed use becomes overuse when there is insufficient evidence to generalize findings and justify the top-down mandate for compliance, or there is lack of attention to local conditions. This latter overuse can occur when supposed "best practices" are universally mandated (Patton 2001). Concern about overuse is ironic since, as the first chapter documented, the profession has been dominated by concern about underuse and nonuse. But as in much of life, you can have too much of a good thing. An unintended consequence of all the focus on increasing use may have contributed to overuse and misuse.

Mechanical use (Patton 2006) is another emergent distinction of increasing concern. Mechanical, or compliance use, refers to going through the motions to meet an evaluation requirement. The evaluation is required, so it is done, but the motivation is compliance and the implementation is mechanical. A number of colleagues who do evaluations in the federal government have encountered this approach, as have I, especially with regard to mandated Program Assessment Rating Tool (PART) reviews, a process mandated by the U.S. Office of Management and Budget (OMB) for all federal programs. PART was developed to help budget examiners and federal managers measure the effectiveness of government programs. It is a 25-item questionnaire divided into four sections: program purpose and design (5 questions); strategic planning (8 questions); program management (7 questions); and program results/ accountability (5 questions). Based on answers to these questions, a score is generated and a program is rated as Effective, Moderately Effective, Adequate, or Results Not Generated. The stakes are high. Results are made public and can affect program budgets and status. So how does mechanical use come into play? A director of a program preparing for a PART says to the evaluator, "Just tell me what I have to do to increase my PART score." Such a director isn't looking to improve the program or make a decision. The object is just to get a decent, acceptable score. The same phenomenon happens in not-for-profit programs when they go mechanically through the motions of complying with a funder's mandated evaluation.

Process Use

Now we turn to a quite different type of use. *Process use* has emerged as one of the most important distinctions in the last decade (Cousins and Shula 2006; Alkin 2005). Process use refers to cognitive, behavioral, program, and organizational changes resulting, either directly or indirectly, from engagement in the evaluation process and learning to think evaluatively (e.g., goals clarification, conceptualizing the program's logic model, identifying evaluation priorities, struggling with measurement issues, participation in design and interpretation). Process use occurs when those involved in the evaluation learn from the evaluation process itself or make program changes based on the evaluation process rather than findings—as, for example, when those involved in the evaluation later say "the impact on our program came not just from the findings but also from going through the thinking process that the evaluation required." Process use also includes the effects of evaluation procedures and operations, for example, the premise that "what gets measured gets done," so establishing measurements and setting targets affects program operations and management focus. These are uses of the evaluation process to affect programs, not use of findings. Process use has become so important that the entire next chapter is devoted to it. I mention it here to be sure it is on the menu when considering use options.

Utilization versus Use versus Influence: The Terminology Debate

> *Words are loaded pistols.*
>
> Jean Paul Sartre,
> philosopher (1905–1980)

The evaluation language we choose and use, consciously or unconsciously, necessarily and inherently shapes perceptions, defines "reality," and affects mutual understanding. Whatever issues in evaluation we seek to understand—types of evaluation, methods, relationships with stakeholders, power, use—a full analysis will lead us to consider the words and concepts that undergird our understandings and actions because language matters (Patton 2000). Deciding on terminology is complicated because two people can infer different connotations from the same word. Early on Carol Weiss (1980, 1981) expressed a preference for *use* rather than *utilization.* She went so far as to propose abandoning the term *utilization* "because of its overtones of instrumental episodic application. People do not utilize research the way that they utilize a hammer." She preferred use instead of utilization to capture the sense that findings "penetrate" decision making through "processes of understanding, accepting, reorienting, adapting, and applying research results to the world of practice." She wanted a more "fluid and diffuse" connotation (Weiss 1981:18). Yet I have quite the opposite reaction to the two terms. Use seems to me more instrumental and episodic in connotation. Taking her analogy, I would argue that people use hammers; they don't "utilize" hammers. But they do utilize evaluations, which connotes to me *a process* of precisely the kind Weiss describes—understanding, accepting, reorienting, adapting, and applying. Use sounds to me more direct, specific, concrete, and moment-in-time. Utilization evokes for me a dynamic process that occurs over time. So I continue to prefer utilization-focused evaluation over use-focused evaluation.

Others have expressed a preference for use instead of utilization simply because the longer word sounds more academic, like jargon, and is too highfalutin (pompous or pretentious). For that reason, I much prefer the verb use instead of utilize, but I make use of both nouns—use and utilization—varying my usage by audience and context.

Karen Kirkhart (2000) wants to abandon both the terms *use* and *utilization* in order to construct an "integrated theory" of evaluation's consequences using the concept of "evaluation influence" as a unifying construct. She defines influence as “the capacity or power of persons or things to produce effects on others by intangible or indirect means.” Kirkhart posits three dimensions of evaluation influence: source of influence (evaluation process or results), intention (intended or unintended), and time (immediate, end-of-cycle, long-term). She is especially anxious to capture effects that are “multidirectional, incremental, unintentional, and instrumental” (p. 7).

Unintended uses are any applications of evaluation findings or processes that were not planned, not predictable, or unforeseen. Kirkhart (2000) cites as an example a program advisory committee that intends to use evaluation results to improve the program, but “the data had unexpected policy implications that led them to initiate a community coalition to advocate for legislative change” (p. 13). I evaluated a leadership program for a philanthropic foundation and the foundation liked the approach so much they supported me to train others in development-oriented utilization-focused evaluation and made it a centerpiece of their evaluation philosophy. Such influence was beyond the scope of anything imagined at the beginning of the process.

Kirkhart’s influence framework has influenced, quite rightly, how research on evaluation’s effects are conceptualized and studied (e.g., Christie 2007; Mark and Henry 2004; Henry 2003; Henry and Mark 2003), especially in calling attention to the importance of looking for unintended effects; examining long-term, incremental, and unanticipated uses of findings; and investigating diverse forms of influence. But the framework is less useful, in my judgment, for informing practice. Alkin (2005) has cogently explained why this is the case.

> Evaluation use typically refers to the impact of the evaluation (findings or process) within the context of the program being evaluated, within some reasonable time frame. Evaluation influence refers to the impact on an external program, which may or may not be related to the program evaluated, or to the impact of the evaluation at some future time. An important distinction between evaluation influence and evaluation use is that evaluators who are concerned with evaluation use can actively pursue a course of action to potentially enhance utilization by recognizing the evaluation factors and attempting to be responsive to them, but evaluation influence is more difficult to predict or to control. (P. 436)

Utilization-focused evaluation is focused on *intended use by intended users.* The emphasis is on intentionality and harnessing that intentionality to enhance utilization. In contrast, evaluation influence emphasizes the indirect aspects of evaluation’s effects over time and outside the program evaluated, things that are largely beyond the evaluator’s control. Utilization-focused evaluators, however, can conduct evaluations in ways that increase use, especially by being intentional about the evaluation’s primary purpose, which is the focus of the next section. Exhibit 4.1 reviews and summarizes the use distinctions discussed above. We turn now to a menu of six distinct evaluation purposes based on varying uses for evaluation *findings*. In the next chapter, we’ll add to this menu a variety of uses of evaluation *processes*.

Six Alternative Evaluation Purposes

> *The purpose of an evaluation conditions the use that can be expected of it.*
>
> Eleanor Chelimsky (1997)

You don’t get very far in studying evaluation before realizing that the field is characterized by enormous diversity. From

EXHIBIT 4.1

Use Distinctions

Direct Intended Uses

Instrumental use occurs when evaluation findings are used to directly inform a decision, improve a program or policy, develop new directions, or contribute to solving a problem; the findings are linked to some subsequent, identifiable action. (Menu 4.2 in this chapter elaborates types of instrumental use.)

Conceptual use occurs when an evaluation influences how key people think about a program or policy, and understand it better in some significant way, but no action or decision flows from the findings. This use is often anticipated and intended by including in the scope of work the expectation of generating "lessons learned" or, more generally, contributing to knowledge.

Process use refers to changes resulting from engagement in the evaluation process and learning to think evaluatively. Process use occurs when those involved in the evaluation learn from the evaluation process itself or make program changes based on the evaluation process rather than findings. Process use also includes the effects of evaluation procedures and operations, for example, the premise that "what gets measured gets done," so establishing measurements and setting targets affects program operations and management focus. (See Chapter 5, Menu 5.1, for different types of process use.)

(Continued)

(Continued)

Longer Term, More Incremental Influences

Influence intentionally broadens thinking about evaluation impacts by attending to “the capacity or power of persons or things to produce effects on others by intangible or indirect means” (Kirkhart 2000:7). Influence draws attention to effects of an evaluation over time and beyond the specific program evaluated. Influence can be intended or unintended, and can flow from either results or the evaluation process.

Enlightenment adds a longer time dimension and connotes a broader policy scope to conceptual use. It involves the gradual percolation of ideas from evaluation into policy discourse, changing understandings, questioning assumptions, evoking new ideas, and altering priorities (Weiss 2004).

Primarily Political Uses

Symbolic use refers to token support for an evaluation process or findings but with no real intent to take either the process or findings seriously. Symbolic use can be helpful when it creates a supportive environment for others to make serious use of evaluation processes and findings.

Legitimative use occurs when evaluation findings are used to support and justify a decision that was already made before the evaluation was ever conducted.

Persuasive use refers to using evaluation findings, often quite selectively, to support one’s position in funding decisions and political debates. This is not necessarily inappropriate, for instance, when evaluation results support a course of action that already has advocates and they appropriately use findings to support their position (Weiss et al. 2005).

Imposed use occurs when those with the power to do so mandate a particular form of evaluation use, usually when those at a higher level of authority require a prescribed use by those at a lower level. For example, a federal requirement that to receive funding a school district curriculum must be on an approved list of “evidence-based” or evaluated programs (Weiss et al. 2005).

Mechanical use, or compliance use, refers to going through the motions to meet an evaluation requirement. The evaluation is required, so it is done, but the motivation is compliance and the implementation is mechanical.

Misuses

Mischievous misuse includes the calculated and intentional suppression, misrepresentation, or unbalanced use of evaluation findings to influence opinions and decisions.

Inadvertent misuse, also called mistaken misuse, occurs when those using findings lack the background or competence to appropriately interpret findings; spend too little time with the results to fully understand them; are swayed by the evaluator’s status, expertise, or personality rather than the findings; or simply lack the sophistication needed for appropriate use.

Overuse occurs when too much emphasis is placed on evaluation findings. For example, weak evaluation results are overused when treated as if they are definitive, or imposed use (see above) occurs with insufficient evidence or lack of attention to local conditions. This latter overuse can occur when supposed “best practices” are universally mandated (Patton 2001).

Nonuses

Nonuse due to misevaluation (Patton 2005b:254), or justified nonuse (Cousins and Shula 2006:282) refers to appropriate nonuse because of weak evidence, a late report, poor evaluator performance, or other failures of the evaluator to adhere to the profession’s standards and principles (see Chapter 1).

Political nonuse occurs when the findings are ignored because they conflict with a potential user's values, prejudices, preferences, and predisposition—so the evaluation is just simply ignored.

Aggressive nonuse, or calculated resistance, refers to situations where an evaluation or evaluator is attacked and use undermined because the results conflict with or raise questions about a preferred position. Resistance to evaluation findings can be a specific example of the more general phenomenon of resistance to change.

Unintended Effects

Unintended uses are any applications of evaluation findings or processes that were not planned, not predictable, or unforeseen.

large-scale, long-term, international comparative designs costing millions of dollars to small, short evaluations of a single component in a local agency, the variety is vast. Contrasts include internal versus external evaluations; outcomes versus process evaluation; experimental designs versus case studies; mandated accountability systems versus voluntary management efforts; academic studies versus informal action research by program staff; and published, polished evaluation reports versus oral briefings and discussions where no written report is ever generated. Then there are combinations and permutations of these contrasting approaches. In the midst of such splendid diversity, any effort to reduce the complexity of evaluation options to a few major categories will inevitably oversimplify. Yet some degree of simplification is needed to make the evaluation design process manageable and facilitate interactions with primary intended users about priority purposes. So let us attempt to heed Thoreau's advice:

> Simplicity, simplicity, simplicity! I say, let your affairs be as two or three, and not a hundred or a thousand.
>
> (*Walden* 1854)

A Menu of Intended Uses Based on Alternative Purposes

The last edition of this book highlighted three primary purposes for evaluation: rendering judgments, facilitating improvements, and generating knowledge. In this edition, I have added three additional purposes based on evolution of the field, feedback from readers, and trends in evaluation practice: accountability, monitoring, and development. I'll explain these additions and their importance as we go along. Different purposes lead to different uses, and that has implications for every aspect of evaluation—design, measurements, analysis, interpretation, reporting, dissemination, and criteria for judging quality.

Summative, Judgment-Oriented Evaluation

Evaluations aimed at determining the overall merit, worth, significance, or value of something are judgment oriented. Merit refers to the intrinsic value of a program, for example, how effective it is in meeting the needs of those it is intended to help. Worth refers to extrinsic value to those outside the program, for example, to the larger community or society. A welfare program that gets jobs for recipients has *merit* for those who move out of poverty and *worth* to society by reducing welfare costs. Judgment-oriented evaluation approaches include summative evaluations aimed at deciding if a program is sufficiently effective to be continued or replicated and comparative ratings or rankings of programs as done by *Consumer Reports*. These judgments are used to inform decisions. In the case of programs, the decisions concern whether to continue a program, expand it, or change it in some major way. In the case of consumer products, the judgments inform decisions about whether to purchase a particular item.

The first clue that intended users are seeking an overall, summative judgment is when you hear the following kinds of questions: Did the program work? Did it attain its goals? Should the program be continued, ended, or expanded to other sites? Did the program provide good value for money? Can the outcomes measured be attributed to the program? Answering these kinds of evaluative questions requires a data-based judgment that some need has been met, some goal attained, or some standard achieved.

In judgment-oriented evaluations, specifying the criteria for judgment is central and critical. Different stakeholders will bring different criteria to the table. During design discussions and negotiations, evaluators may offer additional criteria for judgment beyond those initially thought of by intended users. Clarifying the values that will be the basis for judgment is a central role for evaluators. The standard to be met in this regard has been articulated in the Joint Committee Program Evaluation Standards: "*Values Identification:* The perspectives, procedures, and rationale used to interpret the findings should be carefully described, *so that the bases for value judgments are clear* [italics added]" (Joint Committee 1994:U4).

Summative evaluation constitutes an important purpose distinction in any menu of intended uses. Summative evaluations judge the *overall effectiveness of a program* and are particularly important in making decisions about continuing or terminating an experimental program or demonstration project. As such, summative evaluations are often requested by funders. Summative evaluation contrasts with *formative evaluation*, which focuses on ways of improving and enhancing programs rather than rendering definitive judgment about effectiveness. Michael Scriven (1967:40–43) introduced the summative-formative distinction in discussing evaluation of educational curriculum. The distinction has since become a fundamental evaluation typology.

With widespread use of the summative-formative distinction has come misuse, so it is worth examining Scriven's own definition:

> Summative evaluation of a program (or other evaluand) is conducted *after* completion of the program (for ongoing programs that means after stabilization) and *for* the benefit of some *external* audience or decision-maker (for example, funding agency, oversight office, historian, or future possible users). . . . The decisions it services are most often decisions between these options: export (generalize), increase site support, continue site support, continue with conditions (probationary status), continue with modifications, discontinue. . . . The aim is to report *on* it [the program], not to report *to* it. (Scriven 1991b:340).

Summative evaluation provides data to support a judgment about the program's worth so that a decision can be made about the merit of continuing the program. While Scriven's definition focuses on a single program, summative evaluations of multiple programs occur when, like the products in a *Consumer Reports* test, programs are ranked on a set of criteria such as effectiveness, cost, sustainability, quality characteristics, and so forth. Such data support judgments about the comparative merit or worth of different programs. Exhibit 4.2 provides an example of a summative evaluation.

When decisions are made using evaluative judgments, evaluation results are combined with other considerations to support decision making. Politics, values, competing priorities, the state of knowledge about a problem, the scope of the problem, the history of the program, the availability of resources, public support, and managerial competence all come into play in program and policy decision processes. Evaluation findings, if used at all, are usually one piece of the decision-making pie, not the whole pie. Rhetoric about "data-based decision making" and "evidence-based practice" can give the impression that one simply looks at evaluation results and a straightforward decision follows. *Erase that image from your mind.* That is seldom, if ever, the case. Evaluation findings typically have technical and methodological weaknesses; data must

EXHIBIT 4.2

A Judgment-Oriented Summative Exemplar: Evaluating Home Visitation

The David and Lucile Packard Foundation employed an evaluation-focused grant-making strategy over more than a decade in funding the home visitation approach to supporting child development. The Foundation's rigorous evaluation of the home visitation model over many years was selected as a featured case for teaching evaluation published by *New Directions for Evaluation* (Sherwood 2005). The Packard Foundation first got involved with home visitation because of a grant request in 1987 from a group of school districts in the Salinas Valley of Monterey County, California, to adapt and implement a child development model called parents as teachers (PAT).The program provides education to parents about effective interaction with their children for learning and developmental screening for children in the first 3 years of life. PAT was also planned as an extension of school services that would be available to all parents within the community. As a result, the service population was predominantly low-income and Hispanic parents in the Salinas Valley.

At the time of the program proposal, there was increasing interest nationally in the 0 to 3 age group, early intervention programs to prevent child abuse and neglect and developmental delays among children in high-risk groups, and programs to enhance school readiness. Home visiting as an intervention model crosscut this broad range of child development activity. The general public and policymakers were paying attention to brain development research that highlighted the lasting effects of early childhood experiences.

The Packard Foundation funded a demonstration project of PAT that included evaluation of the PAT model. The highly regarded SRI International conducted the evaluation, which concluded that there were "consistent and strong beneficial effects from PAT participation on virtually all measures included in the evaluation. . . Clearly PAT is an effective intervention for improving parenting knowledge, attitudes, and behaviors and for supporting positive child development" (quoted in Sherwood 2005:64). Based on the evaluation results, the Foundation decided to go forward with a full-scale program and a more comprehensive random assignment evaluation.

be interpreted; other contextual factors must be taken into consideration. In short, evaluation use is a complex process. Utilization-focused evaluation acknowledges and deals with those complexities to increase the likelihood that evaluation findings are appropriately and meaningfully used.

Understanding the Decision Contexts of Potential Users

Those who study evaluation use would be well advised to focus on the decision contexts of the potential users. The reasons include the need to fit evaluation findings into the users' existing construction of reality and the expertise that the potential users bring to the context. High payoff evaluations are likely to be those for which the questions have been framed by a structured process. These are likely to reduce uncertainty about important issues and test assumptions about policy, programs, social needs, and service delivery. To control legitimation: Provide a good enough product, control the spin, and seek utilization where positions have not yet hardened (Leviton 2003:533–34).

In summative, judgment-oriented evaluations, what Scriven (1980) has called "the logic of valuing" rules. Four steps are necessary: (1) Select criteria of merit; (2) set standards of performance; (3) measure performance; and (4) synthesize results into a judgment of value (Shadish, Cook, and Leviton, 1991:73, 83–94). Selecting criteria for judging success can be a complicated and time-consuming process when large numbers of stakeholders are involved. Gary Henry (2002) used a values inquiry approach to identify criteria for success of a public preschool program by surveying four groups of stakeholders: teachers, administrators, parents, and the public. Different values preferences and varying criteria lead to different judgments about success. Jane Davidson (2005) in her "nuts-and-bolts" approach describes six strategies for determining judgment criteria (pp. 105–28). See Exhibit 4.3.

Improvement-Oriented, Formative Evaluation

Using evaluation results to improve a program turns out, in practice, to be fundamentally different from rendering judgment about overall effectiveness, merit, or worth. Improvement-oriented forms of evaluation include formative evaluation, quality enhancement, learning organization approaches, and continuous quality improvement (CQI), among others. What these approaches share is a focus on improvement—making things better—rather than rendering summative judgment. Judgment-oriented evaluation requires preordinate, explicit criteria and values that form the basis for judgment. Improvement-oriented approaches tend to be more open-ended, gathering varieties of data about strengths and weaknesses with the expectation that both will be found and each can be used to inform an ongoing cycle of reflection and innovation. Program management, staff, and sometimes participants tend to be the primary users of improvement-oriented findings, while funders and external decision makers tend to use judgmental evaluation, though I hasten to add that these associations of particular categories of users with specific types of evaluations represent utilization tendencies, not definitional distinctions; any category of user may be involved in any kind of use.

Improvement-oriented evaluations ask the following kinds of questions: What are the program's strengths and weaknesses? To what extent are participants progressing toward the desired outcomes? Which types of participants are making good progress

EXHIBIT 4.3

Six Strategies for Determining the Importance of the Evaluative Criteria

1. Having stakeholders or consumers "vote" on importance
2. Drawing on the knowledge of selected stakeholders
3. Using evidence from the literature
4. Usually specialist judgment
5. Using evidence from the needs and values assessments
6. Using program theory and evidence of causal linkages

SOURCE: Davidson (2005:105–28).

and which types aren't doing so well? What kinds of implementation problems have emerged and how are they being addressed? What's happening that wasn't expected? How are staff and clients interacting? What are staff and participant perceptions of the program? What do they like? Dislike? Want to change? What are perceptions of the program's culture and climate? How are funds being used compared with initial expectations? How is the program's external environment affecting internal operations? Where can efficiencies be realized? What new ideas are emerging that can be tried out and tested?

The flavor of these questions—their nuances, intonation, feel—communicate improvement rather than judgment. Bob Stake's metaphor explaining the difference between summative and formative evaluation can be adapted more generally to the distinction between judgmental evaluation and improvement-oriented evaluation: "When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative" (quoted in Scriven 1991b:169). More generally, anything done to the soup during preparation in the kitchen is improvement oriented; when the soup is served, judgment is rendered, including judgment rendered by the cook that the soup was ready for serving (or at least that preparation time had run out).

The metaphor also helps illustrate that one must be careful to stay focused on intent rather than activities when differentiating purposes. Suppose that those to whom the soup is served are also cooks, and the purpose of their tasting the soup is to offer additional recipe ideas and consider potential variations in seasoning. Then, the fact that the soup has moved from kitchen to table does not mean a change in purpose. Improvement remains the primary agenda. Final judgment awaits another day, a different serving—unless, of course, the collection of cooks suddenly decides that the soup as served to them is already perfect and no further changes should be made. Then, what was supposed to be formative would suddenly have turned out to be summative. And thusly are purposes and uses often confounded in real-world evaluation practice.

Formative evaluation typically connotes collecting data for a specific period of time, usually during the start-up or pilot phase of a project, to improve implementation, solve unanticipated problems, and make sure that participants are progressing toward desired outcomes. Often the purpose of formative evaluation is to get ready for summative evaluation, that is, to get the program's early implementation bugs worked out and the model stabilized so that it can be evaluated summatively to judge merit and worth. Exhibit 4.4 provides an example of how formative evaluation can prepare a program for summative evaluation by connecting these separate and distinct evaluation purposes to separate and distinct stages in the program's development. As the example also shows, the information needed for improvement is typically different from the data needed for summative judgment.

EXHIBIT 4.4

Formative and Summative Evaluation of The Saint Paul Technology for Literacy Center (TLC): A Utilization-Focused Model

TLC was established as a 3-year demonstration project to pilot test the effectiveness of an innovative, computer-based approach to adult literacy. The pilot project was funded by six Minnesota Foundations and the Saint Paul Schools at a cost of $1.3 million. The primary intended users of the evaluation were the school superintendent, senior school officials, and School Board Directors who would determine whether to continue and integrate the project into the district's ongoing community education program. School officials and foundation donors participated actively in designing the evaluation. The evaluation cost $70,300.

After 16 months of formative evaluation, the summative evaluation began. The formative evaluation, conducted by an evaluator hired to be part of the TLC staff, used extensive learner feedback, careful documentation of participation and progress, and staff development activities to specify the TLC model and bring implementation to a point of stability and clarity where it could be summatively evaluated. The summative evaluation, conducted by two independent University of Minnesota social scientists, was planned as the formative evaluation was being conducted.

The summative evaluation began by validating that the specified model was, in fact, being implemented as specified. This involved interviews with staff and students, and observations of the program in operation. Outcomes were measured using the Test of Adult Basic Education administered on a pre-post basis to participant and control groups. The test scores were analyzed for all students who participated in the program for a 3-month period. Results were compared with data available on other adult literacy programs. An extensive cost analysis was also conducted by a University educational economist. The report was completed 6 months prior to the end of the demonstration, in time for decision makers to use the results to determine the future of the program. Retention and attrition data were also analyzed and compared with programs nationally.

Comparisons showed significant gains in reading comprehension and math for the participant group versus no gains for the control group. Adult learners in the program advanced an average of one grade level on the test for every 52.5 hours spent in TLC computer instruction. However, the report cautioned that the results showed great variation: high standard deviations, significant differences between means and medians, ranges of data that include bizarre extremes, and very little correlation between hours spent and progress made. The report concluded, "Each case is relatively unique. TLC has created a highly individualized program where learners can proceed at their own pace based on their own needs and interests. The students come in at very different levels and make very different gains during their TLC work . . . , thus the tremendous variation in progress" (Council on Foundations 1993:142).

Several years after the evaluation, the Council on Foundations commissioned a follow-up study on the evaluation's utility. The Saint Paul Public Schools moved the project from pilot to permanent status. The Superintendent of Schools reported that "the findings of the evaluation and the qualities of the services it had displayed had irrevocably changed the manner in which adult literacy will be addressed throughout the Saint Paul Public Schools" (Council on Foundations 1993:148). TLC also became the basis for the District's new Five Year Plan for Adult Literacy. The evaluation was so well-received by its original philanthropic donors that it led the Saint Paul Foundation to begin and support an Evaluation Fellows program with the University of Minnesota. The independent Council on Foundations follow-up study concluded, "Everyone involved in the evaluation—TLC, funding sources, and evaluators—regards it as a utilization-focused evaluation. . . . The organization and its founders and funders decided what they wanted to learn and instructed the evaluators accordingly" (Council on Foundations 1993:154-55). The formative evaluation was used extensively to develop the program and get it ready for the summative evaluation. The summative evaluation was then used by primary intended users to inform a major decision about the future of computer-based adult literacy. Ten years, later Saint Paul's adult literacy effort continues to be led by TLC's original developer and director.

SOURCES: Turner and Stockdill (1987); Council on Foundations (1993:129–55).

Accountability

> Accountability is a state of, or process for, holding someone to account to someone else for something—that is, being required to justify or explain what has been done. Although accountability is frequently given as a rationale for doing evaluation, there is considerable variation in who is required to answer to whom, concerning what, through what means, and with what consequences. More important, within this range of options, the ways in which evaluation is used for accountability are frequently so poorly conceived and executed that they are likely to be dysfunctional for programs and organizations. (Rogers 2005a:2)

This astute conclusion by Australian Patricia Rogers, the first international recipient of the American Evaluation Association's prestigious Myrdal Award for contributions to evaluation use, frames the challenge of bringing utility to the very political undertaking of supporting accountability. More than a quarter century ago, in positing 95 theses for reform of evaluation, Lee J. Cronbach and associates (1980) at Stanford posited,

> A call for accountability is a sign of pathology in the political system. . . . Accountability emphasizes looking back in order to assign praise or blame; evaluation is better used to understand events and processes for the sake of guiding future activities. (P. 4)

Cronbach's distinction between the uses of accountability and evaluation continues to be debated today. Are accountability systems really evaluative or are they primarily political and managerial? In the last edition of this book, I incorporated accountability within judgment-oriented evaluation. However, in practice, these involve significantly different uses. One important reason for distinguishing and separating judgmental/summative evaluation from accountability is articulated by Rogers (2005a):

> Accountability systems focus on reporting discrepancies between targets and performance to funders, the assumption being that they will use this information in future funding and policy decisions. However, accountability systems rarely provide sufficient information to make it possible for funders to decide if such discrepancies should be followed by decreased funding (as a sanction),

Formative-Summative Confusions

Common misconceptions about the formative-summative distinction

- Formative focuses on process, summative on outcomes. *Not true*. Formative evaluation often gives an early picture of what progress is being made toward desired outcomes and what unanticipated outcomes are emerging. Summative evaluation must describe implementation and processes to discuss and judge the relationship between what was done and what was accomplished.
- Formative is more qualitative while summative is more quantitative. *Not true*. Formative and summative are purpose distinctions, not methods distinctions. The nature and combination of methods used depends on what questions are being asked and what evidentiary criteria are preferred by evaluators and primary intended users as they negotiate the design.
- Summative is judgmental while formative is descriptive. *Not true*. The difference is a matter of degree. Summative evaluation involves a definitive, conclusive judgment of *overall* merit, worth, and value, if possible and the data support such a definitive judgment. Providing formative feedback about what works and doesn't work involves some degree of judgment against criteria related to the notion of what it means for a program to "work," but formative judgments tend to be directed at specific aspects of a program (rather than the overall program) and involve lower stakes decisions than does overall summative judgment. Because of the focus on learning and improvement, formative evaluation typically *feels* less judgmental to staff and participants.
- Summative is definitive while formative is tentative. *Not true*. This depends on the nature of the evidence. While a summative evaluative aims to be definitive, the evidence may not be sufficient to support a definitive judgment. On the other hand, formative evidence about the need for improvement can be quite definitive.
- The formative versus summative distinction is context dependent. *True.* This means that a certain type of evaluation, for example, an impact evaluation, cannot be considered intrinsically summative. An impact evaluation can be used to improve the next stage in the life of a program. Qualitative feedback from participants and in-depth case studies can be used summatively when the results show little or no value from the perspective of intended beneficiaries. Scriven, originator of the distinction, emphasizes that the distinction is "not intrinsic, it's contextual — *mainly a matter of the use to which the evaluation is put* [italics added]. . . . In introducing the distinction between formative and summative, I stressed that this was a difference in roles, not of intrinsic nature. And roles are defined by context" (Scriven 1996:153).

> increased funding (to improve the quality or quantity of services being provided) clock, or termination of the function. (Pp. 3–4)

The accountability function includes *oversight and compliance:* "the assessment of the extent to which a program follows the directives, regulations, mandated standards or any other formal expectations" (Mark, Henry, and Julnes 2000:13).

Performance measurement is a common approach to oversight, compliance, and accountability. Burt Perrin (2002, 1998) has long been a leader in studying the "effective use and misuse of performance measurement." He has been especially adamant about the limitations of performance indicator approaches for evaluation asserting that such data are "useless for decision making and resource allocation" (1998:374). Why? Because a performance indicator alone doesn't tell a decision maker why the results are at a certain level and without knowing why, informed action

The Utility of an Accountability System

The utility of an accountability system depends on who is held accountable, by whom, for what—and how they are held accountable, that is, the extent to which results can be determined and explained, and that there are consequences for failure and rewards for success. The credibility of an accountability system, which greatly affects its utility, depends on the extent to which those held accountable actually have the capacity to achieve those things over which they are held accountable, within the timeframes expected, and that the consequences are proportionately and reasonably aligned with that capacity and those timeframes.

is problematic. In essence, accountability systems serve the purpose of providing *an account of how things are going* but not enough information to inform decisions or solve problems. Those actions require deeper evaluative data than accountability systems usually provide (Bemelmans-Videc, Lonsdale, and Perrin 2007; Mayne 2007; Owen 2007; Perrin 2007).

A comprehensive accountability approach involves both description—What was achieved?—and explanation—How and why was it achieved at the levels attained? To describe is not to explain, and to explain is not to excuse or diminish responsibility. Ideally, description, explanation, and responsibility can be combined to produce an effective and useful accountability system. Description, however, comes first. Having an accurate account of how things are going, including what results are being attained, is essential. Explaining those results and assigning responsibility follow. And that's where it all becomes very political.

Accountability is like a red cape in front of a bull in the political arena where politicians fancy themselves as matadors braving the horns of waste and corruption. Funders and politicians issue shrill calls for accountability (notably for others, not for themselves), and "managing for accountability" (Kearns 1996) has become a rallying cry in both private and public sectors. In its extreme bean-counting manifestation, this can become what Weinberger (2007) has called "The Folly of Accountabalism."

Program and financial audits are aimed at assuring compliance with intended purposes and mandated procedures. The program evaluation units of legislative audit offices, offices of comptrollers and inspectors, and federal agencies such as the OMB have government oversight responsibilities to make sure programs are properly implemented and effective. Reflecting the increased emphasis on accountability in government, in 2004, the legal name of the Congressional oversight agency, GAO, changed its name to the Government Accountability Office instead of the General Accounting Office, a designation it had had for 83 years. The U.S. Government Performance and Results Act of 1993 requires annual performance measurement to "justify" program decisions and budgets. Political leaders in Canada, the United Kingdom, and Australia have been active and vocal in attempting to link performance measurement to budgeting for purposes of accountability (Auditor General of Canada 1993) and these efforts greatly influenced the United States federal approach to accountability (Breul 1994).

Accountability concerns are driven by the following kinds of questions: Are funds being used for intended purposes? Are goals and targets being met? Are indicators showing improvement? Are resources being efficiently allocated? Are problems being handled? Are staff qualified? Are

Accountability for Utilization

GAO, as the largest internal, independent evaluation unit in existence, has a distinguished history of paying attention to how its evaluations are used. Every recommendation in its numerous reports is followed to find out whether its findings are adopted. In 2004, for example, GAO made 1,950 recommendations. In its own internal utilization study, GAO found that 80 percent of its recommendations to improve government operations were implemented from 2000 to 2004 (Mathison 2005:168).

only eligible participants being accepted into the program? Is implementation following the approved plan? Are quality control mechanisms in place and being used?

The varying contexts within which such questions are asked matter a great deal. In government, accountability issues inevitably find their way into debates between those in power and those out of power. In philanthropy, accountability "satisfies the fiduciary responsibility of a foundation to oversee the use of money and to ensure that grant funds were spent according to its terms. Evaluation, therefore, provides the evidence for both grantee and foundation accountability" (Kramer and Bickel 2004:53). For not-for-profit agencies and nongovernmental organizations, accountability is part of good management. In all these contexts, accountability-oriented evaluation is manifesting one of the major historical streams that flow into the large ocean of evaluation: the audit stream (Wisler 1996).

In the public sector, rhetoric about accountability can become particularly strident in the heat of political campaigns. Everyone campaigns against ineffectiveness, waste, and fraud. Yet one person's waste is another's jewel. For years, U.S. Senator William Proxmire of Wisconsin periodically held press conferences in which he announced *Golden Fleece Awards* for government programs he considered especially wasteful. I had the dubious honor of being the evaluator for one such project ridiculed by Proxmire, a project to take higher education administrators into the wilderness to experience, firsthand, experiential education. The program was easy to make fun of: Why should taxpayer dollars be spent for college deans to hike in the woods? Outrageous! What was left out of Proxmire's press release was that the project, supported by the Fund for the Improvement of Postsecondary Education, had been selected in a competitive process and funded because of its innovative approach to rejuvenating burned-out and discouraged administrators, and that many of those administrators returned to their colleges to spearhead curriculum reform. There was lots of room for debate about the merit or worth of the program *depending on one's values and priorities*, but our evaluation found that the funds were spent in accordance with the agency's innovative mandate and many, though not all, participants followed through on the project's goal of providing leadership for educational change. The funding agency found sufficient value that the project was awarded a year-long dissemination grant.

Some criteria, such as fraud and gross incompetence, are sufficiently general and agreed-on that when uncovered and given media attention, they inevitably raise the crescendo of voices lamenting the offending program's lack of accountability. One of my favorite examples comes from a program audit of a weatherization program in Kansas as reported in the newsletter of Legislative Program Evaluators.

> Kansas auditors visited several homes that had been weatherized. At one home, workers had installed 14 storm windows to cut down on air filtration in the house. However, one could literally see through the house because some of the siding had rotted and either pulled away from or fallen off the house. The auditors also found that the agency had nearly 200 extra storm windows in stock. Part of the problem was that the supervisor responsible for measuring storm windows was afraid of heights; he would "eyeball" the size of second-story windows from the ground. . . . If these storm windows did not fit, he ordered new ones. (Hinton 1988:3)

The auditors also found fraud. The program bought windows at inflated prices from a company secretly owned by a program employee. A kickback scheme was uncovered. "The workmanship on most homes was shoddy, bordering on criminal. . . . [For example], workers installing a roof vent used an ax to chop a hole in the roof." Some 20 percent of beneficiaries didn't meet eligibility criteria. Findings such as these are thankfully rare, but they grab headlines when they become public, and they illustrate why accountability will remain a central purpose of many evaluations.

The extent to which concerns about accountability dominate a specific study varies by the role of the evaluator. For auditors, accountability is always primary. Public reports on performance indicators for government programs are accountability driven. Performance measurement follows the mantra that "what gets measured gets done." But for an accountability system to have integrity and credibility, there needs to be some separation between the measuring and the doing, or at least some independent way of verifying the accuracy of internally generated accountability data. Burt Perrin, in a presentation on accountability at the European Evaluation Society annual conference, Seville, Spain, called attention to an article in *Nature* about how statistics reported by China regarding the fish catches by its fishery had been grossly distorted. "Apparently, under the Communist system of matching results with plan, the same bureaucrats were responsible for not only counting the catch but also meeting targets to increase it—so they simply exaggerated the count to match their allotted goals" (*International Herald Tribune* 2001).

What Gets Measured Gets Done

In Poland, as manufacturing shifted from communism to capitalism, performance incentives were introduced and the performance of furniture factories was measured by the tons of furniture shipped. Responding to this incentive system—what gets measured gets done—Poland came to have the heaviest furniture in the world (Perrin 2002:368).

Elliot Stern, president of the International Organization for Cooperation in Evaluation, has long expressed concerned that "most accountability systems encourage a blame culture." He sees this as part of "the wider preoccupation with regulation and control as part of public management today."

> When programmes do not achieve their targets or when policy instruments appear not to work, a first reflex is to identify the guilty party and remove or relocate him. Accountability is after all one of the acknowledged main purposes of evaluation. (Stern 2004:12)

Accountability systems, then, pose special challenges for evaluation, especially in implementing high-quality systems that are useful and credible, and overcoming the tendencies of such systems to become politicized and corrupted. Canada has had some success with such systems (Fraser 2006; Mayne 2006; Schwartz and Mayne 2004). The situation at the federal level in the United States is more problematic as the Bush administration instigated a new accountability system alongside and on top of the existing Clinton administration system. See Exhibit 4.5.

To be useful beyond providing meat for political dog fights, accountability systems need to be designed with utility in mind. Rogers has identified critical characteristics of such a useful system, what she calls *smart accountability:*

> Accountability requires a much more comprehensive explanation of performance, an incentive system that encourages improvement of performance rather than misreport and distortion of it, and a commitment to address learning as well as accountability. In other words, accountability systems need to be a tool for informed judgment and management rather than a substitute. This is the smart accountability that is been increasingly advocated.
>
> Smart accountability includes demonstrating responsible, informed management; including appropriate risk management, such as cautious trials of difficult or new approaches; and a commitment to identify and learn from both successes and mistakes. The incentive system for accountability needs to reward *intelligent failure* (competent implementation of something that has since been found not to work), discourage setting easy targets, discourage simply reporting compliance with processes or targets, and encourage seeking out tough criticism.
>
> The acid test of a good accountability system is that it encourages responsibility and promotes better performance. (Rogers 2005a:4)

Performance Measurement Challenges

Many activities are in the public sector precisely because of measurement problems: If everything was so crystal clear and every benefit so easily attributable, those activities would have been in the private sector long ago.

SOURCE: Mintzberg (1996:76), Strategic Management Scholar

Performance Measurement: A View from the Trenches

I have been working now for about 20 years in the area of evaluation and performance measurement, and I am so discouraged about performance measurement and results reporting and its supposed impact on accountability that I am just about ready to throw in the towel. So I have had to go right back to the basics of reporting and democracy to try to trace a line from what was intended to what we have ended up with . . . .

Performance measurement has been oversold - it makes promises that are not easily kept, and I honestly believe now that it has become a paper exercise for departments, and is too boring and technical for the public or Legislators to have the time or interest to read. What ever happened to good old monitoring?

Karyn Hicks *EvalTalk* posting
Programs Advisor July 28, 2006, Government of the Northwest Territories. Used with permission
Yellowknife, Canada

EXHIBIT 4.5

Accountability: Too Much of a Good Thing?

GPRA and PART as Dueling Banjos

The Clinton/Gore Administration's effort to "reinvent government" led to the 1993 Government Performance and Results Act (GPRA). This major legislation aimed to shift the focus of government decision making and accountability away from a preoccupation with reporting on activities to a focus on the results of those activities, such as real gains in employability, safety, responsiveness, or program quality. Under GPRA, U.S. federal government agencies are required to develop multiyear strategic plans, annual performance plans, and annual performance reports.

In 2001, the U.S. Government Accountability Office (GAO) initiated major reviews of how GPRA was being implemented (www.gao.gov/new.items/gpra/gpra.htm). GAO has continued issuing annual reviews (www.gao.gov/pas/2005) as part of its Performance and Accountability Series. At the beginning of each new Congress, based on its audits and evaluations, GAO identifies federal programs and operations that are "*high risk*" due to their vulnerabilities to fraud, waste, abuse, and mismanagement. GAO has increasingly focused on the need for broad-based transformations to address major economy, efficiency, and effectiveness challenges (GAO 2006a). Those agencies identified as *high risk* receive increased scrutiny both inside and outside government. Follow-up reviews show that GAO's *high risk* evaluations are used to bring about significant change. "Lasting solutions to high-risk problems offer the potential to save billions of dollars, dramatically improve service to the American public, strengthen public confidence and trust in the performance and accountability of our national government, and ensure the ability of government to deliver on its promises" (GAO 2005, Highlights).

Immediately following election in 2000, the Bush administration reiterated a commitment to performance, accountability, and results. To that end, in 2001, the Office of Management and Budget (OMB) began to develop a mechanism called the Program Assessment Rating Tool (PART) to help budget examiners and federal managers measure the effectiveness of government programs. A PART review aims to identify a program's strengths and weaknesses to inform funding and management decisions aimed at making the program more effective. The PART framework aims to evaluate "all factors that affect and reflect program performance including program purpose and design; performance measurement, evaluations, and strategic planning; program management; and program results" (www.whitehouse.gov/omb/part). PART intends to examine program improvements over time and allow comparisons between similar programs. Bill Trochim, Chair, of the American Evaluation Association Public Affairs Committee observed, "PART is one of the more significant evaluation-related items emerging from the US federal government in many years" (Trochim 2006a).

In 2006, OMB launched a Web site (www.ExpectMore.gov) that reports on federal program performance and what is being done to improve results. It opened with nearly 800 PART program assessments. GAO evaluated how federal agencies responded to PART. Their findings focused on implementation rather than evaluation use.

Several agencies struggled to identify appropriate outcome measures and credible data sources before they could evaluate program effectiveness. Evaluation typically competed with other program activities for funds, so managers may be reluctant to reallocate funds to evaluation. Some agency officials thought that evaluations should be targeted to areas of policy significance or uncertainty. However, all four agencies indicated that the visibility of an OMB recommendation brought agency management attention—and sometimes funds—to get the evaluations done. Moreover, by coordinating their evaluation activities, agencies met these challenges by leveraging their evaluation expertise and strategically prioritizing their evaluation resources to the studies that they considered most important (GAO 2006b:3).

(Continued)

(Continued)

Both GPRA and PART involve massive amounts of staff time, money, and paperwork. Both are federal government efforts to increase accountability, evaluate effectiveness, and demonstrate results. How do they relate to each other? Not very well, it turns out. Integration is, at best, a work in progress. They are parallel, often redundant, efforts. The promulgation of competing and redundant government performance measurement systems goes well beyond GPRA and PART and has become a widespread problem stemming from the many different performance measurement approaches and systems introduced at all levels of government (Nicholson-Crotty et al. 2006). Both legislative and executive improvements are proposed regularly, often in recognition that the sheer volume of information reported reduces utility because there is too much to sort through and make sense of. Compliance with mandated reporting trumps meaningfulness. Nor is this simply an American problem. Around the world new performance monitoring systems get created with little sense of what is already in place, with little evaluation of the strengths, weakness, and uses of current information systems, and with inadequate attention to the accuracy, credibility, timeliness, and utility of new systems (Rogers 2006).

The GAO (2004) evaluated how GPRA and PART were being used—an excellent example of a utilization study—and concluded that PART had emerged as a parallel and competing approach with GPRA's Performance Management Framework. Many federal agency officials, they found, viewed PART's program measures as detrimental to and in conflict with their GPRA planning and reporting processes. The relationship between the PART and GPRA was not well-defined, was often confusing to program officials and agency managers, and, ironically, undermined the efforts of both to promote efficiency and accountability, thus defeating the purpose of each, which is, pointedly, to increase efficiency and accountability.

Distinguished public administration scholar Paul Light (2006) reviewed the last six decades of major administrative reforms enacted by the U.S. Congress. He found acceleration in both the number and the variety of reforms attempted, fueled in part by heightened public distrust toward government. Ironically, from an evaluation perspective, part of what drives constant reform, Light found, is a lack of hard evidence about what actually works to improve government performance. New systems are put in place before existing systems have a chance to work, much less be evaluated. Meanwhile, critiques and ideas for still more reforms abound (e.g., Caiden 2006; Kettl et al. 2006; Shipman 2003).

Monitoring: Evaluation's Global Partner

Monitoring is another purpose distinction that is new to this edition. Sometimes, monitoring is subsumed under accountability since both use performance indicators. But that's like treating formative and summative evaluation as the same because they both use data. In fact, performance indicators can serve different purposes, and this is a chapter on purpose distinctions, so it seems to me worth calling attention to the facts that (1) performance indicators can be used for either accountability or ongoing management purposes and (2) these purposes are often in conflict because they involve different primary intended users. Accountability is driven by attention to external stakeholders, those to whom the program is responsible and those who have funded it. Ongoing monitoring serves managers, providing those internal to the program with the information they need to know where their managerial attention is needed.

The other reason for highlighting monitoring as a distinct purpose is that this has become the international norm. In the United States, we talk about evaluation and performance measurement as virtually distinct endeavors. But in developing countries, the standard reference is to "M & E"—monitoring and evaluation. These are close siblings, always together. There are "M & E handbooks," "M & E" conferences, "M & E" workshops. As serendipity would have it, on the very day I was writing this section, an international participant on EvalTalk, the AEA listserv, posted a request for resources on building "M & E capacity." The very first response from an American participant was, "What's M & E?" That sealed the deal. Readers of this book will not have to ask that question.

But there are different approaches to M & E. Ray Rist, coauthor with Jody Zall Kusek (2004) of *Ten Steps to a Results-Based Monitoring and Evaluation System* (see Exhibit 4.6), created the International Program for Development Evaluation Training (IPDET) with his World Bank colleague Linda Morra. That program has trained more development evaluators than any other in the world and the graduates of IPDET, with support and inspiration from Ray, Linda, and others in the international community, have provided the leadership for the International Development Evaluation Association (IDEAS) (www.ideas-int.org).

EXHIBIT 4.6

Ten Steps to a Results-Based Monitoring and Evaluation System

1. Conducting a readiness assessment
2. Agreeing on outcomes to monitor and evaluate
3. Selecting key indicators to monitor outcomes
4. Baseline data on indicators—where are we today?
5. Planning for improvement—selecting results targets
6. Monitoring for results
7. The role of evaluations
8. Reporting findings
9. Using findings
10. Sustaining the M&E system within the organization

SOURCE: Kusek and Rist (2004:25).

Rist travels the world advocating for and training people in a particular kind of M & E system:

> A theoretical distinction needs to be drawn between traditional M&E and results-based M&E. Traditional M&E focuses on the monitoring and evaluation of inputs, activities, and outputs, that is, project or program implementation. There are governments have over time track their expenditures and revenues, staffing levels and resources, program and project activities, numbers of participants, goods and services produced, etc. Indeed, traditional efforts at monitoring have been a function of many governments for many decades or longer. In fact, there is evidence that the ancient Egyptians (5000 B.C.) regularly tracked their government's outputs in grain and livestock production.
>
> Results-based M&E, however, combines the traditional approach of monitoring implementation with the assessment of results. . . . It is this linking of implementation progress (performance) with progress in achieving desired objectives are goals (results) of government policies and programs that makes results-based M&E most useful as a tool for public management (Rist 2006a:4–5)

Most approaches to designing M & E systems intend them to serve both accountability and managerial functions. And therein lies the rub. Policymakers and funders want global, big picture data, what is sometimes called the view from 40,000 feet. Managers need detailed data, the view from 10,000 feet. Aggregating detailed indicators into big picture patterns is one of the major challenges of a performance monitoring system that tries to serve both sets of stakeholders equally well. Still, major texts, while distinguishing between managerial and accountability uses, tend to play down these different uses. Consider how Theodore Poister presents performance monitoring in the influential *Handbook of Practical Program Evaluation* (Wholey, Hatry, and Newcomer 2004):

> Performance monitoring systems are designed to track selected measures of program, agency, or system performance at regular time intervals and report them to managers and other specified audiences on an ongoing basis. Their purpose is to provide objective information to managers and policy makers in an effort to improve decision making and thereby strengthen performance, as well as to provide accountability to a range of stakeholders, such as higher-level management, central executive agencies, governing bodies, funding agencies, accrediting associations, clients and customers, advocacy groups, and the public at large. Thus, performance monitoring systems are critical elements in a variety of approaches to results-oriented management." (Poister 2004:99)

A utilization-focused approach to M & E is less cavalier about such laundry lists of stakeholders and multiple intended uses. Any system will have to set priorities for intended uses by intended users at some point, or risk serving everyone poorly.

The "monitoring and tailoring" approach of Cooley and Bickel (1985) illustrates an approach where school administrators and teachers are the primary intended users. They built a classroom-based information system aimed at systematically tracking daily attendance patterns for individuals, classrooms, and schools. Teachers and administrators could quickly identify attendance problems and intervene before the problems became chronic or overwhelming. Attendance could also be treated as an early warning indicator of other potential problems.

Most monitoring systems look internal (Owen, 1999:239–62). How is program implementation unfolding? What is progress toward desired results? Are we reaching the target population? Are we maintaining quality? Indeed, continuous quality improvement systems (CQI) are one common form of monitoring (Colton 1997). But monitoring systems that include periodic environmental scanning can be especially useful as early warning systems that something in the environment has changed, something that might threaten performance. During a training workshop in South Africa, participants found an M & E metaphor in the vineyards outside Cape Town. The fields of grapevines nestled beneath the green hills are surrounded by fence rows of white roses. Each day the growers inspect the roses. Anything disease or pest that might hard the vines will show up on the roses first. They monitor the roses to decide if action is needed to protect the grapes.

At the policy and resource allocation level, a major challenge has been to connect monitoring to planning and budget cycles (Joyce 1997; Newcomer 1997). Influencing how money is spent may be the ultimate instrumental use for a monitoring system. One long-time dream has been to tie performance results to the budget process, thus increasing attention to results and, hopefully, utility, by increasing the stakes. This is a program-level application of the idea of pay-for-performance in personnel evaluation in which executives and staff who excel get bonuses and special recognition while poor performers get weeded out. This sounds reasonable, even ideal, but proves quite complicated in practice. See Exhibit 4.7 for a review of how the federal PART system has approached the connection to budget.

As evidenced by periodic discussions on the *EvalTalk*, the American Evaluation Association listserv, evaluators disagree about how monitoring and in-depth evaluation studies are related. Hatry et al. (2004) have looked closely and thoughtfully at this issue, bringing great expertise and experience to consider what works. They acknowledge that

> performance monitoring seeks primarily to assess the outcomes of a program without any in-depth examination of the program. . . . In-depth evaluations are considerably more informative and provide considerably more information for major policy and program decisions. . . . We believe these processes are complementary. We believe that performance monitoring can and should be considered an important subset of program evaluation. (p. 676)

The phrase M & E makes the marriage of monitoring and evaluation explicit. In particular, findings from monitoring data can generate questions to be answered by evaluation through more in-depth inquiry, helping to focus and increase the utility of scare evaluation resources. Kusek and Rist (2004) emphasize the integration of monitoring and evaluation in a well-designed, well-implemented, and result-oriented M & E system:

> We want to stress the complementarity of evaluation to monitoring. Each supports the other-even as each asks different questions and will likely make different uses of information and analyses. The immediate implication is that moving to a result-based M&E system requires building an information and analysis system with two components-monitoring and evaluation. Either alone, in the end, is not sufficient. (P. 114)

As always we return to the issue of use. Ongoing and continuous monitoring systems, like all useful evaluation approaches, must be designed to meet the specifiable information needs of identifiable users. A system designed by software, technology, and data experts with little or no serious input and pilot-testing with intended users can be a wonderful system—for the experts who designed it, but not for the intended users.

EXHIBIT 4.7

Follow the Money

A performance monitoring system shows weak results. What are the budget implications of such a finding? Often a primary reason a program has poor results is that it has inadequate resources to achieve quite grandiose goals. If a program is producing poor results, do you kill it or increase its resources so it can improve? Answering this question involves more than a simple report-card grade that the program is good or bad. You need to know why the program is struggling and whether increased resources could be well used.

Consider the situation of a student struggling in a course. Do you just flunk the student or try to provide tutoring and extra help? Does the student have special needs? What else is going on in the student's life? Is the problem in this course part of a long-term pattern of underachievement or is the student's poor performance new? The decision to fail the student or provide extra help will depend, then, on why the student is struggling and an assessment of whether tutoring will help. What does this have to do with government performance monitoring and evaluation?

In the United States, the President's fiscal year 2006 budget elevated the importance of federal PART accountability reviews (see Exhibit 4.5) and increased their visibility by asserting that the budget process was influenced by measures of the success of programs in meeting goals and "identifies which are achieving their intended results and which are not . . . and helps the Administration to reward only those [programs] that succeed" (White House 2006:4). Based on this analysis, the President's budget identified a list of 154 programs slated for deep cuts or elimination because those programs were "not getting results." That sounds straightforward, even laudatory, but here's where the story gets interesting.

OMB Watch is an independent, not-for-profit organization founded in 1983 to increase transparency in the policy-making process. It is funded primarily by philanthropic foundations and has been a thorn in the side of both Republican and Democratic administrations as it has analyzed and evaluated the details of policies and budgets. *OMB Watch* analyzed the list of programs to be cut in the President's 2006 budget and compared program funding requests with the ratings received under the PART. Here is what they found:

> Out of the list of 154 programs to be cut or eliminated, supposedly for lack of results, more than two-thirds have never even been reviewed by the PART. It is unclear what kinds of determinations, if any, the administration used to identify these failing programs when the White House budget staff had yet to assess them.
>
> - Of the 85 programs receiving a top PART score in 2006, the president proposed cutting the budgets of more than 38 percent, including the National Center for Education Statistics.
> - Of all the programs reviewed on the list of 154, nearly 20 percent of programs receiving an "effective" or "moderately effective" PART score—the two highest ratings—were targeted for elimination. Further, 46 percent of programs receiving the middle rating of "adequate" were proposed to be eliminated.
> - Some programs receiving the lowest score were not cut. For instance, the Substance Abuse Prevention and Treatment Block Grant, a program that provides grants to states to address addiction problems, was given the lowest possible rating of "ineffective" but received no reduction in funding. Moreover, the Earned Income Tax Credit Compliance Program—which targets poor people who have claimed the EITC and double-checks their eligibility for the credit—was rated ineffective, yet it received a funding increase. (Hughes and Shull 2005:4).

The analysis of all programs rated under PART since its inception revealed no logical or consistent connections with budget requests. On the face of it, this judgment sounds like strong criticism. But it is only a negative finding when interpreted in the context of the promise to base budget decisions on PART ratings of program performance. As noted earlier, there is good reason to be skeptical about the wisdom of any such simple and mechanical approach to budgeting: highly rated programs get more funds; poorly rated programs get cuts. Performance ratings can and should be one factor in budget decisions, but not the only factor. Those ratings must be interpreted and used within a larger context taking into account factors such as what alternatives are available, what the program has learned about what works and doesn't work that could improve future performance, the track record of managerial competence, how much support the program has among important political constituencies, overall state of the economy and the federal budget, and competing program priorities, to name but a few factors.

Knowledge-Generating Evaluation

> *Whoever undertakes to set himself up as a judge of Truth and Knowledge is shipwrecked by the laughter of the gods.*
>
> —Albert Einstein

In the knowledge age, what could be more useful than contributing to knowledge? Despite Einstein's caution, the evaluation profession has set its sights on knowledge generation, in part because of the great potential for use. The instrumental uses of summative and formative evaluation concern judgments about and improvements for specific programs. Accountability and monitoring also focus typically on performance indicators for a particular program. Knowledge generation, however, changes the unit of analysis as evaluators look across findings from different programs to identify general *patterns of effectiveness*. Knowledge generation, then, has emerged as one of the principal purposes of evaluation (Chelimsky 1997).

As the field of evaluation has matured and a vast number of evaluations has accumulated, the opportunity has arisen to look beyond and across findings about specific programs to formulate generalizations about processes and interventions that make a difference. This involves synthesizing findings from different studies, a strategy the GAO has found useful in providing accumulated wisdom to Congress about how to formulate effective policies and programs (GAO 1992c). A classic example was GAO's report (1992b) on "Adolescent Drug Use Prevention Drug Use Prevention: Common Features of Promising Community Programs." See Exhibit 4.8.

An excellent and important example of synthesis evaluation is Lisbeth Schorr's (1988) *Within Our Reach*, a study of programs aimed at breaking the cycle of poverty. She identified "the lessons of successful programs" as follows (pp. 256–83):

- offering a broad spectrum of services;
- regularly crossing traditional professional and bureaucratic boundaries, i.e., organizational flexibility;
- seeing the child in the context of family and the family in the context of its surroundings, i.e., holistic approaches;

EXHIBIT 4.8

Example of a Knowledge-Oriented Evaluation Synthesis: Common Features of Promising Community Programs Engaged in Adolescent Drug Use Prevention

Six features associated with high levels of participant enthusiasm and attachment:

1. a comprehensive strategy,
2. an indirect approach to drug abuse prevention,
3. the goal of empowering youth,
4. a participatory approach,
5. a culturally sensitive orientation, and
6. highly structured activities.

Six common program problems:

1. maintaining continuity with their participants,
2. coordinating and integrating their service components,
3. providing accessible services,
4. obtaining funds,
5. attracting necessary leadership and staff, and
6. conducting evaluation.

SOURCE: GAO (1992b).

- coherent and easy-to-use services;
- committed, caring, results-oriented staff;
- finding ways to adapt or circumvent traditional professional and bureaucratic limitations to meet client needs;
- professionals redefining their roles to respond to severe needs; and
- overall, intensive, comprehensive, responsive and flexible programming.

These kinds of "lessons" constitute accumulated wisdom—principles of effectiveness—that can be adapted, indeed, must be adapted, to specific programs, organizations, or even broader initiatives like community change (Auspos and Kubisch 2004).

Earlier in this chapter, in reviewing judgment-oriented use, Exhibit 4.2 offered an example of a summative of a Packard Foundation grant for home visitation. The evaluation reached a positive conclusion that led to additional funding. It also led to a significant knowledge building effort that spanned several years as results from multiple home visitation grants, project evaluations, and independent research findings accumulated. As findings from various evaluations and other home-visiting experiments were coming in during the 1996–1998 period, a pattern was emerging of mixed or no significant effects. “The bottom line was small positive effects on a few measures of child development and parenting outcomes for participants who received the expected intensity of service, but very few effects for the overall enrollee groups” (Sherwood 2005:67). What had looked like a promising intervention in the early 1990s had become a disappointment by the end of the decade. Ann Segal, then a senior official in the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, examined the cumulative evidence and concluded that there was no solid evidence that early intervention, via home visiting, with pregnant and parenting teenagers was effective by itself. The lesson she drew was that such programs had been overpromised and that what they set out to accomplish would not work without other, complementary interventions.

> Most home visiting programs promised to do everything–get mothers working, reduce child abuse and neglect, increase literacy, and more. A common sense reading is that these programs aren’t going to get you where you want to go. I take away that the evaluation answer is right–there’s nothing there. *But,* these programs shouldn’t be out there by themselves. You have to hook them onto something stronger. (Segal quoted in Sherwood 2005:70)

This journey from a single program that seemed to have promising outcomes to cumulative evidence that the model is not generally effective is a common evaluation story. Yet the accumulating evidence also shows the importance and value of early childhood interventions (Karoly, Kilburn, and Cannon 2005), for there are models that consistently work, those with better trained home visitors and greater intensity of services. Cronbach and Associates (1980) observed in their 1995 theses that “an evaluation of a particular program is only an episode in the continuing evolution of thought about a problem area” (p. 2). And “in project-by-project evaluation, each study analyzes a spoonful dipped from a sea of uncertainties” (p. 8).

In the philanthropic world, the strategy of synthesizing results from several studies has come to be called “cluster evaluation” (Connor et al. 2004; Millett 1996; Council on Foundations 1993:232–51). A cluster evaluation team visits a number of different grantee projects with a similar focus (e.g., grassroots leadership development) and draws on individual grant evaluations to identify patterns across and lessons from the whole cluster (Sanders 1997; Barley and Jenness 1993; Kellogg Foundation n.d.). The McKnight Foundation commissioned a cluster evaluation of 34 separate grants aimed at aiding families in poverty. One lesson learned was that “effective programs have developed processes and strategies for learning about the *strengths* as well as the needs of families in poverty” (Patton, 1993:10). This “lesson” takes on added meaning when connected with the finding of Independent Sector’s review of “Common Barriers to Effectiveness in the Independent Sector”:

> The deficits model holds that distressed people and communities are “needy”; they’re a collection of problems, pathologies and handicaps; they need doctoring, rehabilitation and fixing of the kind that professionalized services are intended to provide.
>
> The assets model holds that even the most distressed person or community has strengths, abilities and capacities; with investment, their strengths, abilities and capacities can increase. This view is only

> barely allowed to exist in the independent sector, where organizations are made to compete for funds on the basis of "needs" rather than on the basis of "can-do."
>
> The deficit model—seeing the glass half empty—is a barrier to effectiveness in the independent sector. (Mayer 1993:7–8)

The McKnight Foundation cluster evaluation and the Independent Sector study reached similar conclusions concurrently and independently. Such triangulated evaluation findings about principles of effective programming have become the knowledge base of the evaluation profession. Being knowledgeable about patterns of program effectiveness allows evaluators to provide guidance about development of new initiatives, policies, and strategies for implementation. Such contributions constitute the conceptual use of evaluation findings. Efforts of this kind may be considered *research* rather than evaluation, but such research is ultimately evaluative in nature and important to the profession.

Some synthesis evaluations look at large numbers of cases. The World Bank's report on *Reducing Poverty on a Global Scale: Learning and Innovating for Development* draws on more than 100 case studies of poverty reduction worldwide. World Bank analysts identified the main factors that help or hurt in reducing poverty at scale. A whole chapter of the report assesses China's experiences in promoting economic growth and reducing poverty since over the past 25 years, noting that China has achieved the most rapid large-scale poverty reduction in human history (World Bank 2006).

In a report published by the Knowledge for Development (K4D) Program of The World Bank Institute, Zeng (2006) studied *Knowledge, Technology and Cluster-based Growth in Africa* by synthesizing findings from 11 case studies of enterprise clusters in Africa.

> These clusters are able to survive and succeed, mainly because they are able to upgrade their business activities towards more diversity and sophistication and reach a certain scale, through building up a supply-production-distribution value chain, acquiring knowledge and technology (both domestic and foreign) and disseminating and adapting them, building a relatively educated labor force, achieving collective efficiency through joint actions and cooperation, gaining government and institutional support as well as international support (such as EU, World Bank and UN) in some cases. (Pp. 8–9)

"Theory-driven evaluation" is an approach to evaluation that places a priority on testing and contributing to social science theory (Chen 1990, 1989; Chen and Rossi 1987). While theory-driven evaluations can provide program models for summative judgment or ongoing improvement, the connection to social science theory tends to focus on increasing knowledge about how effective programs work in general. Shadish (1987), in this vein, has argued that the understandings gleaned from evaluations ought to contribute to "macrotheories" about "how to produce important social change" (p. 94). Such knowledge-generating efforts focus beyond the effectiveness of a particular program to future program designs and policy formulation in general.

Synthesis evaluations also help us generate knowledge about conducting useful evaluations. The premises of utilization-focused evaluation featured in this book originally emerged from studying 20 federal evaluations (Patton et al. 1977). Those premises were affirmed by Alkin et al. (1979) in the model of evaluation use they developed by analyzing evaluations from different education districts in California and by Wargo (1989) in his "characteristics of successful program evaluations" identified by studying three "unusually successful evaluations of national food and nutrition programs" (p. 71). Alkin, Hofstetter, and Ai (1998, pp. 109–11) identified "Lessons Learned from Stakeholder Approaches" based on their review of research and theory. The Council on Foundations commissioned a synthesis evaluation based on nine case studies of major foundation evaluations to learn lessons about "effective evaluating." (A summary of one of those case studies is presented as Exhibit 4.4 in this chapter.) Among the Council's 35 key lessons learned is this utilization-focused evaluation premise: "Key 6. Make sure the people who can make the most use of the evaluation are involved as stakeholders in planning and carrying out the evaluation" (Council on Foundations 1993:255). Carlsson et al. (1999) studied evaluation use in nine Swedish development project evaluations and concluded, among other lessons, that in developing countries where oral communications are especially important, overreliance on written evaluation reports reduces use and broad dissemination of findings.

Knowledge Generation and High-Quality Lessons Learned

> *Do not be proud of your knowledge. Listen to the ignorant and the wise. Truth may lie as hidden in the earth as copper, or it may be found at play upon the lips of maidens bent above their grindstones.*
>
> Ptah-hotep, Egyptian teacher,
> 2540 BCE

As the knowledge-generating purpose of evaluation has become more prominent, it has become common practice for evaluation reports to include a section on "lessons learned." A common problem when some idea becomes highly popular, in this case, the search for lessons learned, is that the idea loses its substance and meaning. Ricardo Millett, former Director of Evaluation at the W.K. Kellogg Foundation, and I reviewed together the kinds of "lessons learned" that were offered in cluster evaluation reports. We found that the items included under these umbrella labels were so broad and inclusive that the phrases lacked any consistent meaning. As these phrases became widely used, they began to be applied to any kind of insight, evidentially based or not. We began thinking about what would constitute a "high-quality lessons learned" and decided that one's confidence in the transferability or extrapolated relevance of a supposed lesson learned would increase to the extent that it was supported by multiple sources and types of learning. Exhibit 4.9 presents a list of kinds of evidence that could be accumulated to support a proposed lesson learned, making it more worthy of application and adaptation to new settings if it has independent triangulated support from a variety of perspectives. Questions for generating "lessons learned" are also listed.

High-quality lessons learned, then, represent principles extrapolated from multiple sources and independently *triangulated* to increase transferability as cumulative knowledge working hypotheses that can be adapted and applied to new situations, a form of pragmatic utilitarian generalizability, if you will. The internal validity of any single source of knowledge would need to be judged in terms of the criteria appropriate for that type of knowledge. Thus, practitioner wisdom and evaluation studies may be internally validated in different ways. However, when these various types and sources of knowledge cohere, triangulate, and reinforce each other, that very coalescence increases the likelihood of generalizability, perhaps sufficient to justify designation as a *triangulated better practice*, or a *high-quality lesson learned*.

EXHIBIT 4.9

High-Quality Lessons Learned

High-quality lessons learned: triangulated knowledge confirmed from multiple sources that can be applied to inform future action.

Sources for triangulation

1. evaluation findings—patterns across programs
2. basic and applied research findings
3. triangulation of multiple and mixed methods
4. practice wisdom and experience of practitioners
5. experiences reported by program participants/clients/intended beneficiaries
6. expert opinion
7. cross-disciplinary findings and patterns
8. theory as an explanation of the lesson and its mechanism of impact

Assessment criteria

- assessment of the importance of the lesson learned
- strength of the evidence connecting an intervention lesson to desired outcomes attainment
- consistency of findings across sources, methods, and types of evidence

The idea is that the greater the number of supporting sources for a "lesson learned," the more rigorous the supporting evidence, and the greater the *triangulation of supporting sources*, the more confidence one has in the significance and meaningfulness of a lesson learned. Lessons learned with only one type of supporting evidence would be considered a "lessons learned hypothesis." Nested within and cross-referenced to lessons learned should be the actual cases from which practice wisdom and evaluation findings have been drawn. A critical principle here is to maintain the contextual frame for lessons learned, that is, to keep lessons learned grounded in their context. For ongoing learning, the trick is to follow future supposed applications of lessons learned to test their wisdom and relevance over time in action in new settings.

Discussion Questions for Generating High-Quality Lessons Learned

1. What is meant by a "*lesson*"?
2. What is meant by "*learned*"?
3. By whom was the lesson learned?
4. What's the evidence supporting each lesson?
5. What's the evidence the lesson was learned?
6. What are the contextual boundaries around the lesson (i.e., under what conditions does it apply)?
7. Is the lesson specific, substantive and meaningful enough to guide practice in some concrete way?
8. Who else is likely to care about this lesson?
9. What evidence will they want to see?
10. How does this lesson connect with other "lessons"?

Boruch and Petrosino (2004) have observed that

> part of the value of high-end systematic reviews, meta-analyses, and research and syntheses lies in determining where good evidence has been produced on the effects of interventions, where good evidence is absent, and where the evidence is ambiguous—respectively, the dry land, the water, and swamp.(P. 178)

They cited as an example examining hundreds of evaluations of *Scared Straight* evaluations, a program aimed at reducing juvenile delinquency by, among other interventions, scaring young people about what prison life is like. They found that most evaluations concluded that the program successfully reduced delinquent behavior, but most, in their judgment, were also not well designed.

> The authors discovered some dry land by focusing on randomized trials in this assemblage of studies. They found clear evidence that such programs have no discernible positive effect and in some cases even increase the likelihood that you will commit crime. That is, the programs effects are negative despite claims, based on untrustworthy evaluations, to the contrary. (P. 178)

They go on to note that “the value of some systematic reviews lies in establishing that no high-quality evaluations have been carried out on a particular topic.” They provide detailed guidance for conducting high-quality syntheses.

One of the challenges facing the profession of evaluation going forward will be to bring some degree of rigor to such popular notions as “lessons learned” and “best practices.” Such rigor takes on added importance as, increasingly, the substantive contribution of evaluation includes not only how to conduct high-quality evaluations but also generating knowledge based on having learned how to synthesize cross-program findings about patterns of effective interventions, that is, better practices in program design and lessons learned about effective programming generally. The future status and utility of evaluation may depend on the rigor and integrity we bring to these challenges. In the meantime, a little humility might be in order, as we proffer lessons learned.

Developmental Evaluation

> *Whosoever desires constant change must change his conduct with the times.*
>
> —Nicolò Machiavelli (1469–1527)

The last of the six purposes that can affect intended uses of evaluation is program and organizational development. Improvement-oriented, formative evaluation focuses on making an intervention or model better. Developmental evaluation, in contrast, involves changing the intervention, adapting it to changed circumstances, and altering tactics based on emergent conditions. Developmental evaluation is designed to be congruent with and nurture developmental, emergent, innovative, and transformative processes.

Summative judgment about a stable and fixed program intervention is traditionally the ultimate purpose of evaluation. Summative evaluation makes an overall judgment of merit or worth based on efficient goal attainment, replicability, clarity of causal specificity, and generalizability. *None of these traditional criteria are appropriate or even meaningful for highly volatile environments, systems-change-oriented interventions, and emergent social innovations*. Developmentally oriented leaders in organizations and programs don't expect (or even want) to reach the state of "stabilization" required for summative evaluation. Staff in such efforts doesn't aim for a steady state of programming because they're constantly tinkering as participants, conditions, learnings, and context change. They don't aspire to arrive at a fixed model that can be generalized and disseminated. At most, they may discover and articulate principles of intervention and development, but not a replicable model that says, "Do X and you'll get Y." Rather, they aspire to continuous progress, ongoing adaptation, and rapid responsiveness. No sooner do they articulate and clarify some aspect of the process than that very awareness becomes an intervention and acts to change what they do. They don't value traditional characteristics of summative excellence such as standardization of inputs, consistency of treatment, uniformity of outcomes, and clarity of causal linkages. They assume a world of multiple causes, diversity of outcomes, inconsistency of interventions, interactive effects at every level—and they find such a world exciting and desirable. They never expect to conduct a summative evaluation because they don't expect the change initiative—or world—to hold still long enough for summative review. They expect to be forever developing and changing—and they want an evaluation approach that supports development and change. That approach is developmental evaluation.

Moreover, they don't conceive of development and change as necessarily improvements. In addition to the connotation that formative evaluation (improvement-oriented evaluation) is ultimately meant to lead to summative evaluation (Scriven, 1991a, 1991b), formative evaluation carries a bias about making something better rather than making it different. From a developmental perspective informed by complexity science and systems thinking, you do something different because something has changed—your understanding, the characteristics of participants, technology, or the world. Those changes are dictated by your latest understandings and perceptions, but the commitment to change doesn't carry a judgment that what was done before was inadequate or less effective. Change is not necessarily progress. Change is adaptation. Assessing the cold reality of change, social innovators can be heard to say:

> At each stage we did the best we could with what we knew and the resources we had. Now we're at a different place in our development—doing and thinking different things. *That's development.* That's change. That's more than just making a few improvements. (Jean Gornick, former Director of Damiano, a not-for-profit working on poverty alleviation in Duluth, Minnesota; quoted in Westley, Zimmerman, and Patton 2006:179)

Developmental evaluation combines findings use with process use, the focus of the next chapter, so we will continue of discussion of it there. Chapter 8, on alternative ways of engaging in evaluation, will provide several examples of developmental evaluations. I have introduced it in this chapter to include it on the menu of alternative purposes for using findings, all six of which are summarized in Menu 4.1. For each distinct purpose, this menu shows the priority questions asked, common evaluation approaches associated with that purpose, and key factors affecting evaluation use. Menu 4.2 identifies the primary intended users and political stakes for each purpose.

MENU 4.1

Primary Uses of Evaluation Findings

Purpose	*Priority Questions*	*Common Evaluation Approaches*	*Key Factors Affecting Use*
Judgment of *overall* value to inform and support major decision making: Determine the value and future of the program and model.	Does the program meet participants' needs? To what extent does the program have merit? Worth? Does it add value for money? How do outcomes and costs compare with other options? To what extent can outcomes be attributed to the intervention? Is the program theory clear and supported by findings? Is this an especially effective practice that should be funded and disseminated as a model program?	–Summative evaluation –Impact evaluation –Cost-benefit analysis –Theory-driven evaluation	Independence and credibility of the evaluator. Rigor of the design: validity, generalizability. Significance of the findings to decision makers. Timeliness.
Learning: Improve the program.	What works and what doesn't? Strengths and weaknesses? Participant reactions? How do different subgroups respond, that is, what works for whom in what ways and under what conditions? How can outcomes and impacts be increased? How can costs be reduced? How can quality be enhanced?	–Formative evaluation –Quality enhancement –Learning reviews –Reflective practice –Participant feedback –Capacity building –Appreciative inquiry	Creating a learning climate, openness to feedback and change. Trust. Evaluator's skill in facilitating learning. Relevance of findings; actionable.
Accountability: Demonstrate that resources are well-managed and efficiently attain desired results.	Are funds being used for intended purposes? Are goals and targets being met? Are indicators showing improvement? Are resources being efficiently allocated? Are problems being handled? Are staff qualified? Are only eligible participants being accepted into the program? Is implementation following the approved plan? Are quality control mechanisms in place and being used?	–Government and funder mandated reporting –Program audits and inspections –Performance measurement and monitoring –Accreditation and licensing –End of project reports –Scorecards	Transparency. Validity of indicators. Integrity and credibility of the system and those reporting. Balance. Consistency of reporting. Fairness of comparisons.

(Continued)

MENU 4.1 (Continued)

Purpose	*Priority Questions*	*Common Evaluation Approaches*	*Key Factors Affecting Use*
Monitoring: Manage the program, routine reporting, early identification of problems.	Are inputs and processes flowing smoothly? What are participation and drop-out rates? Are these changing? Are outputs being produced as anticipated and scheduled? Where are bottlenecks occurring? What are variations across subgroups or sites?	–Management information systems –Quality control systems and CQI (continuous quality improvement) –Routine reporting and record keeping –performance indicators	Timeliness, regularity, relevance, and consistency of reporting; incentives to input data at field levels and incentives to use the data at management levels; capacity and resources to maintain the system. Appropriate links to accountability system.
Development: Adaptation in complex, emergent, and dynamic conditions.	~~What's happening at the interface between what the program is doing/accomplishing and what's going on the larger world around it? How is the program as an intervention system connected to and affected by larger systems in its environment? What are the trends in those larger systems? What does feedback show about progress in desired directions? What can we control and not control, predict and not predict, measure and not measure, and how do we respond and adapt to what we cannot control, predict, or measure? How do we distinguish signal from noise to determine what to attend to?~~	–Developmental evaluation –Complexity systems –Emergent evaluation –Real-time evaluation –Rapid assessment, rapid feedback –Environmental scanning	Openness. Adaptive capacity. Tolerance for ambiguity and uncertainty ("getting to maybe"). Balancing quality and speed of feedback. Nimble. Integrate and synthesize multiple and conflicting data sources.
Knowledge generation: Enhance general understandings and identify generic principles about effectiveness.	What are general patterns and principles of effectiveness across programs, projects, and sites? What lessons are being learned? How do evaluation findings *triangulate* with research results, social science theory, expert opinion, practitioner wisdom, and participant feedback? What principles can be extracted across results to inform practice?	–Cluster evaluation –Meta-analyses –Synthesis evaluation –Lessons learned –Effective practices studies	Quality and comparability of sources used; quality of synthesis; capacity to extrapolate. Rigor of triangulation. Identifying principles that can inform practice.

NOTE: Menu 5.1 (Chapter 5) presents a corresponding menu, "Uses of Evaluation Logic and Processes," where the impact on the program comes primarily from application of evaluation thinking and engaging in an evaluation process in contrast to impacts that come from using the content of evaluation findings, the focus of this menu.

MENU 4.2

Evaluation Purpose	*Primary Intended Users*	*What's at Stake?*
Overall Summative Judgment	Funders; those charged with making major decisions about the program's future (e.g., a board of directors); policymakers; those interested in adopting the model.	*Very high stakes*—the future of the program can be at stake, though evaluation findings are rarely the only or even primary basis for such decisions.
Formative Improvement and Learning	Program administrators, staff, and participants; those immediately involved day-to-day in the program.	*Moderate stakes*—make adjustments, act on participant feedback; enhance implementation and outcomes. Small changes involve low stakes; major improvements increase the stakes.
Accountability	Those with executive, managerial, legislative, and funding authority and responsibility to make sure that scarce resources are well-managed.	*High stakes*—the more visible the program, the more political the environment, and the more controversial the intervention, the higher the stakes.
Monitoring	Program managers as primary for a management information system: internal accountability as the priority.	*Low stakes*—ongoing, routine management, alert for bottlenecks and blips in indicators that require attention. *Becomes high stakes* when used for external accountability.
Developmental	Social innovators: those involved in bringing about major systems change in dynamic environments.	*Low stakes day-to-day* as tactical, incremental changes are made; *high stakes longer term* and strategically because social innovators aspire to have major impacts.
Knowledge generating	Program designers, planners, modelers, theorists, scholars, and policymakers.	*Moderate to low stakes*—knowledge is accumulated incrementally and cumulatively over time; no single study carries great weight; lessons learned are often principles to inform general practice and design rather than concrete recommendations to be implemented immediately.

Applying Purpose and Use Distinctions

By definition, the six different purposes we've examined—making summative judgments, offering formative improvements, accountability reporting, monitoring systems, generating generic knowledge, and developmental evaluation—can be distinguished fairly clearly. In practice, these purposes can become interrelated, parallel, and simultaneous processes as when internal government evaluators are engaged in ongoing monitoring while also preparing periodic summative reports for annual budget decisions. Or internal evaluators may be working on formative evaluation while external evaluators are conducting a summative evaluation. Many such combinations occur in real-world practice, some of them appropriate, but some of them entangling and confusing what should be distinct purposes, and those entanglements and confusions can affect use. Let me illustrate with an evaluation of an innovative educational program.

Some years ago, the Northwest Regional Educational Laboratory contracted with the Hawaii State Department of Education to evaluate Hawaii's experimental "3-on -2 Program," a team teaching approach in which three teachers worked with two regular classrooms of primary-age children, often in multiage groupings. Walls between classrooms were removed so that three teachers and 40 to 60 children shared one large space. The program was aimed at greater individualization, increased cooperation among teachers, and making more diverse resources available to students.

The Northwest Lab proposed an advocacy-adversary model for summative evaluation (Northwest Regional Educational Laboratory, 1977). Two teams were created; by coin toss one was designated the advocacy, the other the adversary team. The task of the advocacy team was to gather and present data supporting the proposition that Hawaii's 3-on -2 Program was effective and ought to be continued. The adversaries were charged with marshalling all possible evidence demonstrating that the program ought to be terminated.

The advocacy-adversary model was a combination debate/courtroom approach to evaluation (Wolf 1975; Kourilsky 1974; Owens 1973). I became involved as a resource consultant on fieldwork as the two teams were about to begin site visits to observe classrooms. When I arrived on the scene, I immediately felt the exhilaration of the competition. I wrote in my journal,

> *No longer staid academic scholars, these are athletes in a contest that will reveal who is best; these are lawyers prepared to use whatever means necessary to win their case. The teams have become openly secretive about their respective strategies. These are experienced evaluators engaged in a battle not only of data, but also of wits. The prospects are intriguing.*

As the two teams prepared their final reports, a concern emerged among some about the narrow focus of the evaluation. The summative question concerned whether the Hawaii 3-on -2 program should be continued or terminated. Some team members also wanted to offer findings about how to change the program or how to make it better without terminating it. Was it possible that a great amount of time, effort, and money was directed at answering the wrong question? Two participating evaluators summarized the dilemma in their published *post mortem* of the project:

> As we became more and more conversant with the intricacies, both educational and

> political, of the Hawaii 3-on -2 Program, we realized that Hawaii's decision-makers should not be forced to deal with a simple save-it-or-scrap-it choice. Middle ground positions were more sensible. Half-way measures, in this instance, probably made more sense. But there we were, obliged to do battle with our adversary colleagues on the unembellished question of whether to maintain or terminate the 3-on -2 Program (Popham and Carlson 1977:5).

In the course of doing fieldwork, the evaluators had encountered many stakeholders who favored a formative evaluation purpose. These potential users wanted an assessment of strengths and weaknesses with ideas for improvement. Many doubted that the program, given its popularity, could be terminated. They recognized that changes were needed, especially cost reductions, but that fell in the realm of formative not summative evaluation. I had a conversation with one educational policymaker that highlighted the dilemma about appropriate focus. He emphasized that, with a high rate of inflation, a declining school-age population, and reduced federal aid, the program was too expensive to maintain. "That makes it sound like you've already made the decision to terminate the program before the evaluation is completed," I suggested.

"Oh, no!" he protested. "All we've decided is that the program has to be changed. In some schools the program has been very successful and effective. Teachers like it; parents want it; principals support it. How could we terminate such a program? But in other schools it hasn't worked very well. The two-classroom space has been re-divided into what is essentially three self-contained classrooms. We know that. It's the kind of program that has some strong political opposition and some strong political support. So there's no question of terminating the program and no question of keeping it the same."

I felt compelled to point out that the evaluation was focused entirely on whether the program should be continued or terminated. "And that will be very interesting," he agreed. "But afterwards we trust you will give us answers to our practical questions, like how to reduce the size of the program, make it more cost-effective, and increase its overall quality."

Despite such formative concerns from some stakeholders, the evaluation proceeded as originally planned with the focus on the summative evaluation question. But was that the right focus? The evaluation proposal clearly identified the primary intended users as state legislators, members of the State Board of Education, and the superintendent. In a follow-up survey of those education officials (Wright and Sachse 1977), most reported that they got the information they wanted. But the most important evidence that the evaluation focused on the right question came from actions taken following the evaluation when the decision makers decided to eliminate the program.

After it was all over, I had occasion to ask the director of the evaluation whether a shift to a formative focus would have been appropriate. He replied,

> We maintained attention to the information needs of the *true* decision makers, and adhered to those needs in the face of occasional counter positions by other evaluation audiences. . . . If a lesson is to be learned it is this: an evaluator must determine who is making the decisions and keep the information needed by the decision makers as the highest priority. In the case of the Hawaii "3 on 2" evaluation, the presentation of program improvement information would have served to muddle the decision making process. (Nafziger 1979, personal communication)

Choosing among Alternatives

As the Hawaii case illustrates, the formative-summative distinction can be critical. Formative and summative evaluations involve significantly different data collection foci. The same data seldom serve both purposes well. Nor will either a specific formative or summative evaluation necessarily yield generic knowledge (lessons learned) that can be applied to effective programming more generally. It is thus important to identify the primary purpose of the evaluation at the outset: overall judgment of merit or worth, ongoing improvement, or knowledge generation? Is a management information system and/or accountability reporting needed? Is the program poised for significant development in adapting to changed conditions rather than improving within a predetermined and fixed model framework? Decisions about what to do in the evaluation can then be made in accordance with how best to support the evaluation's primary purpose. But this is easier said than done. One frequent reaction to posing alternatives is, "We want to do it all." A comprehensive evaluation, conducted over time and at different levels, may include variations on all six purposes, but for any given evaluation activity, or any particular stage of evaluation, it's critical to have clarity about the priority use of findings.

Consider the evaluation of a leadership program run by a private philanthropic foundation. The original evaluation contract called for 3 years of formative evaluation followed by 2 years of summative evaluation. The program staff and evaluators agreed that the formative evaluation would be for staff and participant use; however, the summative evaluation would be addressed to the foundation's board of directors. The formative evaluation helped shape the curriculum, brought focus to intended outcomes, and became the basis for the redesign of follow-up activities and workshops. As time came to make the transition from formative to summative evaluation, the foundation's president got cold feet about having the evaluators meet directly with the board of directors. The evaluators insisted on interacting directly with these primary users to lay the groundwork for genuinely summative decision making. Senior staff decided that no summative decision was imminent, so the evaluation continued in a formative mode and the design was changed accordingly. As a matter of ethics, the evaluators made sure that the chair of the board was involved in these negotiations and that the board agreed to the change in focus. There really was no summative decision on the horizon because the foundation had a long-term commitment to the leadership program. However, the program was facing some major new challenges in dealing with a large influx of immigrants in the area it served, and with major economic and political changes that affected the training leaders needed. Thus, the program moved from formative to developmental evaluation to create a substantially new approach based on changing conditions.

Now, consider a different case, the evaluation of an innovative school, the Saturn School, in Saint Paul, Minnesota. Again, the original evaluation design called for 3 years of formative evaluation followed by 2, final years with a summative focus. The formative evaluation revealed some implementation and outcome problems, including lower-than-desired scores on district-mandated standardized tests. The formative evaluation report, meant only for internal discussion to support program improvement, got into the newspapers with glaring headlines about problems and low test scores. The evaluation's visibility and public reporting put pressure on senior district officials to make summative decisions about the program despite earlier assurances that the program would have a full 5 years before such decisions were made. The formative evaluation essentially became summative when it hit the newspapers and district decision makers felt a need to make major decisions to show they were on top of things (accountability thus coming to the fore). Much to the chagrin of staff and program supporters, including many parents, the shift in purpose led to personnel changes and top-down, forced program changes. Many of those involved in openly and honestly sharing concerns in what they thought was an internal, formative process felt betrayed by the changed use from formative to summative, with heavy accountability overtones.

Sometimes, however, program staff like such a reversal of intended use as when, for example, evaluators produce a formative report that is largely positive and staff want to disseminate the results as if they were summative, even though the methods of the formative evaluation were aimed only at capturing initial perceptions of program progress, not at rendering an overall judgment of merit or worth. Keeping formative evaluations formative, and summative evaluations summative, is an ongoing challenge, not a one-time decision. When contextual conditions merit or mandate a shift in focus, evaluators need to work with intended users to fully understand the consequences of such a change. We'll discuss these issues again in the chapter on situational responsiveness and evaluator roles.

A knowledge-generating evaluation can also experience tugs and pulls into other purposes. A national foundation funded a cluster evaluation in which a team of evaluators would assemble data from some 30 different projects and identify lessons for effective community-based health programming—essentially a knowledge-generating evaluation. The cluster evaluation team had no responsibility to gather data to improve specific programs or make summative judgments. Each separate project had its own evaluation for those purposes. The cluster evaluation was intended to look for patterns of effectiveness (and barriers to same) across projects. Yet during site visits, individual projects provided cluster evaluators with a great deal of formative feedback that they wanted communicated to the foundation, and individual grantees were hungry for feedback and comparative insights about how well they were doing and ways they might improve. As the evaluation approached time for a final report, senior foundation officials and trustees asked for summative conclusions about the overall effectiveness of the entire program area as part of rethinking funding priorities and strategies. They also asked the evaluators to design a routine reporting and monitoring system for the cluster grantees. Thus, a knowledge-generating evaluation got caught up in pressures to adapt to meet demands for formative, summative, and monitoring uses.

In results-oriented M & E systems, the relationship of monitoring to evaluation is often ambiguous. Rist (2006a, 2006b) argues that we are moving from "studies to streams," by which he means that organizations are increasingly relying on systems not individual evaluators to produce evaluative knowledge. Episodic and stand-alone evaluations, which dominated the early days of the profession, are becoming a thing of the past, he argues. He sees monitoring and evaluation as merging as evaluations increasingly integrate multiple streams of information, using information produced by nonevaluators, and drawing on databases that are continuous and virtual. With managers faced with time frames that are immediate, analysis is continuous, and data collection goes on at multiple levels by multiple stakeholders. He foresees partnerships being dominant in collecting, analyzing, and sharing evaluative knowledge (rather than evaluators acting alone and controlling the evaluation process) and the Internet becoming the new information glue in support of increased transparency of evaluative knowledge. M & E can then support continuous organizational adaptation and improvement (Rist and Stames 2006). In this vision of M & E, monitoring systems will generate evaluation questions which, as they are answered with specific inquires, will feed back into and improve monitoring, yielding a continuous cycle of improvements, the results of which can be documented to meet accountability needs and demands. It's an inspiring vision. Thus far, as I read the evidence and listen to evaluators describe their experiences from around the world, it's a vision that is far from being realized. More often, as soon as accountability mandates are introduced, and they're introduced early and authoritatively, the tail wags the dog, and everyone focuses on meeting accountability demands, effectively undercutting the learning and improvement agenda, and limiting managerial willingness and capability to take risks that might attract opposition or resistance. It's not enough to create results-oriented monitoring systems. An organizational culture and climate must be created to support the appropriate and effective use of such systems. That gets us into one form of process use, organizational development, which is the focus of the next chapter.

Evaluation Use and Decision Making

We began this chapter by noting that early in the emergence of the profession, evaluators aspired to have their findings used to inform decision making. While the development of the profession has yielded more—and more nuanced—distinctions about types of evaluation uses and the alternative purposes they serve, the aspiration to inform and influence decisions remains alluring. To find out whether such potential use might be realistic, evaluators need to push intended users to be clear about what, if any, decisions are expected to be influenced by an evaluation. It is worth repeating that none of the federal health decision makers we interviewed about evaluation use, the results of which were reported at the beginning of this chapter, had been involved in a utilization-focused process. That is, none of them had carefully considered how the evaluation would be used in advance of data collection. My experiences in pushing decision makers and intended users to be more intentional and prescient about evaluation use *during the design phase* have taught me that it is possible to significantly increase the degree of influence evaluations have. Doing so, however, requires persistence in asking the following kinds of questions: What decisions, if any, is the evaluation expected to influence? What is at stake? When will decisions be made? By whom? What other factors (values, politics, personalities, promises already made) will affect the decision making? How much influence do you expect the evaluation to have? What needs to be done to achieve that level of influence? How will we know afterward if the evaluation was used as intended? (In effect, how can use be measured?) Exhibit 4.10 highlights questions to use in determining an evaluation's potential for concrete and specific instrumental use in informing decision making.

EXHIBIT 4.10

Questions to Ask of Intended Users to Establish an Evaluation's Intended Influence on Forthcoming Decisions

What decisions, if any, are the evaluation findings expected to influence?

(There may not be any, in which case the evaluation's purpose may be simply to generate knowledge for conceptual use and future enlightenment. If, however, the evaluation is expected to influence decisions, clearly distinguish summative decisions about program funding, continuation or expansion from formative decisions about program improvement, and ongoing development.)

When will decisions be made? By whom? When, then, must the evaluation findings be presented to be timely and influential?

What is at stake in the decisions? For whom? What controversies or issues surround the decisions?

What's the history and context of the decision-making process?

What other factors (values, politics, personalities, promises already made) will affect the decision making? What might happen to make the decision irrelevant or keep it from being made? In other words, how volatile is the decision-making environment?

How much influence do you expect the evaluation to have—*realistically*?

To what extent has the outcome of the decision already been determined?

What data and findings are needed to support decision making?

What needs to be done to achieve that level of influence?

(Include special attention to which stakeholders to involve for the evaluation to have the expected degree of influence.)

How will we know afterwards if the evaluation was used as intended?

(In effect, how can use be measured?)

Making Menu Selections: Connecting Decisions to Uses

Where the answers to the evaluator's questions indicate that a major decision about program merit, worth, continuation, expansion, dissemination, and/or funding is at stake, then the evaluation should be designed to render overall judgment—summative judgment. The design should be sufficiently rigorous and the data collected should be sufficiently credible that a summative decision can be made. The findings must be available in time to influence this kind of major decision.

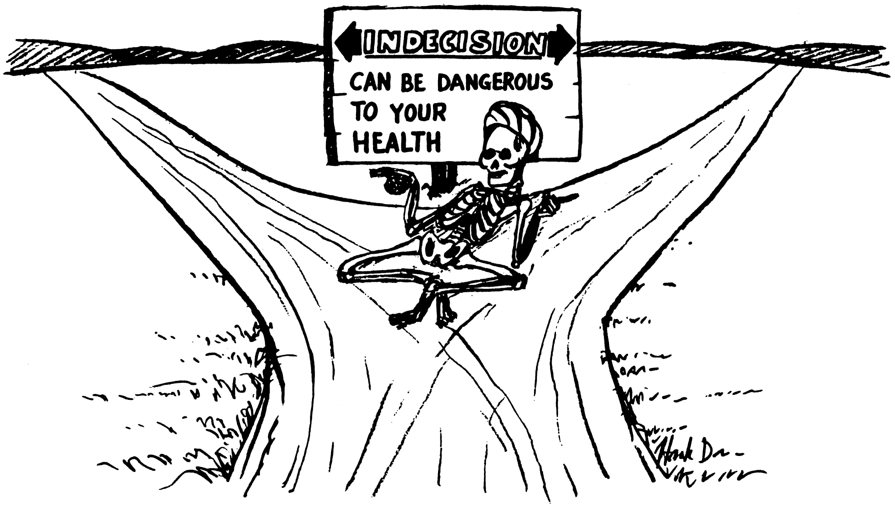
Where the dialogue with primary intended users indicates an interest in identifying strengths and weaknesses, clarifying the program's model, and generally working at increased effectiveness, the evaluation should be framed to support improvement-oriented decision making. Skills in offering formative feedback and creating an environment of mutual respect and trust between the evaluator and staff will be as important as actual findings.

Where the intended users are more concerned about generating knowledge for formulating future programs than with making decisions about current programs, then some form of synthesis or cluster evaluation will be most appropriate to discover generic principles of effectiveness.

Likewise, the evaluator can review accountability concerns, the potential role of a monitoring system, and the degree of interest in developmental evaluation.

The six options I've presented are by no means inherently conflicting purposes, and some evaluations strive to incorporate aspects of different approaches, as in M & E. But in my experience, one purpose is likely to become the dominant motif and prevail as the *primary* purpose informing design decisions and priority uses; or else, different aspects of an evaluation are designed, compartmentalized, and sequenced to address these contrasting purposes. I also find that confusion among these quite different purposes, *or failure to prioritize them*, is often the source of problems and misunderstandings along the way, and can become disastrous at the end when it turns out that different intended users had different expectations and priorities.

In helping intended users select from the evaluation purposes menu, and thereby focus the evaluation, evaluators may encounter some reluctance to make a commitment. I worked with one director who proudly displayed this sign on his desk: "My decision is maybe—and that's final." Unfortunately, the sign was all too accurate. He wanted me to decide what kind of evaluation should be done. After several frustrating attempts to narrow the evaluation's focus, I presented what I titled a "MAYBE DESIGN." I laid out cost estimates for an all-encompassing evaluation that included formative, summative, knowledge-generating, accountability, monitoring, and developmental components looking at all aspects of the program. Putting dollars and timelines to the choices expedited the decision making considerably. He decided not to undertake any evaluation "at this time."



I was relieved. I had become skeptical about the potential for doing anything useful. Had I succumbed to the temptation to become the decision maker, an evaluation would have been done, but it would have been my evaluation, not his. I'm convinced he would have waffled over using the findings as he waffled over deciding what kind of evaluation to do.

Thus, in utilization-focused evaluation, the choice of not dining at all is always on the menu. It's better to find out before preparing the meal that those invited to the banquet are not really hungry. Take your feast elsewhere, where it will be savored.

Follow-Up Exercises

1. Identify an actual program. Describe the program and its context. Specify the specific primary evaluation questions that would guide an evaluation endeavor under each of the six purposes in Menu 4.1.

2. For the program identified in Question 1, or another program, use Menu 4.2 to identify the specific intended users by name and position in the center column, "primary intended users." Then, assess the stakes (Column 3) for those intended users. How do the stakes for the primary intended users you've identified compare with the norms described in Column 3 of Menu 4.2?

3. Search the news, the Internet, evaluation journals, and other sources to find (a) an example of instrumental use of an evaluation, (b) an example of conceptual use of an evaluation, and (c) example of persuasive use. Describe each use and its context. To what extent and in what ways do you consider the use as appropriate and meaningful? Explain the basis for your judgments.

4. Search the news, the Internet, evaluation journals, and other sources to find what you consider an example of misuse of an evaluation. Describe the misuse, the context, and the consequences. What, if anything, might have been done, in your judgment, to prevent or reduce the misuse?

5. Use Exhibit 4.11, Questions to Ask of Intended Users to Establish an Evaluation's Intended Influence on Forthcoming Decisions, to interact with a real-life program manager. Approach that manager as a simulation of a real evaluation consultation in which you will be assisting in designing a decision-oriented evaluation. Record the highlights of the interaction and comment on what it reveals about decision-oriented, instrumental use.

6. Conduct your own utilization study. Use the inquiry questions on use at the beginning of this chapter. Identify at least two different evaluations to follow up. Interview both a program person and the evaluator, if possible. Find out how those evaluations were used. Compare your findings with those presented from the federal health evaluation and more recent distinctions of types and degrees of use.