1

Evaluation Use

Both Challenge and Mandate

The human condition: insidious prejudice, stultifying fear of the unknown, contagious avoidance, beguiling distortion of reality, awesomely selective perception, stupefying self-deception, profane rationalization, massive avoidance of truth—all marvels of evolution's selection of the fittest. Evaluation is our collective effort to outwit these human propensities—when we choose to use it.

—Halcolm

On a cold November morning in Minnesota, some 15 coffee-clutching people in various states of wakefulness have gathered to discuss evaluation of a county welfare-to-work program. Citizen advisory board representatives are present; the county board and state representatives have arrived; and the internal evaluator is busy passing out handouts and setting up the PowerPoint presentation. We are assembled at this early hour to review the past year's evaluation.

The evaluator begins by reviewing the problems getting started—fuzzy program goals, uncertain funding, incomplete program files, and software inadequacies. Data collection problems included staff resistance to doing additional paperwork, difficulty finding clients for follow-up interviews, and inconsistent state and county information systems. The evaluation was further hampered by management problems, staff turnover, unclear decision-making hierarchies, political undercurrents, trying to do too much, and an impossibly short timeline for reporting. Despite the problems, the evaluation had been completed and, putting the best face on a difficult situation,

the evaluator explains that "the findings are tentative to be sure, but more than we knew a year ago."

Advisory board members are clearly disappointed. One says, "The data just aren't solid enough." A county commissioner explains why Board decisions have been contrary to evaluation recommendations: "We didn't really get the information we needed when we wanted it and it wasn't what we wanted when we got it." The room is filled with disappointment, frustration, defensiveness, cynicism, and more than a little anger. There are charges, countercharges, budget threats, moments of planning, and longer moments of explaining away problems. The chairperson ends the meeting in exasperation, lamenting: "What do we have to do to get evaluation results we can actually use?"

This book is an outgrowth of, and an answer to, that question.

Evaluation Use as a Critical Societal Issue

If the scene I have described were unique, it would merely represent a frustrating professional problem for the people involved. But if that scene is repeated over and over on many mornings, with many advisory boards, then the question of evaluation use would become what eminent sociologist C. Wright Mills called a critical public issue:

> Issues have to do with matters that transcend these local environments of the individual and the range of his inner life. They have to do with the organization of many such milieux into the institutions of an historical society as a whole. . . . An issue, in fact, often involves a crisis in institutional arrangements (Mills 1959:8–9).

In my judgment, the challenge of using evaluation in appropriate and meaningful ways represents just such a crisis in institutional arrangements. How evaluations are used affects the spending of billions of dollars to fight problems of poverty, disease, ignorance, joblessness, mental anguish, crime, hunger, and inequality. How are programs that combat these societal ills to be judged? How does one distinguish effective from ineffective programs? And how can evaluations be conducted in ways that lead to use? How do we avoid producing reports that gather dust on bookshelves, unread and unused? These are the questions this book addresses, not just in general, but within a particular framework: *utilization-focused evaluation*.

But first, what are these things called "evaluations" that we hope to see used?

Evaluation as Defined in the *Encyclopedia of Evaluation*

Evaluation is an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit, worth, significance, or quality of a program, product, person, policy, proposal, or plan. Conclusions made in evaluations encompass both an empirical aspect (that something is the case) and a normative aspect (judgment about the value of something). It is the value feature that distinguishes evaluation from other types of inquiry, such as basic science research, clinical epidemiology, investigative journalism, or public polling. (Fournier 2005a:140)

To evaluate something means determining its merit, worth, value, or significance. Program or project evaluations typically involve making the following kinds of judgments: How effective is the program? To what extent has the program been implemented as expected? Were the program's goals achieved? What outcomes and results were achieved by the program? To what extent and in what ways did program participants benefit, if at all? What needs of participants were met? What unanticipated consequences resulted from the program? What are the strengths and weaknesses of the program, and how can it be improved? What worked and what didn't work? What has been learned in this program that might be useful to other programs? To what extent do the benefits of the program provide sufficient value to justify the costs of the program? Should the program's funding be maintained as is, increased, or decreased? Evaluations, then, typically describe and assess what was intended (goals and objectives), what happened that was unintended, what was actually implemented, and what outcomes and results were achieved. The evaluator will then discuss the implications of these findings, sometimes including items for future action and recommendations. In the simplest terms, evaluations are said to answer three questions:

What?

So What?

Now What?

Sometimes these evaluation questions are answered in formal reports. Some evaluative judgments flow from analyzing and discussing data from a program's information system without producing a formal report; indeed, increasingly findings emerge as "the real-time production of streams of evaluative knowledge" (Rist 2006a:6–7; Stame 2006b:vii), rather than as discrete, stand-alone studies. Some evaluation reports are entirely internal to an organization for use by staff and administrators to support ongoing managerial decision making. Other evaluation reports are published or posted on the Internet to meet an obligation for public accountability or to share lessons learned.

The issue of evaluation use has emerged at the interface between science and action, between knowing and doing. It raises fundamental questions about human rationality, decision making, and knowledge applied to creation of a better world. And the issue is not just a concern of researchers. Sometimes it reaches the larger public as in this classic newspaper headline, "Agency Evaluation Reports Disregarded by Legislators Who Requested Them" (see Exhibit 1.1).

A Broader Perspective: Using Information in the Knowledge Age

The challenge of evaluation use epitomizes the more general challenge of knowledge use in our times. Our age—*the Age of Information, Knowledge, and Communications*—has developed the capacity to generate, store, retrieve, transmit, and instantaneously communicate massive amounts of information. Our problem is keeping up with, sorting out, absorbing, prioritizing, and *using* information. Our technological capacity for gathering and computerizing information now far exceeds our human ability to process and make sense out of it all. We're constantly faced with deciding what's worth knowing and what to ignore.

Evaluators are "knowledge workers," a term the great management scholar and consultant Peter Drucker introduced to

What? So What? Now What?

Glenda H. Eoyang, Executive Director of the Human Systems Dynamics Institute in Minnesota, describes how she uses this evaluative framework in managing her own organization.

At the Institute, we use three simple questions to help us distinguish emergencies from the merely emergent, to analyze multiple factors in the moment, and to align our diverse actions toward shared goals. These questions, though simple, are deeply powerful as we shape our work together toward adaptive action.

WHAT? What do we see? What does data tell us? What are the indicators of change or stability? What cues can we capture to see changing patterns as they emerge?

SO WHAT? So, what sense can we make of emerging data? What does it mean to us in this moment and in the future? What effect are current changes likely to have on us, our clients, our extended network, and our field of inquiry and action?

NOW WHAT? What are our options? What are our resources? When and how can we act—individually or collectively—to optimize opportunities in this moment and the next?

We and our clients have used these questions to move together toward decisive action.

A social service agency faced radical changes in public policy that would have a direct effect on their clients and the resources they had available to meet clients' needs. What? So what? Now what?

A medical technology company focused on getting processes under control and ensuring lean, high quality product development and deployment procedures. What? So what? Now what?

An organization in the midst of internal transformation faced backlash from disgruntled workers. What? So what? Now what?

A group of attorneys and their support staff recognized patterns of negative attitudes and disruptive relationships that sucked their energies and distracted them from productive work. What? So what? Now what?

In each of these cases, the three questions helped leadership focus on critical options and effective actions. What emerged was not a sophisticated and complicated plan for an unknowable future. No. What did emerge was a shared understanding of emerging challenges and clear focus on actions that could shift emergencies into emergent possibilities. (Eoyang 2006b)

SOURCE: Reprinted with permission of Glenda Eoyang.

describe anyone who produces knowledge, ideas, and information in contrast to more tangible products and services (Drucker 2003; Lowenstein 2006). The challenge we face is not just producing knowledge but the even greater challenge of getting people to use the knowledge we produce.

Getting people to use what is known has become a critical concern across the different knowledge sectors of society. A major specialty in medicine (compliance research) is dedicated to understanding why so many people don't follow their doctor's orders. Common problems of information use

EXHIBIT 1.1

Newspaper Column on Evaluation Use

Agency Evaluation Reports Disregarded by Legislators Who Had Requested Them

Minnesota lawmakers who mandated that state agencies spend a lot of employee hours and money developing performance evaluation reports pretty much ignored them. . . . The official word from the state legislative auditor's evaluation of the performance evaluation process: Legislators who asked for the reports did not pay much attention to them. They were often full of boring and insignificant details. . . .

Thousands of employee hours and one million taxpayer dollars went into writing the 21 major state agency performance evaluation reports. The auditor reports the sad results:

- Only three of 21 state commissioners thought that the performance reports helped the governor make budget choices regarding their agencies.
- Only seven of 21 agencies were satisfied with the attention given the reports in the House committees reviewing their programs and budgets. And only one agency was satisfied with the attention it received in the Senate.

Agency heads also complained to legislative committees this year that the 1993 law mandating the reports was particularly painful because departments had to prepare new two-year budget requests and program justifications at the same time. That "dual" responsibility resulted in bureaucratic paperwork factories running overtime.

"Our experience is that few, if any, legislators have actually read the valuable information contained in our report . . . ," one agency head told auditors. "The benefits of performance reporting will not materialize if one of the principal audiences is uninterested," said another.

"If the Legislature is not serious about making the report 'the key document' in the budget decision process, it serves little value outside the agency," said a third department head.

Mandating the reports and ignoring them looks like another misguided venture by the 201-member Minnesota Legislature. It is the fifth largest Legislature in the nation and during much of the early part of this year's five-month session had little to do. With time on their hands, lawmakers could have devoted more time to evaluation reports. But if the reports were dull and of little value in evaluating successes of programs, can they be blamed for not reading them?

Gary Dawson, "State Journal" column
Saint Paul Pioneer Press August 7, 1995: 4B

SOURCE: Reprinted with permission from the Saint Paul Pioneer Press.

underlie trying to get people to use seat belts, quit smoking, begin exercising, eat properly, and pay attention to evaluation findings. In the fields of nutrition, energy conservation, education, criminal justice, financial investment, human services, corporate management, public administration, philanthropy, international development—the list could go on and on—a central problem, often *the* central problem, is getting people to apply what is already known.

In agriculture, a major activity of university extension services is trying to get farmers to adopt new scientific methods.

Experienced agricultural extension agents like to tell the story of a young agent telling a farmer about the latest food production techniques. As he begins to offer advice the farmer interrupts him and says, "No sense telling me all those new ideas, young man. I'm not doing half of what I know I should be doing now."

I remember talking with a time management trainer who had done a follow-up study of people who had taken her workshop series. Few were applying the time management techniques they had learned. When she compared the graduates of her time management training with a sample of nonparticipants, the differences were not in how people in each group managed their time. The time management graduates had quickly fallen back into old habits. The difference was *the graduates felt much guiltier about how they wasted time*.

Research on adolescent pregnancy illustrates another dimension of the knowledge use problem. In a classic study, adolescent-health specialist Michael Resnik (1984) interviewed teenagers who became pregnant. He found very few cases in which the problem was a lack of information about contraception, about pregnancy, or about how to avoid pregnancies. The problem was that teens just didn't apply what they knew. "There is an incredible gap between the knowledge and the application of that knowledge. In so many instances it's heartbreaking—they have the knowledge, the awareness, and the understanding, but somehow it doesn't apply to them" (p. 15).

Sometimes the stakes are incredibly high, as in using information to prevent genocide. Between April and June 1994, an estimated 800,000 Rwandans were killed in the space of 100 days. Lieutenant General Roméo Dallaire headed the small UN Peacekeeping Force in Rwanda. He filed detailed reports about the unspeakable horrors he and his troops witnessed. He documented the geographic scope of the massacre and the numbers of people being slaughtered. In reporting these findings to the UN officials and Western governments, Dallaire pleaded for more peacekeepers and additional trucks to transport his woefully ill-equipped force. He sought authority to seize Hutu arms caches, but the narrow UN mandate didn't allow him to disarm the militias. As bodies filled the streets and rivers, the general tried in vain to attract the world's attention to what was going on. In an assessment that military experts now accept as realistic, Dallaire argued that with 5,000 well-equipped soldiers and a free hand to fight Hutu power, he could bring the genocide to a rapid halt. The United Nations, constrained by the domestic and international politics of Security Council members, ignored him. He asked the United States to block the Hutu radio transmissions that were provoking and guiding the massacre. The Clinton administration refused to do even that.

Instead, following the deaths of 10 Belgian peacekeepers assigned to protect the President of Rwanda, Dallaire's forces were cut to a mere 500 men, far too few to make a difference as one of the most horrific genocides in modern history unfolded. Dallaire, frustrated and disheartened by the passive attitude of world leaders, repeatedly confronted his superiors, trying to get them to deal with the data about what was going on, all to no avail. The international community occupied itself with arguing about the definition of genocide, placing blame elsewhere, and finding reasons not to intervene.

The highly respected Danish international development agency, Danida, sponsored a major retrospective evaluation of the Rwanda genocide seeking to extract lessons that might help the world avoid future such tragedies (Danida 2005). The United Nations (1996) undertook its own investigation and Dallaire and Beardsley (2004) have provided their own account of what happened and why. The point, for our purposes, is that the Rwanda story included the refusal of international agencies and world leaders to take seriously and use the data they were given. Those who had the responsibility and capacity to act failed to pay attention to the evidence Dallaire provided them about the deteriorating situation and the consequences of a failure to act. While his efforts involved the highest stakes possible—saving human lives—evaluators across a broad range of sectors face the daily challenge of getting decision makers to take evidence of ineffectiveness seriously and act on the implications of the evidence. It was precisely this larger relevance of the Rwanda example that led to Dalliare being invited to keynote 2,330 evaluation professionals from 55 countries at the joint Canadian Evaluation Society and the American Evaluation Association (AEA) international conference in Toronto in 2005. Following the keynote, he was awarded the Presidents' Prize for *Speaking Truth to Power.* This award symbolizes one of the most important roles evaluators can be called on to play, a role that goes beyond technical competence and methodological rigor, a role that recognizes the inherently political nature of evaluation in a world where knowledge is power—the role of speaking truth to power.

The High Stakes of Evaluation Use

When the space shuttle Columbia disintegrated on February 1, 2002, killing all seven astronauts aboard, a comprehensive independent investigation ensued by a 13-member board of inquiry. While the direct mechanical problem was damage caused by a foam tile that came loose during liftoff, the more basic cause, investigators concluded, was the National Aeronautics and Space Administration's (NASA's) own culture, a culture of complacency nurtured by a string of successes since the 1986 Challenger disaster, which also killed seven. This led to a habit of relaxing safety standards to meet financial and time constraints, for example, defining a problem as insignificant so as not to require a fix that would cause delay. The Columbia Accident Investigation Board (2003) concluded in its 248-page report that the space agency lacked effective checks and balances, did not have an independent safety program, and had not demonstrated the characteristics of a learning organization.

In addition to detailing the technical factors behind Columbia's breakup just minutes before its scheduled landing at the end of a 16-day science mission, the board's report laid out the cultural factors behind NASA's failings. It said NASA mission managers fell into the habit of accepting as normal some flaws in the shuttle system and tended to ignore, not recognize, or *not want to hear about* such problems even though they might foreshadow catastrophe. Such repeating patterns meant that flawed practices embedded in NASA's organizational system continued for years and made substantial contributions to both accidents, the report concluded. During Columbia's last mission, NASA managers missed

opportunities to evaluate possible damage to the craft's heat shield from a strike on the left wing by flying foam insulation. Such insulation strikes had occurred on previous missions and, the report said, engineers within NASA had documented the dangers involved, but the evidence they submitted and the accompanying warnings they sent up the chain-of-command were ignored. This attitude of ignoring data that led to conclusions they didn't like also contributed to the lack of interest among NASA managers in getting spy satellite photos of Columbia, images that might have identified the extent of damage on the shuttle. Over time, NASA managers had come to accept more and more risk in order to meet scheduled launch deadlines. But most of all, the report concluded, there was ineffective leadership that discouraged dissenting views on safety issues, ignored the evaluation findings of safety engineers, and ultimately created blind spots about the risk to the space shuttle of the foam insulation impact.

Sometimes, as in the examples of the Rwanda genocide and the Columbia shuttle disaster, important data are ignored. In other cases, the data-generating process itself is distorted and manipulated to create biased and distorted findings. Regardless of what one thinks of the U.S. invasion of Iraq to depose Saddam Hussein, both those who supported the war and those who opposed it have come to agree that the intelligence used to justify the invasion was deeply flawed and systematically distorted (U.S. Senate Select Committee on Intelligence 2004). Under intense political pressure to show sufficient grounds for military action, those charged with analyzing and evaluating intelligence data began doing what is sometimes called cherry-picking or stovepiping—selecting and passing on only those data that support preconceived positions and ignoring or repressing all contrary evidence (Hersh 2003). This is a problem of the *misuse of evaluation findings*, the shadow side of the challenge of increasing utility.

Getting evaluations used begins with having valid, accurate, relevant, and balanced findings *that are worth using*. Then, as the Rwanda example demonstrates, those with data have to get the attention of those who will make crucial decisions. As the NASA story shows, this is not just a matter of reaching individual decision makers but dealing with the whole culture of organizations to create a learning environment that is receptive to data-based decision making. And as the problem of Iraq prewar intelligence shows, evaluators committed to accurate and balanced reporting will have to deal with political obstacles and speak truth to power.

These are high-profile examples of evaluation neglect and misuse, but the challenges of getting evaluations used aren't limited to such obviously high stakes national and international initiatives. Look at today's local news. What decisions are being reported about programs and policies in your community—decisions by city councils, school boards, county commissions, legislative committees, not-for-profit agencies, philanthropic foundations, and businesses? What does the news story tell you about the data that informed those decisions? What evidence was used? What was the quality of that evidence? How were data generated and presented as part of the decision-making process?

Decisions abound. Policy choices are all around us. Programs aimed at solving problems exist in every sector of society. And behind every one of these decisions, policy choices, and program initiatives

is an evaluation story. What evaluative evidence, if any, was used in the decision making? What was the quality of that evidence? Asking these questions is at the foundation of *utilization-focused evaluation.*

These questions also have relevance at the personal level. Some degree of evaluative thinking is inherently involved in every decision you make. How did you decide what computer to purchase? Or what course to take? How do you and those you know make decisions about dating, marriage, nutrition, exercise, lifestyle, where to live, what to do for leisure, who to vote for (and whether to vote at all), what movie to see (and whether you thought it was any good), what books or magazines to read, when to see a doctor, and so on. We are all evaluators. But we are not all good at it, not always systematic or thoughtful, careful about seeking out and weighing evidence, and explicit about the criteria and values that underpin our interpretations of whatever evidence we have. Thus, as we consider how to enhance program decision making through systematic evaluation, you may pause now and again to consider the implications of this way of thinking for decisions you make in your personal life. Or maybe not. How will you decide?

These examples of the challenges of putting knowledge to use are meant to set a general context for the specific concern of this book: generating high-quality and highly relevant evaluation findings and then actually getting those findings used for program decision making and improvement. Although the problem of information use remains central to our age, we are not without knowledge about what to do. We've learned a few things about overcoming our human resistance to new knowledge and change, and over the past three decades of professional evaluation practice, we've learned a great deal about how to increase evaluation use. Before presenting what's been learned, let's set the context.

Historical Precedents

Today's professional evaluators stand on the shoulders of much earlier practitioners though they didn't necessarily call what they did evaluation. The emperor of China established formal proficiency testing for public officials some 4,000 years ago. The book of Daniel in the Old Testament of the *Bible* opens with the story of an educational program evaluation in which King Nebuchadnezzar of Babylon created a 3-year civil service training program for Hebrew youth after his capture of Jerusalem. When Daniel objected to eating the King's meat and wine, the program director, Melzar, agreed to an experimental comparison to evaluate how eating a kosher diet might affect the "countenance" of Daniel and his friends, Hananiah, Mishael, and Azariah. When they remained healthy after the pilot test period, he agreed to a permanent change for those who eschewed the Babylonian diet—the earliest documentation of using evaluation findings to change a program's design.

Once you start looking, you can turn up all kinds of historical precedents for evaluation.

The great Lewis and Clark expedition through the central and western American wilderness had as its purpose evaluating the suitability of the interior rivers for transportation and the value of the land for settlement. The Louisiana Purchase, which they would reconnoiter, covered more than 2 million square kilometers of land

Evaluating a Venture in Colonial America

In the 1730s, the settlement of the colony of Georgia by the English poor began as a philanthropic venture in colonial America complete with a detailed proposal ("blueprint"), grandiose goals, quite measurable outcomes, testable hypotheses about how to achieve outcomes (what we'd call today a "theory of change"), annual plans, internal evaluation reports from the staff to the Board of Trustees, independent site visits, multiple and conflicting stakeholders, divisive politics, bureaucratic ineptitude and micromanaging, pointed participant feedback about problems, a major problem with dropouts, ongoing efforts at project improvement, and ultimately, a judgment that the experiment failed, accompanied by a lofty explanation from the funders about why they were compelled to pull the plug, to wit:

"At first it was a trial, now it is an experiment; and certainly no man or society need be ashamed to own, that from unforeseen emergencies their hypothesis did misgive; and no person of judgment would censure for want of success where the proposal was probable; but all the world would exclaim against that person or society who, through mistaken notions of honor or positiveness of temper, would persist in pushing an experiment contrary to all probability, to the ruin of the adventurers." (Boorstin 1958:96)

extending from the Mississippi River to the Rocky Mountains, essentially doubling the size of the United States. They would travel all the way to the Pacific Ocean. On June 20, 1803, President Thomas Jefferson launched the exploration thusly:

> The Object of your mission is to explore the Missouri river & such principal streams of it as by its course and communication with the waters of the Pacific ocean, whether the Columbia, Oregon, Colorado or any other river may offer the most direct and practicable water communication across this continent for the purpose of commerce.

The reports Lewis and Clark sent back to Jefferson were, for all intents and purposes, evaluation reports addressing the objectives set forth and much, much more. In effect, President Jefferson needed to find out what he had purchased, what we might call a *retrospective evaluation*. The reports of Lewis and Clark went well beyond their narrow, stated mission and included extensive inventories of plants and animals, including many new species, details about indigenous peoples, maps of the land, indeed, everything they did and all that they encountered over a 3-year period through lands that later became 11 states. They might be awarded a posthumous Guinness World Record for the evaluation that most exceeded its original scope of work. The impressive Saint Louis Gateway Arch on the banks of the Mississippi River, commemorating the Westward Expansion opened up by the Lewis and Clark Expedition, could be considered a monument to the impact of evaluation findings.

You get the idea. Just because something wasn't officially labeled an evaluation report doesn't mean it didn't serve an evaluative function. Lewis and Clark gathered and reported extensive data to judge the merit, worth, significance, and value of the Louisiana Territory. Their descriptive data and judgments affected congressional policy, executive directives, and federal appropriations. If it reads like an evaluation, serves evaluative functions, and gets used like an evaluation, we might call it an evaluation.

Thomas Jefferson was also among the recipients of another evaluation report, this one well before he became president, indeed, before the American Revolution. Jefferson and other founding fathers of the United States had become knowledgeable about and impressed with the Iroquois republic, a Native American people in the Northeastern part of North America, which had continuously existed since the fourteenth or fifteenth century. The Iroquois Constitution, known as "The Great Law of Peace," was an orally transmitted constitution for the union of five (later six) Indian nations: Mohawk, Onondagam Seneca, Oneida, Cayuga, and the Tscarora. Jefferson, Benjamin Franklin, John Adams, and George Washington were all familiar with the Iroquois polity and were influenced by its key ideas and processes in conceptualizing American government (Johansen 1998; 1987; Idarius 1998). In 1774, the Virginia Colony offered the Iroquois Confederacy scholarships to send six of their young men to Williamsburg College to be educated. The Iroquois Chiefs responded that they had already had some of their young men attend such a college and their evaluation of the results did not predispose them to accept the Virginia offer, which they, in fact, declined:

> Several of our Young People were formally brought up at the Colleges of the Northern Provinces; they were instructed in all of your Sciences; but, when they returned to us, they were bad runners, ignorant of every means of living in the Woods, unable to bear either Cold or Hunger, knew neither how to build a Cabin, take a Deer, or kill an Enemy, spoke our language imperfectly, were therefore neither fit for Hunters, Warriors, nor Counsellors, they were totally good for nothing. (Hopkins 1898:240)

Now that's a clear, evidence-based, mince-no-words, evaluative judgment!

You never know where an evaluation report may turn up. Doing archival research in Tanzania, I found scores of reports by anthropologists, colonial managers, and English academics describing and assessing various and sundry failed attempts to settle the nomadic cattle-herding Wagogo people of the Dodoma Region in what had been central Tanganyika. My assignment was to wrestle lessons learned from the many failed settlement schemes spanning some 50 years of colonial rule so that the modern democratic government under President Julius Nyerere could find a new, more humane, and effective approach. None of the reports were titled "evaluations," but all of them were fundamentally evaluative.

I stumbled across an explicit but still secret evaluation visiting the Hiroshima Museum while conducting evaluation training at Hiroshima University. One of the exhibits there describes how the American military's "Target Committee" selected Hiroshima for trying out the first atomic bomb. Allied forces were engaged in heavy bombing throughout Japan, especially in and around Tokyo. Since the destructive power of the atomic bomb was unknown, a small number of major Japanese cities were excluded from routine bombing and carefully photographed with inventories of buildings, infrastructure, and industries. On August 6, 1945, the nuclear weapon *Little Boy* was dropped on Hiroshima by the *Enola Gay*, a U.S. Air Force B-29 bomber, which was altered specifically to hold the bomb, killing an estimated 80,000 people and heavily damaging 80 percent of the city. An American military team subsequently completed a full evaluation of the bomb's damage and impact but, according to the Museum display, that report has never been made

public. Whether and how it was used is also, therefore, unknown.

Using Evaluative Thinking: A Transdisciplinary Perspective

Experience in thinking can be won, like all experience in doing something, only through practice.

—Philosopher Hannah Arendt (1963:4)

The historical examples of evaluations I have just reviewed provide some sense of the long and diverse history of evaluation reporting, but even more important, they illustrate the centrality of *evaluative thinking* in human affairs and inquiries of all kinds. Using evaluative thinking and reasoning is ultimately more important and has more far-reaching implications than merely using evaluation reports. This is why eminent philosopher and evaluation theorist Michael Scriven (2005b, 2004) has characterized evaluation as a *transdiscipline*, because every discipline, profession, and field engages in some form of evaluation, the most prominent example being, perhaps, evaluations of students taking courses and completing disciplinary programs of study, and refereed journals in which new research is evaluated by peers to determine if it is worthy of publication. Evaluation is a discipline that serves other disciplines even as it is a discipline unto itself; thus its emergent transdisciplinary status (Coryn and Hattie 2006). Statistics, logic, and evaluation are examples of transdisciplines in that their methods, ways of thinking, and knowledge base are used in other areas of inquiry, e.g., education, health, social work, engineering, environmental studies, and so on (Mathison 2005:422). In studying evaluation use, then, we will be looking at not only the use of evaluation findings and reports but also what it means to use evaluative thinking.

The Emergence of Program Evaluation as a Formal Field of Professional Practice

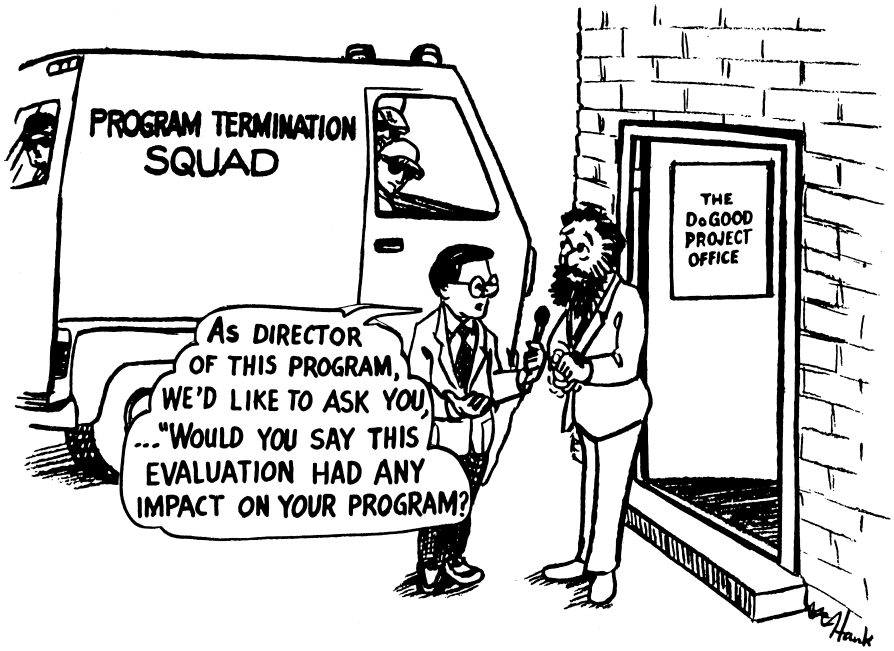
There is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things. Because the innovator has for enemies all those who have done well under the old conditions and lukewarm defenders in those who may do well under the new.

—Advice from *The Prince* (1513) Niccolo Machiavelli (1469–1527)

While evaluative thinking, inquiry, and judgments are as old as and inherent to our human species, formal and systematic evaluation as a field of professional practice is relatively recent. Like many poor people, evaluation in the United States has grown up in the "projects"—federal projects spawned by the Great Society legislation of the 1960s. When the federal government of the United States began to take a major role in alleviating poverty, hunger, and joblessness during the Depression of the 1930s, the closest thing to evaluation was the employment of a few jobless academics to write program histories. Some important evaluations began to be done after World War II, for example, an evaluation of the First Salzberg Seminar conducted by distinguished anthropologist Margaret Mead for the W. K. Kellogg Foundation in 1947 (Russon and Ryback 2003; Greene 2003; Patton 2003). In 1959, the U.S. Department of Health, Education, and Welfare published

guidelines for evaluation (Herzog 1959). Preeminent evaluation researcher and author Carol Weiss has recounted finding a number of published evaluation studies from the late 1950s and early 1960s that informed her own first evaluation effort (Weiss 2004:163). But it was not until the massive federal expenditures on an awesome assortment of programs during the 1960s and 1970s that accountability in government began to mean more than financial audits or political head counts of opponents and proponents. Demand for systematic empirical evaluation of the effectiveness of government programs grew as government programs grew (Shadish and Luellen 2005; Aucoin and Heinzman 2000; House 1993; Wye and Sonnichsen 1992). At the same time, fear of, resistance to, and a backlash against evaluation accompanied evaluation's growth as some program staff and agency managers looked on evaluation as a personal attack and feared that evaluation was merely a ruse for what was really a program termination agenda.

Educational evaluation accompanied the expansion of access to public schooling. Joseph Rue's comparative study of spelling performance by 33,000 students in 1897 was a precursor of educational evaluation, which remains dominated by achievement testing. During the cold war, after the Soviet Union launched Sputnik in 1957, calls for better educational assessments accompanied a critique born of fear that the education gap was even larger than the "missile gap." Demand for independent evaluation accelerated with the growing realization, in the years after the 1954 Supreme Court *Brown* decision requiring racial integration of schools, that "separate and unequal" was still the norm rather than the exception. Passage of the U.S. Elementary and Secondary Education Act in 1965 contributed greatly to more



comprehensive approaches to evaluation. The massive influx of federal money aimed at desegregation, innovation, compensatory education, greater equality of opportunity, teacher training, and higher student achievement was accompanied by calls for evaluation data to assess the effects on the nation's children: To what extent did these changes really make an educational difference?

But education was only one arena in the War on Poverty of the 1960s. Great Society programs from the Office of Economic Opportunity were aimed at nothing less than the elimination of poverty. The creation of large-scale federal health programs, including community mental health centers, was coupled with a mandate for evaluation, often at a level of 1 percent to 3 percent of program budgets. Other major programs were created in housing, employment, services integration, community planning, urban renewal, welfare, and so on—the whole of which came to be referred to as "butter" (in contrast to "guns") expenditures. In the 1970s, these Great Society programs collided head on with the Vietnam War, rising inflation, increasing taxes, and the fall from glory of Keynesian economics. All in all, it was what sociologists and social historians, with a penchant for understatement, would characterize as "a period of rapid social and economic change."

Program evaluation as a distinct field of professional practice was born of two lessons from this period of large-scale social experimentation and government intervention: first, the realization that there is not enough money to do all the things that need doing and, second, even if there were enough money, it takes more than money to solve complex human and social problems. As not everything can be done, there must be a basis for deciding which things are worth doing. Enter evaluation.

High Hopes for Evaluation

> *One of the most appealing ideas of our century is the notion that science can be put to work to provide solutions to social problems.*
>
> —Political Sociologist Hans Zetterman (quoted in Suchman [1967:1])

Evaluation and Rationality

The great sociologist Max Weber (1864–1920), founder of organizational sociology, predicted that modern institutions would be the foundation of ever-increasing rationality in human affairs. "Modernity, Weber said, is the progressive disenchantment of the world. Superstitions disappear; cultures grow more homogeneous; life becomes increasing rational" (Menand 2006:84). Evaluation epitomizes Weber's vision of rationality in the modern world. Donald Campbell (1917–1996) picked up the mantle of Weber's work on the sociology of science and rearticulated his vision of modernity, explicitly incorporating evaluation as a cornerstone of rationality, and expressed in the ideal of an "experimenting society":

> It would be an *active society* preferring exploratory innovation to inaction. . . . It will be an *evolutionary*, *learning society*. . . . It will be an *honest society*, committed to *reality testing*, to self-criticism, to avoiding self-deception. It will say it like it is, face up to the facts, be undefensive and open in self-presentation. (Campbell 1999:13)

The ascendance of applied social and behavioral sciences was driven by hope that knowledge could be used rationally to make the world a better place, that is, that social sciences would yield *practical*

knowledge (Stehr 1992). In 1961, Harvard-educated President John F. Kennedy welcomed scientists to the White House as never before. Scientific perspectives were taken into account in the writing of new social legislation. Economists, historians, psychologists, political scientists, and sociologists were all welcomed into the public arena to share in the reshaping of modern postindustrial society. They dreamed of and worked for a new order of rationality in government—a rationality undergirded by social scientists who, if not exemplifying Plato's philosopher-kings themselves, were at least ministers to philosopher-kings. Carol Weiss has captured the optimism of that period:

> There was much hoopla about the rationality that social science would bring to the untidy world of government. It would provide hard data for planning . . . and give cause-and-effect theories for policy making, so that statesmen would know which variables to alter in order to effect the desired outcomes. It would bring to the assessment of alternative policies a knowledge of relative costs and benefits so that decision-makers could select the options with the highest payoff. And once policies were in operation, it would provide objective evaluation of their effectiveness so that necessary modifications could be made to improve performance. (Weiss 1977:4)

While pragmatists turned to evaluation as a commonsensical way to figure out what works and is worth funding, visionaries were conceptualizing evaluation as the centerpiece of a new kind of society: "the experimenting society." Donald T. Campbell gave voice to this vision in his 1971 address to the American Psychological Association:

> The experimenting society will be one which will vigorously try out proposed solutions to recurrent problems, which will make hard-headed and multidimensional evaluations of the outcomes, and which will move on to other alternatives when evaluation shows one reform to have been ineffective or harmful.
>
> We do not have such a society today. (Campbell 1991:223)

Early visions for evaluation, then, focused on evaluation's expected role in guiding funding decisions and differentiating the wheat from the chaff in federal programs. But as evaluations were implemented, a new role emerged: helping improve programs as they were implemented. The Great Society programs floundered on a host of problems: management weaknesses, cultural issues, and failure to take into account the enormously complex systems that contributed to poverty. Wanting to help is not the same as knowing how to help; likewise, having the money to help is not the same as knowing how to spend money in a helpful way. Many War on Poverty programs turned out to be patronizing, controlling, dependency generating, insulting, inadequate, misguided, overpromised, wasteful, and mismanaged. Evaluators were called on not only to offer final judgments about the overall effectiveness of programs but also to gather process data and provide feedback to help solve problems along the way.

By the mid-1970s, interest in evaluation had grown to the point where two professional organizations were established: the academically oriented Evaluation Research Society and the practitioner-oriented Evaluation Network. In 1984, they merged as the American Evaluation Association. By that time, interest in evaluation had become international with establishment of the Canadian Evaluation Society and the Australasian Evaluation Society.

One manifestation of the scope, pervasiveness, and penetration of the high hopes for evaluation is the number of evaluation studies conducted. As early as 1976, the

Congressional Sourcebook on Federal Program Evaluations contained 1,700 citations of program evaluation reports issued by 18 U.S. executive branch agencies and the General Accounting Office (GAO) during fiscal years 1973 through 1975 (Office of Program Analysis, GAO 1976:1). In 1977, federal agencies spent $64 million on program evaluation and more than $1.1 billion on social research and development (Abramson 1978). The third edition of the Compendium of Health and Human Services Evaluation Studies (HHS 1983) contained 1,435 entries. The fourth volume of the U.S. Comptroller General's directory of Federal Evaluations (GAO 1981) identified 1,429 evaluative studies from various U.S. federal agencies completed in fiscal year 1980. While the large number of and substantial funding for evaluations suggested great prosperity and acceptance, under the surface and behind the scenes a crisis was building—a utilization crisis.

Reality Check: Evaluations Largely Unused

By the end of the 1960s, it was becoming clear that evaluations of "Great Society" social programs were largely ignored or politicized. The utopian hopes for a scientific and rational society had somehow failed to be realized. The landing of the first human on the moon came and went, but poverty persisted despite the 1960s "war" on it—and research was still not being used as the basis for government decision making. While all types of applied social science suffered from underuse (Weiss 1977, 1972a), nonuse seemed to be particularly characteristic of evaluation studies. Ernest House (1972) put it this way: "Producing data is one thing! Getting it used is quite another" (p. 412). Williams and Evans (1969) wrote that "in the final analysis, the test of the effectiveness of outcome data is its impact on implemented policy. By this standard, there is a dearth of successful evaluation studies" (p. 119). Wholey et al. (1970) concluded that "the recent literature is unanimous in announcing the general failure of evaluation to affect decision making in a significant way" (p. 46). They went on to note that their own study "found the same absence of successful evaluations noted by other authors" (Wholey et al. 1970:48). There was little evidence to indicate that government planning offices had succeeded in linking social research and decision making. Seymour Deitchman (1976), in his *Tale of Social Research and Bureaucracy*, did not mince words: "The impact of the research on the most important affairs of state was, with few exceptions, nil" (p. 390). Weidman et al. (1973) concluded that "on those rare occasions when evaluations studies have been used . . . the little use that has occurred [has been] fortuitous rather than planned" (p. 15). In 1972, the eminent evaluation scholar Carol Weiss viewed underutilization as one of the foremost problems in evaluation research: "A review of evaluation experience suggests that evaluation results have not exerted significant influence on program decisions" (Weiss 1972c:10–11).

This conclusion was echoed by four prominent commissions and study committees: the U.S. House of Representatives Committee on Government Operations, Research and Technical Programs Subcommittee (1967); the Young Committee report published by the National Academy of Sciences (1968); the Report of the Special Commission on the Social Sciences for the National Science Foundation (1968); and the Social Science Research Council's prospective on the Behavioral and Social Sciences (1969).

British economist L. J. Sharpe (1977) reviewed the European literature and commission reports on use of social scientific

An Evaluation Report Disappears into the Void—And an Area of Inquiry Is Born

Sociologist and Harvard professor Carol Weiss is recognized in the *Encyclopedia of Evaluation* as the "Founding Mother" of evaluation (Mathison 2005:449). She was also the first to give prominence to the issue of evaluation use, a deep-seated interest arising from her experience in the 1960s evaluating a government program that was part of the "War on Poverty."

"I was asked to evaluate a program in central Harlem. One of the program's goals was to bring black college students from universities in the south to work in central Harlem, to work in the schools, the hospitals and social agencies. They were trained and then they spent the year working in the community. When I finished my evaluation of the Harlem program, the report came out in 3 volumes. We sent copies of the report to Washington: I never heard a word from them! I had the feeling I could have just dumped it into the ocean and it would have made no difference. So, I asked myself: 'Why did they support and fund this evaluation if they were not going to pay any attention to it?' That's how I got interested in the uses of research: What was going on? What could researchers—or anyone else—do to encourage people to pay more attention to research?" (www.gse.harvard.edu/news/features/weiss09102001.html)

Weiss subsequently began studying and writing about knowledge utilization (Weiss 1977) and became one of the most influential contributors to our understandings of evaluation use, policy formulation, and organizational decision making (Alkin 2004). She has been one of the most visible and influential voices for the idea that cumulative evaluative evidence can contribute to significant program and policy changes expressed in the aphorism: *In Evidence Lies Change* (Graff and Christou 2001). Bottom line: "Utility is what evaluation is all about" (Weiss 2004:161).

knowledge and reached a decidedly gloomy conclusion:

> We are brought face to face with the fact that it has proved very difficult to uncover many instances where social science research has had a clear and direct effect on policy even when it has been specifically commissioned by government. (P. 45)

Ronald Havelock (1980) of the Knowledge Transfer Institute generalized that "there is a gap between the world of research and the world of routine organizational practice, regardless of the field" (p. 13). The same conclusions came forth time and again from different fields:

> At the moment there seems to be no indication that evaluation, although the law of the land, contributes anything to educational practice, other than headaches for the researcher, threats for the innovators and depressing articles for journals devoted to evaluation. (Rippey 1973:9)

More recent utilization studies continue to show low levels of research use in government decision making (Landry, Lamari, and Amara 2003)

It can hardly come as a surprise, then, that support for evaluation began to decline. During the Reagan Administration in the 1980s, the U.S. GAO found that federal evaluation received fewer resources and that "findings from both large and small studies have become less easily available for use by the Congress and the public" (GAO 1987:4). In both 1988 and 1992, the GAO prepared status reports on program evaluation to inform changing executive branch administrations at the federal level.

> We found a 22-percent decline in the number of professional staff in agency program

> evaluation units between 1980 and 1984. A follow-up study of 15 units that had been active in 1980 showed an additional 12-percent decline in the number of professional staff between 1984 and 1988. Funds for program evaluation also dropped substantially between 1980 and 1984 (down by 37 percent in constant 1980 dollars). . . . Discussions with the Office of Management and Budget offer no indication that the executive branch investment in program evaluation showed any meaningful overall increase from 1988 to 1992. (GAO 1992a:7)

The GAO went on to conclude that its 1988 recommendations to enhance the federal government's evaluation function had gone unheeded: "The effort to rebuild the government's evaluation capacity that we called for in our 1988 transition series report has not been carried out" (GAO 1992a:7). Here, ironically, we have an evaluation report on evaluation going unused.

In 1995, the GAO provided another report to the U.S. Senate on Program Evaluation subtitled "Improving the Flow of Information to Congress." GAO analysts conducted follow-up case studies of three major federal program evaluations: the Comprehensive Child Development Program, the Community Health Centers Program, and the Chapter 1 Elementary and Secondary Education Act aimed at providing compensatory education services to low-income students. The analysts concluded that

> lack of information does not appear to be the main problem. Rather, the problem seems to be that available information is not organized and communicated effectively. Much of the available information did not reach the [appropriate Senate] Committee, or reached it in a form that was too highly aggregated to be useful or that was difficult to digest. (GAO 1995:39)

Many factors affect evaluation use in Congress, but politics is always a dominant factor (Chelimsky 2007; 2006a, 2006b; Julnes and Rog 2007; Mohan and Sullivan 2007). Evaluation use throughout the U.S. federal government continued its spiral of decline through the early 1990s (Popham 1995; Wargo 1995; Chelimsky 1992). In many federal agencies, the emphasis shifted from program evaluation to inspection, auditing, and investigations (Smith 1992; Hendricks, Mangano, and Moran 1990). Then came attention to and adoption of performance monitoring for accountability and the picture changed dramatically.

New Directions in Accountability: Reinventing Government

A predominant theme of the 1995 International Evaluation Conference in Vancouver was worldwide interest in reducing government programs and making remaining programs more effective and accountable. Decline in support for government programs was fueled by the widespread belief that such efforts were ineffective and wasteful. While the Great Society and War on Poverty programs of the 1960s had been founded on good intentions and high expectations, they came to be perceived as a failure. The "needs assessments" that had provided the rationales for those original programs had found that the poor, the sick, the homeless, and the uneducated—the needy of all kinds—needed services. So services and programs were created. Thirty years down the road from those original efforts, and billions of dollars later, most social indicators revealed little improvement. Poverty statistics, rates of homelessness, hard core unemployment and underemployment, multigenerational welfare recipients, urban degradation, and

crime rates combined to raise questions about the effectiveness of services. Reports on effective programs (e.g., Guttmann and Sussman 1995; Kennedy School of Government 1995; Schorr 1988) received relatively little media attention compared with the relentless press about waste and ineffectiveness (Wortman 1995). In the 1990s, growing concerns about federal budget deficits and runaway entitlement costs intensified the debate about the effectiveness of government programs. Both conservatives and liberals were faced with public demands to know what had been achieved by all the programs created and all the money spent. The call for greater accountability became a watershed flowing at every level—national, state, and local; public sector, not-for-profit agencies, and the private sector (Mohan and Sullivan 2007; Chelimsky 2006a; Harvard Family Research Project 1996a, 1996b).

Clear answers were not forthcoming. Few programs could provide data on results achieved and outcomes attained. Internal accountability had come to center on how funds were spent (inputs monitoring), eligibility requirements (who gets services and client characteristics), how many people get services, what activities they participate in, and how many complete the program. These indicators of inputs, client characteristics, activities, and outputs (program completion) measured whether providers were following government rules and regulations rather than whether desired results were being achieved. Control had come to be exercised through audits, licensing, and service contracts rather than through measuring outcomes. The consequence was to make providers and practitioners compliance-oriented rather than results-focused. Programs were rewarded for doing the paperwork well rather than making a difference in clients' lives.

Public skepticism turned to deep-seated cynicism. Polling data showed a widespread perception that "nothing works." As an aside, and in all fairness, this perception is not unique to the late twentieth century. In the nineteenth century, Spencer traced 32 acts of the British Parliament and discovered that 29 produced effects contrary to those intended (Edison 1983:5). Given today's public cynicism, three effective programs out of 32 might be considered a pretty good record.

More damning still, in modern times, the perception has grown that no relationship exists between the amount of money spent on a problem and the results accomplished, an observation made with a sense of despair by economist John Brandl in his keynote address to the AEA in New Orleans in 1988. Brandl, a professor in the Hubert H. Humphrey Institute of Public Affairs at the University of Minnesota (formerly its Director), was present at the creation of many human services programs during his days at the old Department of Health, Education and Welfare (HEW). He created the interdisciplinary Evaluation Methodology training program at the University of Minnesota. He later moved from being a policy analyst to being a policy formulator as a Minnesota state legislator. His opinions carried the weight of both scholarship and experience. In his keynote address to professional evaluators, he opined that no demonstrable relationship exists between program funding levels and impact, that is, between inputs and outputs; more money spent does not mean higher quality or greater results.

In a later article, Brandl updated his analysis. While his immediate focus was on Minnesota state government, his comments

characterize general concerns about the effectiveness of government programs in the 1990s:

> The great government bureaucracies of Minnesota and the rest of America today are failing for the same reason that the formerly Communist governments in Europe fell a few years ago. . . . There is no systematic accountability. People are not regularly inspired to do good work, rewarded for outstanding performance, or penalized for not accomplishing their tasks.
>
> In bureaus, people are expected to do well because the rules tell them to do so. Indeed, often in bureaus here and abroad, able, idealistic workers become disillusioned and burned-out by a system that is not oriented to produce excellent results. No infusion of management was ever going to make operations of the Lenin shipyard in Gdansk effective.
>
> Maybe—I would say surely—until systematic accountability is built into government, no management improvements will do the job. (Brandl 1994:13A)

Similar indictments of government effectiveness became the foundation for efforts at Performance Monitoring, Total Quality Management, Reengineering Government, Management by Objectives (MBO), Reinventing Government, and Managing for Results. Such public sector initiatives made greater accountability and performance monitoring, and increased use of evaluation, central to reform in U.S. federal and state governments, as well as governments around the world, notably Australia, Canada, New Zealand, and the United Kingdom (Moynihan 2006; Rogers 2006; Sears 2006). In this vein, Exhibit 1.2 illustrates the premises for results-oriented government as promulgated by Osborne and Gaebler (1992) in their influential and best-selling book *Reinventing Government*.

In the United States, 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

EXHIBIT 1.2

Premises of Reinventing Government

- What gets measured gets done.
- If you don't measure results, you can't tell success from failure.
- If you can't see success, you can't reward it.
- If you can't reward success, you're probably rewarding failure.
- If you can't see success, you can't learn from it.
- If you can't recognize failure, you can't correct it.
- If you can demonstrate results, you can win public support.

From Osborne and Gaebler (1992, chap. 5)

develop multiyear strategic plans, annual performance plans, and annual performance reports.

It is now an entrenched part of American politics that each new presidential administration will initiate new performance monitoring and accountability requirements. The Bush administration focused a good portion of its campaign rhetoric on 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 in order to inform funding and management decisions aimed at making the program more effective. The PART framework sets as its goal an evaluation of "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 aims to examine program improvements over time and allow comparisons between similar programs. William Trochim (2006a), 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." In Chapter 4, we shall examine the utility of these accountability initiatives.

Seemingly endless administrative reforms with a focus on accountability are by no means limited to the U.S. government. It is indicative of the political power and public appeal of accountability-oriented government reforms that one of the first things

the newly elected Conservative government of Prime Minister Harper did in Canada was pass a 255-page "Accountability Act." In one review of the Act, political observer Robin Sears (2006) concluded that it was one more effort in a long tradition of trying "to tame the twin nightmares of every modern democracy: lousy management of public spending, and a broad conviction among voters that insiders get favours from government" (p. 19). Making government accountability meaningful, credible, and useful is one of the challenges facing all modern democracies (Chelimsky 2006a, 2006b).

> I was working with a major, long-established organization. In a meeting with senior management to get the evaluation off to a good start, I asked them to tell me about an evaluation that had been useful to them, to start exploring the features that make evaluation useful. There was a long, nervous silence until one of them said, "None, really."
>
> "Then I guess we'll have to do things quite differently," I said.
>
> —An experienced evaluator

Misuse of Evaluations

Utilization-focused evaluation can be located between two extremes. One extreme, as just discussed, is the oversimplified image of analyzing evaluation findings then mechanically making instantaneous decisions based on those findings, for example, the simplistic expectation that PART effectiveness scores (or any simple grading system that categorizes results) should nicely match budget allocations (high scores equal more funds, low scores mean program termination). Real-world evaluation use, we shall find, is more complex, nuanced, and interpretative. Moving from data to action involves treading a path fraught with obstacles. Evaluators who successfully facilitate use of findings need technical skill, to be sure, but they also need to be good communicators, have political savvy, understand how organizations function, and know how to work with a variety of people with different learning and decision-making styles and competing interests.

If one extreme is an image of simple, mechanical, and immediate use, the other extreme is ignoring evaluation findings altogether, or worse, misusing them. Evaluation findings are not going to be methodologically or technically perfect. Debates about focus, measurement challenges, design weaknesses, sampling problems, and controversies about what the data mean are the rule rather than the exception. The world is a messy place. Programs are messy and complex. Studying the world and evaluating programs is difficult because, in doing so, we encounter what William James (1950) famously called "one great blooming, buzzing confusion" (p. 488). Clear, precise, certain, and noncontroversial findings are elusive, a modern chimera, especially on matters about which people have differing opinions and perspectives, which is just about everything. The imperfections of research designs and the difficulties of moving from evaluation findings to action are not, however, reasons to ignore evaluation findings altogether, or worse yet, manipulate the findings and interpretations to support preconceived positions and biases. Let's distinguish then, right here in the first chapter, between seriously taking evaluation findings into account as part of a complex and multifaceted process of deliberation versus ignoring findings altogether because the person

getting the findings doesn't like how they came out, or that person manipulating the findings to make them come out the way he or she wants them to be.

Thus, we face not only the challenge of increasing evaluation use. We also must be concerned with misuse, deception, and abuse. Marv Alkin (1990, 2004), an early theorist of user-oriented evaluation, has long emphasized that evaluators must attend to appropriate use, not just amount of use, and be concerned about misuse (Christie and Alkin 1999; Alkin and Coyle 1988). Ernest House, one of the most astute observers of how the evaluation profession has developed, observed in this regard: "Results from poorly conceived studies have frequently been given wide publicity, and findings from good studies have been improperly used" (1990a:26). The field faces a dual challenge then: supporting and enhancing appropriate uses while also working to eliminate improper uses (Patton 2005b).

In 2004, Philip A. Cooney, chief of staff for the White House Council on Environmental Quality, repeatedly edited government climate reports to play down links between such emissions and global warming. Before joining the Bush White House, Cooney had been a lobbyist at the American Petroleum Institute, the largest trade group representing the interests of the oil industry, where he led the oil industry's fight against limits on greenhouse gases. He was trained as a lawyer with a bachelor's degree in economics, but with no scientific training. News accounts (e.g., *New York Times*, June 8, 2005) reported that Cooney removed or adjusted descriptions of climate research that government scientists and their supervisors had already approved. The dozens of changes, while sometimes as subtle as the insertion of the phrase "significant and fundamental" before the word "uncertainties," tended to raise doubts about findings, despite a consensus among climate experts that the findings were robust. In one instance, he changed an October 2002 draft of a regularly published summary of government climate research, "Our Changing Planet," by adding the word "extremely" to this sentence: "The attribution of the causes of biological and ecological changes to climate change or variability is extremely difficult." In a section on the need for research into how warming might change water availability and flooding, he crossed out a paragraph describing the projected reduction of mountain glaciers and snowpack.

Such distortions don't just happen with politically motivated advisors protecting national policies. Evaluators at a local level regularly report efforts by program staff, administrators, and elected officials to alter their findings and conclusions. As I was working on this very section, I received a phone call from an evaluation colleague in a small rural community who had just received a request from an agency director to rewrite an evaluation report "with a more positive tone" and leave out some of the negative quotations from participants. She wanted help with language that would diplomatically but firmly explain that such alterations would be unethical.

There is irony here that, in a broad historical context, is worth noting. When the first edition of this book was published in 1978, just as the field of evaluation was emerging, the primary concern was getting anyone to pay any attention to evaluations and take findings seriously. In the last quarter century, a sea change has occurred in political rhetoric. Now, in the information and knowledge age, politicians, policymakers, business leaders, and not-for-profit advocates have learned that the public expects them to address problems through a research and

evaluation lens. Public debates regularly include these questions: What do we actually know about this issue? What does the research show? What are evaluation findings about the effectiveness of attempted interventions and solutions? From 1981 to 2006, the frequency of articles about "accountability" in the *New York Times* increased some fivefold, from 124 to 624, a rate by 2006 of at least one per day (John Bare, Vice President, The Arthur M. Blank Foundation, Atlanta, Georgia, 2007 personal communication).

The irony is that as evaluation has become more prominent and more used, it has also become more subject to manipulation and abuse. Thus, critics of the Bush administration consider the Cooney manipulation of climate research to be business-as-usual in politics rather than an exception. The Centers for Disease Control, long the world's leading source for high-quality, credible research and evaluation, has been subject to such manipulation. Data about the ineffectiveness of abstinence-only sex education programs have been manipulated and suppressed; and clear evidence about the effectiveness of condoms in preventing HIV-AIDS has been denigrated. On the National Cancer Institute Web site, evidence about a correlation between abortions and cancer was fabricated and disseminated. There are a great many other examples, not least of which was the manipulation and distortion of intelligence about weapons of mass destruction to justify the Iraq invasion (Specter 2006). All political groups attempt to support their preferred ideological positions by championing empirical findings that support their beliefs and denigrating evidence that runs counter to their beliefs. In both the Clinton and the Bush administrations, findings about the effectiveness of needle exchanges to prevent HIV transmission were dismissed. Scholarship published by the National Society for the Study of Education has documented both the uses and the misuses of data for educational accountability, especially in the standards-based *No Child Left Behind* initiative of the U.S. federal government (Herman and Haertel 2005); misuses have resulted from lack of competence, inadequate resources, political pressures, and, in some cases, premeditation and corruption.

These examples illustrate some of the political and ethical challenges evaluators face in working to get evaluation findings taken seriously and used (Chelimsky 1995b). We'll look at these issues in depth in later chapters, especially how evaluators can meet their responsibility to assure the integrity and honesty of evaluations as called for in the Guiding Principles adopted by the AEA.

> Evaluators display honesty and integrity in their own behavior and attempt to ensure the honesty and integrity of the entire evaluation process. (AEA Task Force on Guiding Principles for Evaluators 1995; see Exhibit 1.3)

Standards of Excellence for Evaluation

Concerns about ethics, the quality of evaluations, and making evaluations useful undergirded an early effort by professional evaluators to articulate standards of practice. To appreciate the importance of the standards, let's begin with some context. Prior to adoption of standards, many researchers took the position that their responsibility was merely to design studies, collect data, and publish findings; what decision makers did with those findings was not their problem. This stance removed from the evaluation researcher any responsibility for fostering use and placed all the blame for nonuse or underutilization on decision makers.

EXHIBIT 1.3

Guiding Principles for Evaluators

Systematic Inquiry
Evaluators conduct systematic, data-based inquiries about what is being evaluated.

Competence
Evaluators provide competent performance to stakeholders.

Integrity/Honesty
Evaluators display honesty and integrity in their own behavior, and attempt to ensure the honesty and integrity of the entire evaluation process.

Respect for People
Evaluators respect the security, dignity, and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.

Responsibilities for General and Public Welfare
Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.

American Evaluation Association (AEA), 1995
Task Force on Guiding Principles for Evaluators
(See also Shadish, Newman, Scheirer, and Wye 1995)

For detailed elaboration and discussion of the specific Guiding Principles adopted by the American Evaluation Association, see www.eval.org/Publications/GuidingPrinciples.asp

Academic aloofness from the messy world in which research findings are translated into action has long been a characteristic of basic scientific research. Before the field of evaluation generated its own standards in the late 1970s, criteria for judging evaluations were based on the quality standards of traditional social and behavioral sciences, namely, technical quality and methodological rigor. Use was ignored. Methods decisions dominated the evaluation design process. Methodological rigor meant experimental designs, quantitative data, and sophisticated statistical analysis. Whether decision makers understood such analyses was not the researcher's problem. Validity, reliability, measurability, and generalizability were the dimensions that received the greatest attention in judging evaluation research proposals and reports (e.g., Bernstein and Freeman 1975). Indeed, evaluators concerned about increasing a study's usefulness often called for ever more methodologically rigorous evaluations to increase the validity of findings, thereby hoping to compel decision makers to take findings seriously.

By the late 1970s, however, it was becoming clear that greater methodological rigor was not solving the use problem. Program

staff and funders were becoming openly skeptical about spending scarce funds on evaluations they couldn't understand and/or found irrelevant. Evaluators were being asked to be "accountable" just as program staff members were supposed to be accountable. The questions emerged with uncomfortable directness: Who will evaluate the evaluators? How will evaluation be evaluated? It was in this context that professional evaluators began discussing standards.

The most comprehensive effort at developing standards was hammered out over 5 years by a 17-member committee appointed by 12 professional organizations, with input from hundreds of practicing evaluation professionals. The standards published by the Joint Committee on Standards in 1981 dramatically reflected the ways in which the practice of evaluation had matured. Just prior to publication, Dan Stufflebeam, Chair of the Committee, summarized the committee's work as follows:

> The standards that will be published essentially call for evaluations that have four features. These are *utility, feasibility, propriety* and *accuracy*. And I think it is interesting that the Joint Committee decided on that particular order. Their rationale is that an evaluation should not be done at all if there is no prospect for its being useful to some audience. Second, it should not be done if it is not feasible to conduct it in political terms, or practicality terms, or cost effectiveness terms. Third, they do not think it should be done if we cannot demonstrate that it will be conducted fairly and ethically. Finally, if we can demonstrate that an evaluation will have utility, will be feasible and will be proper in its conduct, then they said we could turn to the difficult matters of the technical adequacy of the evaluation. (Stufflebeam 1980:90)

In 1994 and 2008, revised standards were published following extensive reviews spanning several years (Stufflebeam 2007; Patton 1994b). While some changes were made in the individual standards, the overarching framework of four primary criteria remained unchanged: *utility*, *feasibility*, *propriety*, and *accuracy* (see Exhibit 1.4). Specific standards have also been adapted to various international contexts (Russon and Russon 2004), but the overall framework has translated well cross-culturally. Taking the standards seriously has meant looking at the world quite differently. Unlike the traditionally aloof stance of purely academic researchers, professional evaluators are challenged to take responsibility for use. No more can we play the game of blaming the resistant decision maker. If evaluations are ignored or misused, we have to look at where our own practices and processes may have been inadequate. *Implementation of a utility-focused, feasibility-conscious, propriety-oriented, and accuracy-based evaluation requires situational responsiveness, methodological flexibility, multiple evaluator roles, political sophistication, and substantial doses of creativity, all elements of utilization-focused evaluation.*

Daniel Stufflebeam (2001), the guiding leader of the standards movement in evaluation, undertook a comprehensive, exhaustive, and independent review of how 22 different evaluation approaches stack up against the standards. No one was better positioned by knowledge, experience, prestige within the profession, and commitment to the standards to undertake such a challenging endeavor. He concluded, "Of the variety of evaluation approaches that emerged during the twentieth century, nine can be identified as strongest and most promising for continued use and development." Utilization-focused evaluation was among those nine, with the highest rating for adherence to the utility standards (p. 80).

EXHIBIT 1.4

Standards for Evaluation

UTILITY
The Utility Standards are intended to ensure that an evaluation will serve the practical information needs of intended users.

FEASIBILITY
The Feasibility Standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.

PROPRIETY
The Propriety Standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

ACCURACY
The Accuracy Standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated.

(Joint Committee on Standards for Educational Evaluation 1994)

For the full set of detailed standards, see www.wmich.edu/evalctr/jc

Worldwide Surge in Demand for Evaluation

Interest in evaluation has surged in the new millennium, including a proliferation of different models and approaches (Stufflebeam and Shinkfield 2007). But no trend has been more important to evaluation in the last decade than its expanding global reach. In the 1970s and 1980s, professional evaluation associations began to appear: the Canadian Evaluation Society, the Australasian Evaluation Society, and the AEA. In 1995, evaluation professionals from 61 countries around the world came together at the first truly international evaluation conference in Vancouver, British Columbia. Ten years later, a second international conference in Toronto attracted 2,330 evaluation professionals from around the world. The 1990s also gave rise to the European Evaluation Society (founded in 1994 in the Hague) and the African Evaluation Association (founded in 1999 in Nairobi and having held its fourth continent-wide conference in Niamey, Niger, in 2007). Now there are more than 60 national evaluation associations around the world, including Japan, Malaysia, Sri Lanka, Mongolia, Russia, Brazil, Colombia, Peru, South Africa, Zimbabwe, Niger, and New Zealand, to name but a few examples. In 2003 in Lima, Peru, the inaugural meeting of the new International Organization for Cooperation in Evaluation (IOCE) was held as an umbrella networking and support initiative for national and regional evaluation

associations around the world. The International Development Evaluation Association (IDEAS) was formed in Beijing in 2002 to support evaluators with special interests in developing countries; its first biennial conference was held in New Delhi in 2005. The Network for Monitoring, Evaluation, and Systematization of Latin America and the Caribbean (ReLAC) was formed in 2005 in Peru.

Commemorating 20 Years of Evaluation Scholarship

In 2005, the *Canadian Journal of Program Evaluation* (CJPE) published a special 20th anniversary issue (www.cjpe.ca). The volume featured articles on the state of the art of evaluation in Canada in a variety of domains of practice, including health, education, child welfare, social services, and government as well as two independent content analyses of CJPE since the publication of Volume 1, Number 1 in 1986. All issues are available online.

In 2007, the American Evaluation Association journal *New Directions for Evaluation* celebrated its 20th anniversary with a review of enduring issues: judging interpretations, theory-based evaluation, participatory evaluation, and cultural issues (Cousins and Whitmore 2007; Datta 2007a; King 2007b; Leviton 2007; Lipsey 2007c; Madison 2007; Mark 2007; Mathison 2007; Rogers 2007; Schwandt 2007a).

Evaluation capacity can be a crucial part of what the World Bank calls "the knowledge-based economy." The Bank's Knowledge Assessment Methodology (KAM 2006) is an interactive benchmarking tool aimed at helping countries identify the challenges and opportunities they face in making the transition to the knowledge-based economy. The World Bank, through its International Program for Development Evaluation Training, also offers annually, month-long evaluation training for people throughout the developing world (Cousins 2006a).

International agencies have developed comprehensive guidelines for the conduct of evaluation (e.g., United Nations Development Programme 2007; United Nations Evaluation Group 2007, 2005a, 2005b; Danida 2006; Independent Evaluation Group 2006; International Organization for Management 2006; Organisation for Economic Co-operation and Development 2006; World Food Programme 2006; International Fund for Agricultural Development 2002). Various national associations have reviewed and adapted the Joint Committee Standards to their own socio-political contexts as the African Evaluation Association (AfrEA) did in adopting African Evaluation Guidelines in 2007 (AfrEA 2007; Russon and Russon 2004). Evaluation texts and anthologies are available for specific countries and languages, e.g., Italy (Stame 2007), France (Ridde and Dagenais 2007, Japan (Patton and Nagao 2000), and New Zealand (Lunt, Davidson, and McKegg 2003).

Such globally interconnected efforts made it possible for evaluation strategies and approaches to be shared worldwide. Thus, the globalization of evaluation supports our working together to increase our international understanding about factors that support program effectiveness and evaluation use. International perspectives also challenge Western definitions and cultural assumptions about how evaluations ought to be conducted and how quality ought to be judged. As the evaluation standards are

translated into different languages, national associations are adding their own cultural nuances and adapting practices to fit local political, social, organizational, economic, and cultural contexts (Stufflebeam 2004b).

Governments around the world are building new systems for monitoring and evaluation, aiming to adapt results-based management and performance measurement to support development (Rist and Stame 2006). International agencies have also begun using evaluation to assess the full range of development efforts under way in developing countries. Most major international organizations have their own evaluation units with guidelines, protocols, conferences, training opportunities, Web sites, and resource specialists. In his keynote address to the international conference in Vancouver, Masafumi Nagao (1995), a cofounder of Japan's evaluation society, challenged evaluators to think globally even as they evaluate locally, that is, to consider how international forces and trends affect project outcomes even in small and remote communities. This book will include attention to how utilization-focused evaluation offers a process for adapting evaluation processes to address multicultural and international issues and constituencies.

The International Evaluation Challenge

In 2005, distinguished international leaders meeting in Bellagio, Italy, committed their support for impact evaluations of social programs in developing countries. Participants noted that in 2005, donor countries committed $34 billion to aid projects addressing health, education, and poverty in the developing world, but evaluation of results was rare and inadequate. Developing countries themselves spent hundreds of billions more on similar programs. The leaders from international agencies, governments, research organizations, and philanthropic foundations endorsed five principles for action.

1. Impact studies are beneficial
2. Knowledge is a public good
3. A collective initiative to promote impact studies is needed
4. The quality of impact studies is essential
5. The initiative should be complementary, strategic, transparent, and independent

(Evaluation Gap Working Group 2006)

The challenges and opportunities for evaluation extend well beyond government-supported programming. Because of the enormous size and importance of government efforts, program evaluation is inevitably affected by trends in the public sector, but evaluation has also been growing in importance in the private and independent sectors. Corporations, philanthropic foundations, not-for-profit agencies, and nongovernmental organizations (NGOs) worldwide are increasingly turning to evaluators for help in enhancing their effectiveness.

All this ferment means that evaluation has become a many-splendored thing—a rich tapestry of models, methods, issues, approaches, variations, definitions, jargon, concepts, theories, and practices. And therein lies the rub. How does one sort through the many competing and contradictory messages

about how to conduct evaluations? The answer in this book is to stay focused on the issue of use—conducting evaluations that are useful and actually get used. And as Carol Weiss (1998b) observed in her keynote address to the AEA annual conference, the challenge is not just increasing use, "but more effective utilization, use for improving daily program practice and also use for making larger changes in policy and programming" (p. 30).

From Problem to Solution: Toward Use in Practice

The future of evaluation is tied to the future effectiveness of programs. Indictments of program effectiveness are, underneath, also indictments of evaluation. The original promise of evaluation was that it would point the way to effective programming. Later, that promise broadened to include providing ongoing feedback for improvements during implementation. Evaluation cannot be considered to have fulfilled its promise if, as is increasingly the case, the general perception is that few programs have attained desired outcomes, that "nothing works."

As this introduction and historical overview closes, we are called back to the early morning scene that opened this chapter: decision makers lamenting the disappointing results of an evaluation, complaining that the findings did not tell them what they needed to know. For their part, evaluators complain about many things as well, but for a long time their most common complaint has been that their findings are ignored (Weiss 1972d:319). The question from those who believe in the importance and potential utility of evaluation remains: "What has to be done to get results that are appropriately and meaningfully used?" This question has taken center stage as program evaluation has emerged as a distinct field of professional practice—and that is the question this book answers. In doing so, we recognize a lineage that extends much farther back than the more recent establishment of the evaluation profession.

> Scientists now calculate that all living human beings are related to a single woman who lived roughly 150,000 years ago in Africa, a "mitochondrial Eve." . . . all humanity is linked to Eve through an unbroken chain of mothers. (Shreeve 2006:62)

And, adds Halcolm, she was an evaluator.

Follow-Up Exercises

1. Scan recent issues of local and/or national newspapers. Look for articles that report evaluation findings. Write a critique of the press report. Can you tell what was evaluated? Are the methods used in the evaluation discussed? How clear are the findings from the evaluation? Can you tell how the findings have been or will be used? How balanced and comprehensive is the press report?

2. See if you can locate the actual evaluation report discussed in the newspaper (Question 1 above). Many evaluation reports are posted on the Internet. Write your own newspaper report based on what you consider important in the evaluation. How is your press report different from the one you found in the newspaper? Why? What does this tell you about the challenges of disseminating evaluation findings to the general public?

3. Find the Web site for one of the international or national evaluation associations. Review the site and its offerings.

4. Most major philanthropic foundations, federal agencies, and international

organizations have Web sites with access to evaluation policies, guidelines, and reports. Visit the evaluation sections of at least one government site and one nongovernmental site. Write a comparison of their information about and approaches to evaluation.

5. Review the full list of Program Evaluation Standards (www.wmich.edu/evalctr/jc). (a) Conduct a cultural assumptions analysis of the standards. What specific standards, if any, strike you as particularly Western in orientation? (b) Select and discuss at least two standards that seem to you unclear, that is, you aren't sure what you would have to do in an evaluation to meet that particular standard.