1

THE PURPOSES OF EVALUATION IN A DEMOCRATIC SOCIETY

Eleanor Chelimsky

The evaluation of public policies, programs, and practices seems to be an intrinsic part of democratic government for four reasons. It reports information about government performance that the public needs to know. It adds new data to the existing stock of knowledge required for government action. It develops an analytical capability within agencies that moves them away from territoriality and toward a culture of learning. And, more generally, its spirit of skepticism and willingness to embrace dissent help keep the government honest. Evaluation thus serves many purposes, and it is common to find that what may have begun, say, as an accountability study of government performance, ends up dominated by a different purpose or at least includes other purposes as an integral part of the evaluation.

Many evaluators, however, disagree about the viability of the various purposes, often favoring one over another. Some evaluators say that evaluation is valuable only when it measures accountability, that is, when it maximizes tax resources by holding policy-makers and program managers accountable for the merit and worth of their policies and programs. Others see evaluation as valuable only when it generates knowledge, when it brings new or more profound understandings in some specific area of public endeavor. Still others believe that without evaluation capacity in government agencies, nothing good can happen: for them, evaluation is valuable only when it improves institutions, moving them from performance measurement and self-evaluation to the learning organization.

This would not be much of a problem if we evaluators took a more inclusive stance about why we do what we do. After all, these three perspectives are not mutually exclusive. But people who engage in one type of evaluation

often condemn the work of others who engage in a different kind. There have been veritable battles in the literature between warriors claiming, on the one hand, that "the only real evaluation is ultimately a fundamental judgment of merit or worth," and, on the other, that there is a whole "menu of evaluation purposes" – seven and still expanding – that have equal importance with merit and worth (Patton, 1996).

Looking at evaluation only from an evaluator's perspective might cause us to underestimate, misinterpret, or rule out purposes for evaluation that we would recognize as valid if we saw them from a different, broader perspective. We're not unlike those ants, asked to write a zoology paper, who divided the animal kingdom into two classes: the kind, gentle beasts such as the lion, tiger or jackal, and the ferocious ones like the chicken, duck or goose.

Still, it's not really surprising that we have such disparate perceptions of evaluation's purpose, and that we have not spent much time examining where it fits in government. Government in the United States became a prime player in evaluation starting only in the 1960s, whereas evaluation itself developed incrementally over more than 100 years. That lengthy development, and especially the dissimilar paths through which it evolved, could not help but influence the distinctions we make when we look at evaluation purpose today.

Varieties of Public Policy Evaluation in the United States

A first strand of evaluative development can be traced to government agricultural research begun in the early 1900s. The purpose was knowledge enhancement, to find out which agricultural practices would lead to the largest crop yield. Experimental design and statistical analysis techniques were applied, with advances coming from social scientists and statisticians, many of whom worked in universities as well as government.

By the 1950s, large-scale retrospective evaluations of the merit-and-worth type were being performed, using survey and computer-assisted techniques. Carefully evaluated demonstration programs came along in the 1960s, responding to the government's efforts to examine social programs' effectiveness in, say, moving people out of poverty or reducing crime. This path of evaluative development, focusing also on education, public health, equality of opportunity and other areas, drew on learning from a wide array of fields, including psychology, sociology, economics, political science and anthropology.

A second evaluative strand began during the 1950s, with efforts to rationalize the resource allocation and management of defense programs. Born in a think tank (the Rand Corporation) but developed within government, this eventually grew into the Department of Defense's Planning, Programming, and Budgeting System (PPBS), and was later expanded to other agencies under Lyndon Johnson. Focused on management improvement and the development of institutional capability, PPBS also had an underlying concern with questions of merit and worth. The thrust was to plan for program cost-effectiveness, and then to evaluate whether this had been achieved. Developed largely by economists and political scientists, the system used techniques such as policy analysis, cost-benefit, cost-effectiveness and systems analysis (Rhoads, 1978). Over time, these techniques implanted themselves into general evaluation practice (Rossi & Freeman, 1985).

These two strands of evaluative inquiry are much less distinct today. It is common to find techniques from both strands used in a single study. Nevertheless, the differing purposes, settings, disciplines, and mindsets (not to mention

differing evaluative questions, methods, and goals) have made for a climate in which it is easy for evaluators to claim exclusive value for one evaluation purpose over another.

When I first began evaluating weapons and military budget systems at NATO in 1966, however, I had no such exclusive views. My idea was to use any and all plausible theories, any and all proven methods, any and all types of analysis that could help answer the questions I was asked by Defense or State Department officials (USNATO was a combined mission of the two agencies). The purpose of my work was always perfectly clear, based on the policy or program need expressed in the question; the work itself, however, was not.

My biggest problem was that the data from NATO nations were often non-comparable. There were few standard definitions of categories and items; missing entries were common; costs depended on the sometimes wildly-varying currency rates used in the country calculations; and individual NATO ministries had a vested interest in keeping their defense figures as obscure as possible, often denying access to information I needed simply to adjust the data. Seemingly straightforward questions that policy-makers needed answers to – like "Why do military pension costs appear to be spiking in some countries but not in others?" – turned into data-sifting nightmares simply to determine whether the differences were real.

Using a clunky, government-issue calculator, a closetful of columnar pads, and a stack of No. 2 pencils, I scratched my way, slowly and laboriously, tabulating the entries of 15 nations. I sat on the floor, surrounded by giant yellow spreadsheets, sharpening pencils and making adjustments. It was not a comfortable process, but at least we evaluators at NATO were always told why our information was needed and what management or policy-making purpose it was intended to serve.

After we finally published a few studies and our work gained some credibility within NATO, we began trying to spread our evaluation methods to other national offices. We had both the altruistic aim of developing a more widespread analytical capability across NATO and the unabashedly self-interested one of being able to count on better data to support our studies.

Eventually, our work led to some basic policy questions about the impacts of the real drawdown of forces and resources we had found (a drawdown that had hitherto been camouflaged under a welter of indecipherable data). We were asked to evaluate the effects of some of these reductive actions by individual NATO nations on the readiness of NATO as a whole. These were, in fact, accountability studies, intended to illuminate the results of covert decisions and unfulfilled responsibilities, to link cause and effect.

Because of this experience, then, it has always seemed sensible to me that evaluation should have at least these three purposes: to gain new knowledge (by performing studies to answer questions about unknowns in a program or policy area); to improve agency capability (by using evaluation to improve problem-solving, database development, analytical skills, management practices, and the like); and to determine accountability (by performing studies that measure policy or program effectiveness or efficiency, assign responsibilities for successes and failures, and lay out options for improvement and correction). Indeed, these three purposes arise even in organizations that have oversight responsibilities and might be expected by some observers to rely mostly on accountability evaluations. In practice, accountability evaluations often depend on or give way to evaluations with another purpose.

Later, after a symposium looking at how the Congress, seven federal agencies, and the

Office of Management and Budget had actually used evaluations, I saw the same three purposes emerge from the analysis of the proceedings (Chelimsky, 1977b, 1978). Again, in 1995, at the international Vancouver conference of five evaluation societies, a look at the evaluations presented gave rise to the same view: the three major purposes were still there, and no single purpose could account for the body of evaluations in the field (Chelimsky, 1996; Chelimsky & Shadish, 1997).

This is not, of course, to suggest that no other purposes exist or could exist (indeed, I believe other chapters in this part address alternative possible purposes for evaluation, such as social justice). My point is that claiming a unique purpose for evaluation flies in the face of past and current practice. Nor is it my argument that any of the three purposes just discussed is typically present in a pristine state within an evaluation. Rather, these purposes rely on, and are intertwined with, each other in many different ways. Yet claimants for a unique purpose rarely admit how much, say, accountability evaluations depend on the work of preceding knowledge or development evaluations that have built the databases and descriptions of prior experience needed to establish accountability. Likewise, proponents of knowledge or development purposes don't often recognize the extent to which earlier or pending accountability studies create the climate of interest or political pressure that make their recommendations more likely to be heard and used.

Because so many evaluations have multiple purposes, then (even though a single purpose may dominate the others), I have chosen not to focus on any one of them here, but instead to examine their intermingling. In my experience, this better reflects the reality of program and policy evaluation in the public sphere.

But before looking at how these purposes work together to answer evaluation questions and facilitate the use of their findings in that murky, complex and painful process known as public decision-making, it may be important to go beyond evidence from experience or practice to understand why these particular purposes have emerged. Is there a larger basis for their presence, a structural core from which they spring? Can we deduce some necessity for their existence and how they fit together? To try to answer these questions, we need to look again at evaluative purpose, but this time from a political or governmental viewpoint, including the principles that dictate the need for evaluation in a democracy. I address this largely from the vantage point of the United States, but later in the chapter examine generalizability to other countries.

Governmental Structure and Congressional Oversight

In most democracies, and certainly in the United States, we find a government whose functions are split across three branches (legislative, executive, and judicial). Such a structure has the political goal of keeping too much power from accumulating in any one place. At least in the US, it is a structure born of distrust: distrust based on past experience with a coercive autocracy.

Such a structure is not, of course, without its disadvantages. Walls generated by individual branches and agencies to protect their independence also generate suspicion and secrecy. Fragmentation carries a host of impediments; sharing of information across agencies, for example, is rare. Most democracies are far away from efficient government performance. But conversely, calls to improve performance, to "make the trains run on time," as in fascist Italy, often turn out to be little more than disguised attempts to weaken individual rights or freedom. So there is some

tension between optimal performance in government and the preservation of liberty.

After the American Revolution, the framers of the US Constitution came up against this very tension: the need for a government to have enough power to govern, versus the distrust expressed by citizens and states of any overarching central authority that could become abusive or corrupt. The Constitution thus produced a divided governmental structure, featuring both a separation of powers among executive, legislative, and judicial branches, and a distribution of powers between federal and state levels of government. Although a far-reaching compromise between distrust and need had thus been achieved, it was so tenuous, vague, and ambiguous that the argument about it has never really been resolved (Ellis, 2002).

This architecture of "checks and balances" was, of course, intended to be an organizational bulwark – built on both external and internal controls – against too much centralized power. Madison wrote of "the necessary partition of power among the several departments" as part of an external control structure. He also called for internal controls to be established "by so contriving the interior structure of the government as that its several constituent parts may, by their mutual relations, be the means of keeping each other in their proper places." This is nothing less than a clarion call to agency independence. That is, "each department should have a will of its own; and consequently should be so constituted that the members of each should have as little agency as possible in the appointment of the members of the others" (Madison, 1788).

Yet the framers had few illusions about the dangers that could accompany too much agency independence, not to mention "the inevitable corruptions that could result when unseen rulers congregate in distant places" (Ellis, 2002). So they envisaged a check against agency autonomy in the form of congressional oversight, that is, the particular "authority to supervise the administration of government" which has led to so many confrontations over the years about accountability and secrecy in the executive branch (Jewell & Patterson, 1966).

Congress has oversight authority over the judicial branch, which it exerts, for example, through Senate approval of judicial appointments and the power to establish federal courts and prescribe their jurisdiction. Congress supervises the executive branch, through such mechanisms as the appropriations power, senatorial approval of nominations to executive office, and, notably, the investigation of how past legislation has been implemented (Jewell & Patterson, 1966). Through the Congressional Budget Control and Impoundment Act of 1974, Congress "greatly strengthened its resources for scrutiny and control of government programs and activities" (Bradshaw & Pring, 1981). Specifically, the Congress lodged within one of its own agencies (the General Accounting Office, or USGAO, now called the Government Accountability Office) the authority and responsibility to perform evaluations in support of congressional oversight.

In short, the US governmental structure builds in democratic protections through exterior and interior controls, agency independence, and congressional oversight. This structure is itself a careful equilibrium, a compromise arrived at between partisans of strong and limited government. Despite its apparent precarity, the framers believed their structure would stand against inevitable efforts to abuse or usurp power, but only if it could be vigorously supported by a well-informed public: that is, an electorate with enough distrust to sniff out problems in government, and also enough willingness and capability to correct them once they became known.

The People as Guarantors of Democracy, and the Information Function of Government

"One of the principal functions of a legislature is to inform the people about the activities of their government" (Bradshaw & Pring, 1981) – another idea that comes down to us from the framers. Madison recognized that organizational controls and the like were essentially adjuncts to the real power of the citizenry, which lies behind any governmental structure in a democracy: "A dependence on the people," Madison (1788) wrote, "is, no doubt, the primary control on the government." Jefferson went further in grappling with the issue of how even a vigilant population might become aware of a distant government's excesses or omissions and take steps to correct them. Writing to Abigail Adams in 1804, he noted the two political parties' disagreement about ensuring that the people should function as the best censor of government: "One side fears most the ignorance of the people; the other, the selfishness of rulers independent of them" (Jefferson, 1946).

Jefferson's view was that ignorance was curable:

> I am persuaded myself that the good sense of the people will always be found to be the best army. They may be led astray for a moment, but will soon correct themselves . . . . The way to prevent these [errors] of the people is to give them full information of their affairs through the channel of the public papers, and to contrive that these papers should penetrate the whole mass of the people . . . . I mean that every man should receive these papers and be capable of reading them. (Jefferson, 1946)

The press and education were thus the chief means through which Jefferson sought to involve the people. It is no accident that Jefferson considered his founding of the University of Virginia as one of the three main achievements of his life – along with the authorship of the Declaration of Independence and that of Virginia's statute for religious freedom (Bronowski & Mazlish, 1993). Jefferson put his faith especially in the spread of accurate information through science and research, and he believed knowledge should inform the activities of government: "Science is more important in a republican than in any other government" (Jefferson, 1946). As President, he conceived the Lewis and Clark expedition, which had as its primary purpose the acquisition of new knowledge (Hicks et al., 1964). Before that, under Vice President Jefferson's urging, the census – first begun only to inform on the size of the population – expanded its horizons to include details about the lifespan of Americans that could be used for social measures to improve longevity (Chelimsky, 1985). And it is no accident either that Jefferson so fervently championed a Bill of Rights that guaranteed, among other freedoms, freedom of the press, against the opinion of Hamilton.

Indeed, Hamilton wrote that he saw no need for a Bill of Rights, no need for concern about how to keep the citizenry well informed, and no need to guarantee freedom of the press, because individual state officials would take care "to apprise the community" of any government abuses that might be taking place. Also, nearby citizens could be trusted to warn those at a distance, "to sound the alarm when necessary, and to point out the actors in any pernicious project" (Hamilton, 1788).

Those of us who have experienced the impediments to the free flow of information brought by agency secrecy, or by general inattention or negligence, can have little confidence in Hamilton's view. His conception is so unrealistic – coming as it does from the most realistic of the framers – that it arouses suspicion it may have been put forward less to inform the public

than to relieve the government of having to do so. Once again, this debate was part of the larger argument between advocates of strong versus limited government (resolved in this case in favor of the Bill of Rights).

To sum up, the democratic protections involving structure, independence, and control are joined by the ultimate protection of reliance on an informed public. These protections bring us a framework of issues that together delineate the form and function, the substance and shape of the governmental need for evaluation.

A Framework of Governmental Issues Relevant to Evaluation

These issues, five in number, are as follows:

- The structure of government is fragmented, with functions and powers divided between federal and state levels and among executive, legislative, and judicial branches to guard against over-centralization of political power.
- Each level and branch, and each department or agency within a branch, is expected to protect its allotted powers and independence against incursion by others.
- Walls, suspicion and secrecy co-exist around these divisions in government and around the prescribed independence of branches and agencies.
- As a check on too much executive independence and too little transparency, Congress uses its oversight authority to scrutinize and control executive branch activities.
- Because the people need knowledge to serve as the "primary control on government," a critical task for the legislative and other branches is to inform citizens about the activities of their government.

This framework sets up the principles and context from which we can infer the place of evaluation in government. Evaluation sits at the heart of the continuing tension between the need to govern and to be distrustful of government. As Joseph Ellis writes, the debate about strong versus limited government "was not resolved so much as built into the fabric of our national identity. If that means the United States is founded on a contradiction, then so be it . . . . We have been living with it successfully for over two hundred years" (Ellis, 2002). Indeed, it is that very contradiction which establishes the legitimacy of evaluation.

The Governmental Need for Evaluation

We can deduce, then, that a democratic government such as the US needs evaluation for four purposes:

(1) To support congressional oversight;

(2) To build a stronger knowledge base for policy-making;

(3) To help agencies develop improved capabilities for policy and program planning, implementation and analysis of results, as well as greater openness and a more learning-oriented direction in their practice;

(4) To strengthen public information about government activities through dissemination of evaluation findings.

How exactly does evaluation serve these purposes?

Supporting Congressional Oversight

Congressman Bolling of Missouri used to speak about oversight with knowledge and wit:

> The problem I've always found with congressional oversight is that most of the oversight done by congressional committees is the wrong kind. It's the kind that looks back and decides how many mistakes other people you disagree with have made. It very seldom involves dealing

> with the skillful and knowledgeable monitoring of programs that committees have been responsible for. . . . I don't think that what I'm talking about is a little important; I think it's absolutely crucial. . . . In the long run, it's absolutely crucial to whether the democratic process will continue to work. This is the place where the democratic process in the United States is going to stand or fall, [based] on whether some day we are actually going to act as we talk on oversight. (Bolling, 1978)

Evaluation, of course, has a natural role in the "skillful and knowledgeable monitoring of programs," but also in the establishment of accountability within a jealous universe of competing priorities and prerogatives. Were legislative purposes appropriately served by a given executive branch policy or program? Were funds expended wisely, or at least efficiently? Evaluation in its retrospective mode can measure and account for the effectiveness and efficiency of an implemented policy or program. And it can also prospectively assess the likelihood, based on past performance and other data, that public funds will be usefully expended on some proposed new activity.

Indeed, in the Congressional Budget Control and Impoundment Act of 1974, Congress explicitly recognized the value of using evaluations in accountability assessments. As Senator Brock noted, "If these programs that we write, enact and administer are not subject to on-going oversight using evaluative techniques, they're just not going to do the job" (Brock, 1975).

Advancing Knowledge

Here the democratic purpose, derived from the information/education role of government, is to increase understanding about the factors underlying public problems, about the "fit" between these factors and the policy or program solutions proposed, and about the theory and logic (or lack thereof) that lie behind an implemented intervention. What are the different causes for homelessness? Why is a technology successful in Asia but not in Africa? Which policies and programs might best address problems of delinquency, based on which theory?

The importance of these evaluations to good government is incalculable. Without them, government takes high-stakes risks in moving ahead with large programs. We should not forget Patrick Moynihan, mourning the paucity of evidence brought to the plans and programs of the War on Poverty: "This is the essential fact. The government did not know what it was doing. It had a theory. Or rather a set of theories. Nothing more" (Moynihan, 1969). This lack of sound foundation not only hurt the success of the individual programs, but also caused the War on Poverty as a whole to be attacked on the grounds of imprudence and wastefulness (Moynihan, 1969).

Unfortunately, knowledge arrives according to its own timetable, and this timetable may not coincide with prevailing political winds. Still, the evidence brought by knowledge evaluations, when they are available, can be critical for the success of both ongoing and future government interventions.

Helping Agencies

Both accountability and knowledge evaluators draw on program data and other information available only within agencies. But for evaluators to collect and use that information, agencies must first allow them access to it. Also, if relevant actions are to be taken as a result of policy or program problems found by evaluators, it is often the agency managers who must take them. In other words, agencies need to possess an evaluative (and self-evaluative) capability and culture if accountability and knowledge evaluations are to be meaningful.

But agencies need evaluative capability for their own purposes, as we saw with PPBS, and also with the more recent Government Performance and Results Act (GPRA) of 1994. Alice Rivlin noted that, under GPRA, "every agency must present a clear picture of its goals, the links between those goals and how it spends its money, and its performance – what it produces for the American people. GPRA gives agencies the chance to tell their story in a credible way, to communicate the value of agency and program activities to OMB [the Office of Management and Budget], to Congress, and to the public" (Wholey, 1997).

Whatever the fate of GPRA, it seems clear that the ongoing development of institutional integrity and capability is as important a public need as the advancement of knowledge or the assessment of accountability. Indeed, from a practical viewpoint, given the evaluator's need for access to people and data, it will always be difficult to perform either knowledge or accountability evaluations without strong evaluative development within agencies.

Informing the People

Finally, insofar as evaluations are intended either to increase knowledge or to account for the activities of government, their publication contributes to the goal of an informed public. Evaluations report on the successes and failures of policies and programs. They speak to the probity and integrity of government practices. Importantly, they also enhance that transparency in government which allows accurate information to emerge. Evaluations that are credible, comprehensibly written, intelligently disseminated, and well reported by the press help make people aware of what is happening in government.

Development evaluations are less likely to be published or to be useful in informing the public. Such studies must walk a fine line between the twin goals of agency discretion and public service through improvements in practice. On the other hand, the information function served by accountability and knowledge evaluations means that, from the perspective of government need, the ultimate client or user of these evaluations is the public. This casts a slightly different light on the meaning of evaluation use. It also puts a premium on the independence of evaluators, on the accuracy and credibility of their product, on the free dissemination of their findings, and on the need for evaluators to fight barriers (such as denied access to data or unnecessary secrecy) which have the potential to distort their information.

Evaluation is not alone, of course, in the job of reporting about the government. Since the time of Madison and Jefferson, an enormous information industry has grown up – including a vastly expanded press, network and cable television, and the internet – whose predominant purpose is to inform the public. In the course of bringing the news, these media are often primary transmitters of evaluative information, frequently with evaluators briefing journalists on the scope and substance of their findings. There are also other forms of analysis, such as auditing or legal or budgetary analysis; while these may overlap with evaluation, they are not usually competitive. Evaluators are in a unique position to answer questions like, "What happened and why?" or "What difference did the policy or program make?" This work, properly disseminated, contributes powerfully both to transparency in government and to the overall democratic enterprise of informing the public.

In summary, there are four functions of government for which evaluation is needed (oversight, knowledge enhancement, agency development, and public information), demonstrating that determining merit or worth cannot be the sole purpose of public policy

evaluation. From a governmental viewpoint, evaluation serves: to obtain accurate information for one branch of government on the activities and accountability of another; to increase knowledge about the underlying bases for legislating, as well as for implementing legislative mandates, in a vast array of subject areas; to inform the public about the successes and failures of government endeavors; and to help develop within agencies the orientation toward challenge and improvement that allows evaluations to be done and their recommendations implemented.

The purposes listed above, then, are not markedly different from the ones derived by analysis of the 1976 symposium and the 1995 international conference in Vancouver (i.e., accountability, knowledge, and development). But the public information function must be added. Still, informing the public is not a fourth purpose of evaluation at the same level as the others; rather, it concerns the requirement for publication and dissemination of the findings of accountability and knowledge evaluations. Further, in a democracy the three purposes can be seen to flow from a universe of checks and balances, and to fit together in relation to that universe. It's true, of course, as noted earlier, that these purposes may not be exhaustive. But it's also true that the integrated response offered to the political and information needs of a democratic government by these three evaluation purposes confers upon them both authenticity and legitimacy.

As I re-examine the evaluations we did at the Program Evaluation and Methodology Division (PEMD) of the GAO, I see a body of work that is entirely consistent with these three purposes. During our 14 years, we were privileged to perform nearly 300 evaluations of all three types, mostly at congressional request. Over time, it became increasingly apparent to us that the support of congressional oversight is integrally tied to knowledge and development work, as well as to the dissemination of information to the public.

The Experience at PEMD

I arrived at GAO in 1980, invited by then-Comptroller General Elmer Staats, to lead the agency's new Institute for Program Evaluation. The GAO is a legislative agency established by the Budget and Accounting Act of 1921, which also created the Office of Management and Budget. This Act provided for considerable independence of the Comptroller General of GAO, who is appointed to a 15-year term by the President of the United States and can be removed only by resolution of both Houses of Congress. The original objectives of the agency had been mostly auditive and investigatory, but the Congressional Budget Control and Impoundment Act of 1974 broadened GAO's responsibilities in program evaluation while retaining its remarkable independence. During the period I worked there (1980–1994), I had many occasions to be grateful for that independence, and for the strength and determination with which the agency defended it. The unit I would organize and direct at GAO, which I staffed largely with social scientists, soon evolved from an "institute" to a regular GAO division in 1983, referred to as PEMD.

My original expectation, based on my fairly shallow grasp of congressional oversight needs at that time, was that the Congress would be more interested in accountability studies than in anything else. But as I look back over our work, I find at least as many examples of evaluations for development or knowledge as for accountability. It's also the case that many of our studies had more than one purpose, or else grew one from another (as in the case of accountability evaluations that had to be

followed by knowledge or development work to remedy some of the problems or gaps in information we had uncovered). Nevertheless, looking at the overall body of work, a particular purpose typically dominated the others in most of our evaluations.

Accountability Studies

Although some evaluators seem to believe that accountability studies are rare, at PEMD we did a great many of them. This surely had something to do with the strength and determination of the opposition in Congress to many policies of the Reagan and G.H.W. Bush Administrations. Accountability may be the terrible swift sword of democracy, but it requires a vigorous Congress to wield it.

Accountability studies are taken seriously in government. At very least, the general climate for doing an accountability study will be strained and its progress slow; often the context is more like the fog of war surrounding a pitched battle. We did accountability studies at PEMD for the Congress in almost every imaginable subject area (health, defense, public assistance, education, transportation, the environment, and more). Perhaps the most important one, in terms of size, scope, quality, and enduring agony, was our study of the United States' strategic nuclear triad.

Evaluation of the Nuclear Triad

In April, 1990, Chairman Dante Fascell, of the House Committee on Foreign Affairs, asked us to evaluate the major modernization programs proposed by the Department of Defense (DOD) for the strategic nuclear triad. "Triad" refers to the three methods of delivering nuclear retaliation: by land, sea, or air.

The Policy Question. The Chairman asked the following basic policy question: "In the face of the budget deficit and the changing context of East-West relations" (the Berlin Wall had fallen in November, 1989), "how can Congress best provide for the strategic security of the United States?" Fascell wrote further, "As the United States and the Soviet Union reach new agreements on strategic arms reductions, Congress will be making important decisions concerning the size and quality of the air/land/sea components of our strategic offensive force structure" (Fascell, 1992a). He asked us to focus on the effectiveness, cost, policy, and arms control implications for each component of the triad and any likely nuclear upgrades. The breathtaking nature of this request is apparent, but is even clearer considering that in 1990 the systems and their upgrades amounted to an estimated $350 billion. Difficulties also were immediately apparent, not least of which was the highly classified status of most of the documents needed for our work. (All of the following discussion is derived from unclassified source material.)

The Evaluation Questions, Design, and Measures. For the triad study, we decided to take some time to examine carefully the rationales underlying the various systems and upgrades before setting up our evaluation design. With the agreement of the Committee on Foreign Affairs, we translated the policy question into seven evaluation questions (around which we would structure separate reports, allowing us to have at least some findings ready for the coming congressional policy and budgetary debates). In summary, these questions involved assessing (1) the vulnerability of the sea leg's nuclear-powered ballistic missile submarines (SSBNS) and (2) the land leg's silo-based intercontinental ballistic missiles (ICBMs); (3) the relative effectiveness of ICBMs versus submarine-launched ballistic missiles (SLBMs); (4) the air leg's proposed upgrades in terms of improved capacity, relative

to existing systems; (5) the comparative costs of the proposed upgrades; (6) existing capabilities for addressing the threat posed by strategic relocatable targets (SRTs); and (7) strategic capabilities in France and the United Kingdom. The nuclear weapons systems and proposed upgrades we eventually included in the evaluation were the major ones (e.g., for the air leg, we examined the B-52G and B-52H, B-lB and B-2 bombers, as well as the ALCM, ACM, SRAM A, and SRAM II missiles). We knew we would have to assess all systems under a full range of threat scenarios, moving from total surprise attack to strategic warning.

Our basic design strategy was to develop a framework for comparison. Because we found no earlier comparative studies by DOD or others on which to build, we had to develop our own set of measurements. Our approach was to examine DOD's own conclusions about: the performance of the various triad weapons systems; the costs of the upgrades being proposed; and the size and nature of the Soviet threat. We then looked for the qualitative and quantitative evidence needed to support and validate these DOD conclusions.

The quantitative data came from a wide variety of data sources – about 250 major technical reports in all. We collected our qualitative data through interviews. We visited field sites, military commands and bases, as well as program offices. In addition to the special advisory board we constituted, we consulted military and civilian experts in a range of agencies, universities, and think tanks. In all, we did more than 200 extensive interviews.

To compare system costs across strategic program upgrades, our unit of analysis was the 30-year life-cycle (i.e., we included not just R&D and procurement, but also operations and support costs for every system). To compare system effectiveness, we used seven different measures: (1) survivability against both offensive and defensive threats, for both platforms and weapons (e.g., submarines or bombers and their missiles); (2) delivery system performance (i.e., accuracy, range and payload, which is the number of weapons carried by a single platform); (3) warhead yield and reliability (i.e., the probability that the warhead will detonate as intended); (4) weapon system reliability (i.e., the combined reliability of all the component processes, from platform launch to warhead detonation); (5) flexibility across a number of dimensions, including retargeting, recall, and impact on arms control; (6) communications (e.g., connectivity between command authority and platforms; and (7) responsiveness (i.e., alert rate and time-to-target).

In short, this was not a simple evaluation in its conception, in its execution, or in its logistics. In addition, its accountability character raised hackles at DOD. Still, we managed to study performance and cost within weapons systems, between existing weapons systems and their proposed upgrades, across weapons systems within a leg, and across legs, thanks to a stellar staff that included Kwai-Cheung Chan, Brett Haan, Rob Orwin, Jim Solomon, Jonathan Turnin, and Winslow Wheeler. All of us, I think, realized the importance of this evaluation and felt quite comfortable when the time came to set down our findings and conclusions.

Study Conclusions. Our first conclusion was that, on balance, the sea leg and its weapons systems emerged as the most cost-effective of the legs and systems. Second, the air leg continued to have a vital role in the triad context. Because strategic bombers are recallable (as missiles are not), and because they are virtually incapable of effecting a surprise attack, they add a critically important stabilizing character to the overall nuclear

force. We also concluded that, within DOD, there was a troubling dearth of the comparative studies needed to show whether a proposed system is justified in terms of the threat it faces, its performance capabilities vis-à-vis other systems, and its relative costs.

From an accountability perspective, this evaluation was successful in many ways. It showed Members of Congress where cost savings could be achieved, if they wanted to make them, and it confirmed the quality of the sea leg's platforms and missiles. It justified the continued existence and modernization of the air leg. And it illuminated DOD's planning and program processes in such a way as to reveal major weaknesses in congressional oversight.

Indeed, within DOD, we found many instances of dubious support for claims of weapons systems' high performance; insufficient and often unrealistic testing; understated cost; incomplete or unrepresentative reporting; lack of systematic comparison of new systems against the systems they were to replace; and unconvincing rationales for their development in the first place. Where mature programs were concerned, on the other hand, we found that their performance was often understated and inappropriate claims of obsolescence were made.

The study was very difficult to carry out, which is hardly unusual in accountability studies. DOD resisted our efforts, and we often had to work with missing or sketchy or unconvincing data in critical areas of the evaluation. Nevertheless, the work clearly corresponds to the framers' vision of checks and balances and, especially, oversight and accountability. It corresponded less well, however, with their idea of providing information about the workings' of government to the citizenry.

Informing the People. Here we ran into trouble from two directions. First, DOD decided to classify all our reports in their entirety (a departure from past experience where only partial classification had been the norm). This meant the whole evaluation would be inaccessible to the press and the public. Second, the study's sponsor, Chairman Fascell, announced his intention to retire from Congress at the end of the year, apparently signifying there would be no hearing either. So it seemed the triad evaluation would be publicly inaccessible. Of course, the Congress itself would be informed, since all members had access to the classified briefings arranged by the Committee on Foreign Affairs (Fascell, 1992a).

After three months of steady negotiation, DOD officials agreed to an unclassified summary statement, 15 pages long. This had the heavy burden of standing in for several thousand pages of text. We then published all of our reports in classified format (USGAO, 1992a–h) and sent Chairman Fascell the unclassified summary statement, which he published in the Congressional Record (Fascell, 1992b).

Press interest began slowly, mostly by specialized trade papers like *Defense Week* or *The Navy Times*. Coverage spread across the country over the next few months, with some serious debate and editorial commentary in the major newspapers. Then, out of the blue, I received a letter from Chairman John Glenn of the Senate Committee on Governmental Affairs, quoting extensively from our summary statement. He invited me to testify at a hearing the Committee wanted to prepare on our report (to look into the management of government and its effectiveness, rather than at strategic nuclear retaliatory systems).

When this hearing occurred, in June, it triggered a nationwide explosion of interest in our evaluation, for and against, in the best tradition of defense debate. From a public information viewpoint, it would be hard to imagine better coverage. However, the hearing also

changed the way our study was perceived, to some degree, because its focus was on oversight and accountability issues rather than the policy and budgetary concerns that had inspired the evaluations. This did mean a new look at our data, and perhaps a slightly different presentation of the findings, but it distorted none of the findings and brought a richer public debate than would have occurred otherwise.

Two final points on the triad evaluation. First, despite the fact that DOD differed strenuously with us on many findings, the first Bush Administration's actions in fact mirrored some of the major recommendations in early drafts of our general summary report. To take but two of several examples, we questioned the need for either SICBM or Peacekeeper Rail Garrison: Both were cancelled by President Bush. We noticed that insufficient tests of the Minuteman IIs precluded any confidence in estimates of the missile's reliability: President Bush decommissioned the entire Minuteman II force. Second, by the time of Senator Glenn's hearing, the Clinton Administration had come to power, Les Aspin had replaced Dick Cheney, and Deputy Secretary of Defense William Perry testified along with me at the hearing. I was stunned to hear Perry say: "Now let me comment briefly on the GAO report. It is a very formidable, substantial undertaking, and it will be used – it is being used – as a very important input to our own planning of strategic forces. On balance, we think it is an excellent report, objectively done, and agree with most of the conclusions in the report" (Perry, 1993).

Drawing on the triad study and other experience, I conclude that accountability evaluations are not merely needed; they are absolutely essential for the effectiveness of congressional oversight. They can be performed, no matter how resistant the agency; and they may even, with help and luck, have extremely happy outcomes.

Development Studies

As I mentioned earlier, we often did developmental work in PEMD to help agencies strengthen their evaluation capabilities after one of our own studies showed that there were, say, technical skill areas or data problems that needed attention. In looking at how DOD assessed operational effectiveness using computer simulations, for example, we had observed problems in evaluating the credibility of results. So we developed a framework for use by DOD analysts to aid in assessing simulation strengths and weaknesses (USGAO, 1987a).

In the same way, a 1986 study sought to determine whether the construction grants program (on which $39 billion had been spent) was making improvements in the quality of the nation's waters. It turned up a staggering absence of effectiveness evaluations at EPA. We therefore undertook some developmental research to come up with guidelines that could help EPA determine whether the program was doing any good. Our method used only the data and software already available within EPA, and we not only developed the method but also tested it in case studies that measured the effects of upgrades in four wastewater treatment plants (USGAO, 1986a).

Having criticized the Department of Health and Human Services' process of initiating new rules for Medicare (in this case, the prospective payment system changes involving fixed per-case payments for diagnosis-related groups) without evaluating them, we developed for agency use a two-part evaluation plan for determining the effects of the prospective payment system on patients in post-hospital care (USGAO, 1986b). This effort began a close PEMD relationship with Senator John Heinz, Chairman of the Senate Special Committee on Aging, and started us on a continuing series of reports about the effectiveness of various

medical treatments, processes and practices. In particular, this body of work, led by Lois-ellin Datta and supported by Senators John Heinz, George Mitchell, John Glenn, and David Pryor, helped bring about the creation in 1989 of the Agency for Health Care Policy and Research (now called the Agency for Health Quality Research), which has since made a distinguished contribution to the spread of effectiveness evaluations in medicine.

Another developmental endeavor of ours (begun in 1982, based on what we were coming to see as a basic need in some agencies) was the preparation of papers on evaluation methodology. These were essentially "how-to" reports, and they included volumes on evaluation design, questionnaire development, statistical sampling, structured interview techniques, quantitative data analysis, case study evaluation, and methods for synthesis research. Although we had designed these for agency use and published them with discretion, we were soon overwhelmed by requests for these papers, not only from agencies (both national and international), but also from universities here and abroad.

Finally, we put steady pressure: on agencies, to develop their evaluation capabilities and report on the effectiveness of their programs (USGAO, 1982, 1987b); on the Office of Management and Budget to demand strong evaluations from the agencies (USGAO, 1990); on individual congressional oversight committees, to insist on better evaluations from the agencies (USGAO, 1988a, 1991, 1993); and on the Congress as a whole to recognize the importance, for both legislative policy-making and oversight, of a thoughtful, vigorous, and courageous evaluation function in the executive branch (USGAO, 1988b, 1992i).

Overall, we spent considerable time and effort in PEMD doing development work, not just in pursuit of some abstract idea of general excellence in government, but because it was an integral part of every other kind of evaluation. More profoundly, perhaps, our experience with the triad study shows what a forceful tool accountability evaluations can be in opening agencies to public scrutiny. But sometimes force equates to overkill. Developmental evaluations act more softly to achieve transparency, and they encounter fewer obstacles to acceptance and use.

Knowledge Studies

Public policy evaluations also need to fulfill the Jeffersonian concept of bringing a better knowledge base to government. These kinds of evaluations not only examine the effects of programs, policies, and practices; they also assess their underlying assumptions, that is, those beliefs enshrined in the hearts and minds of officials and practitioners that may not stand up under examination. This testing of underlying assumptions turns out to be a very important part of knowledge evaluation. It is that component of skeptical observation Max Weber called *Entzauberung*, the demystification of dominant ideas and theories by means of empirical testing and analysis.

We did this kind of work often for congressional committees. In fact, a good example can be found in my testimony on the triad evaluation for Senator Glenn. A matrix at the end shows prevailing assumptions or beliefs about each of the triad legs (on issues like performance, vulnerability, and ease of communications) and compares them with the findings that emerged from a serious look at the data (Senate Committee on Governmental Affairs, 1993). Some studies, however, were entirely Weberian in nature. The one I discuss here is a 1987 evaluation (led by Richard Barnes and Roy Jones) which examined the idea that raising the minimum drinking age improves highway safety (USGAO, 1987c).

Evaluation of the Effects of Drinking Age Laws

In October, 1985, we received a letter from Congressman James Oberstar, Chairman of the Subcommittee on Oversight for the Committee on Public Works and Transportation. The Subcommittee wanted to know whether existing research supported the idea that raising minimum drinking-age laws improved highway safety. The Subcommittee noted "the frequency with which evaluations that are submitted for the record support opposing conclusions, even though they use similar data bases and assumptions" (USGAO, 1987c). This was a fascinating question, with an *Entzauberung* of its own, demolishing the hallowed belief that legislators are uninterested in research data and methods. The Subcommittee was actually asking us to tell them "what constitutes a 'good' evaluation."

The Policy Question. Of course, the Subcommittee came to us not only because they wanted to acquire knowledge, but also because they were in the middle of a political battle. This battle was part of a much longer war, with recent roots in the 1933 Repeal of Prohibition, but going back to the unresolved constitutional debate about state versus federal power. Repeal granted the states substantial power to regulate the purchase and possession of liquor, though The Highway Safety Act of 1966 returned some of that power to the federal government. Then, in the early 1970s, the 26th Amendment extended the right to vote to 18-year-olds, prompting 29 states to lower their minimum drinking age from 21 to 18.

By the 1980s, a documented increase in alcohol-related fatal crashes among younger drivers led to congressional enactment of the 1984 Uniform National Minimum Drinking-Age Law. This law included a controversial provision reducing the amount of federal highway aid to states that did not enact a legal minimum drinking age of 21. Although many states then passed "age 21" laws to avoid losing federal highway funds, others remained reluctant to do so. In September 1984, South Dakota brought suit against the Secretary of Transportation, asking that the Uniform National Drinking-Age Law be declared unconstitutional on the grounds that it violated states' rights. South Dakota also asserted there was no scientific evidence showing that raising the minimum age reduced alcohol-related traffic accidents. This lawsuit was being watched with passionate interest by lobbying groups on both sides of the issue. Many state legislatures, intensively jawboned by the liquor and restaurant lobbies and by college students, reportedly were planning again to repeal their age-21 laws.

This was the context in which we were asked to do our study. The policy question was "Does raising the minimum drinking age improve highway safety?" A corollary question was of equal policy importance: "How good is the research supporting each side of the issue?" The ability to distinguish a "good" evaluation from a "bad" one had taken on immense political significance.

The Evaluation Question, Methods, and Measures. We derived a number of evaluation questions from our discussions with the Subcommittee, but the main one was more or less the same as the policy question: Does raising the minimum drinking age result in a change in alcohol-related motor vehicle fatalities, injuries and crashes among the age group affected by the law? (Other questions related to effects on consumption, as well as displacement effects in other age groups.)

Given the large body of evaluation studies to be examined, we decided, in accord with the Subcommittee, to use the evaluation synthesis method, which had been one of the fruits of our development work (USGAO, 1983). The

literature search uncovered 400 documents, of which 82 were evaluations of the effects of changing the minimum drinking age. Of these 82, 33 were directed at lowering the drinking age, leaving us with 49 evaluations as the basis for beginning our work.

We formed a review panel, including independent experts along with our staff, to develop rating criteria and review studies of direct relevance to the evaluation questions. The panel developed criteria for two generic types of studies: cross-sectional (comparing two or more defined groups at a single point in time) and pre–post (comparing groups at two or more points in time). We rated all studies in terms of five criteria: (1) the existence and adequacy of comparison groups; (2) the source data used; (3) the appropriateness and comparability of measures used; (4) the appropriateness of methods for taking chance into account; and (5) the extent to which a study controlled for other factors and provided quantitative measures of difference. For pre–post evaluations, we also looked for: (6) data that were comparable; and (7) controls for the non-independence of measures (autocorrelation, seasonality, and the like).

To assess the quality of the 49 evaluations, three raters reviewed each study independently, and then met to reconcile differences in individual ratings of "acceptable," "questionable," or "unacceptable." An unacceptable rating was typically given to evaluations failing to meet two or more criteria. Among the 49 studies, 28 thus dropped out, leaving us with 21 on which we based our findings. Every step in this process was carefully documented (because of past experience with a synthesis in which some evaluators were unhappy with our ratings).

Study Conclusions. We finished our study in the summer of 1986. The most important conclusion was that raising the drinking age does have direct, sizable effects on reducing traffic accidents among 18- to 20-year-olds, on average, across the states. We found statistically significant reductions ranging from 5% to 28% in "driver-fatal" crashes.

The Subcommittee held a hearing in September 1986, in which we were praised for our industry and grilled on our methodology (House Committee on Public Works and Transportation, 1986). Why did we throw out this evaluation or that one, for example, which found no effect? What is a cause-and-effect question and why do you need comparison groups to answer it? What are the advantages of cross-sectional versus time-series types of studies?

The four-hour hearing delved into questions of methodology and the foundations of conclusions to a degree I would never see again in any other hearing. Overall, the report was well received and the whole experience was positive in many ways: using a new method successfully, working well and easily with the Subcommittee, enduring little or no political pressure during the course of the work, answering a major policy question, and – critically important in this case – delivering the work on time. Also, with respect to disseminating the findings publicly, we turned out to be luckier than we could ever have imagined.

Informing the People. The hearing, transmitted nationwide on television (C-Span broadcast it twice a day for more than a week), brought in a large response from viewers. But the debate was far from over. While we were doing our work, the South Dakota lawsuit was moving up to the Supreme Court calendar for the October term of 1986. When the Supreme Court issued its ruling in 1987, it went against South Dakota. Our congressional hearing of September 1986 was one of the "legislative materials" used by the Court, and our work was examined both for its conclusion on the

effect of raising the minimum drinking age, and for its judgments about the methodological soundness of the various studies (Supreme Court of the United States, 1987).

The press and television coverage was extensive in every state. Editorials proliferated. As a result of the Supreme Court decision, by 1989 all 50 states had a minimum drinking age of 21. The Department of Transportation credited our evaluation with saving an estimated 1000 youthful lives in 1988, and a follow-on study of Tennessee showed a 38% decline in the death rate among 19- to 20-year-olds as a result of the legal increase in the drinking age (USGAO, 1989).

In summary, this knowledge evaluation fits well into the Jeffersonian vision of research as support for policy-making – though Jefferson probably would have been less than happy about the defeat handed to state sovereignty. Still, there is little doubt that the dissemination of information was exemplary. In this case, at least, the people knew what their government was doing.

Overall, the cumulative experience – not just with the illustrative cases presented above, but with accountability, development, and knowledge studies generally, as well as the evidence furnished by national symposia and by the international Vancouver conference – shows clearly that these different kinds of evaluations are all needed, and that together they derive legitimacy from their function of support to open democratic government. But to what degree is this experience transferable to countries with different types of government?

Issues of Generalizability

In considering whether these questions of form, function, and legitimacy are applicable to other nations, two issues appear germane:

- whether the nation is a democracy;
- if it is, whether its government is structured so as to oblige officials and politicians to tolerate dissent.

My own experience of evaluation in countries other than the US is not as extensive as I would wish, but what I have grasped – in working with NATO, the European Commission, the World Bank, and national audit offices of many authoritarian countries – is that non-democratic societies furnish sometimes insuperable obstacles to serious evaluation. This is largely because strong studies can threaten regimes. Evaluations showing improvement in the GDP or in the unemployment rate, for example, tend to strengthen the position of those in power; conversely, unfavorable conclusions may bring calls for change and/or reform. Where political controversies are the norm, this is merely a passing problem. But when there are no free newspapers, only a single political party, and a belief that citizens exist to serve the state (not vice versa), then calls for change and reform are usually unacceptable to those in power.

In one case (China, after the Hundred Flowers of 1958), the routine development of a statistical series brought down the agency that generated it. When critics of the new regime were able to point out that per capita income had undergone a long decline after the advent of the People's Republic in 1949, the carefully run State Statistical Bureau, which had developed both the data and their evaluation, fell into disfavor. For many years after 1958, aggregated data were not regularly produced in China (Chelimsky, 1977a; Chen & Galenson, 1969).

When I gave a talk in Beijing about evaluation practice (in December, 1988, under the auspices of the Chinese Auditor General, Lu Peijian), I was asked by one of the state planning economists present what I thought the chances

were for developing a PEMD-like evaluation shop in China. I responded with another question: Could an evaluator really examine the short- and long-term effects of the single-child-per-family policy in China, without putting his or her freedom in jeopardy? Those seemed like pretty high stakes for evaluators. Yet, as we learned at the Vancouver conference, China is moving toward a national evaluation function.

Again, in Colombia (which was creating a legislatively mandated system of evaluation in 1991, when I spoke at their Santa Marta conference), I was startled to hear how Colombian justice officials planned to use evaluations: to prosecute and punish evaluators when ex-post results didn't square with ex-ante estimates. Such a mismatch, they claimed, showed the evaluators had "either cheated or made mistakes" in their analysis. In vain, I pointed out that this might create something of a chill in the evaluation profession if you could end up in jail for making an error in your cost-benefit calculations.

The truth is that authoritarian leaders don't, as a rule, think they have much to learn from research; the ability to express opinions or to publish is in short supply in their countries; and Machiavellian lion-and-fox tactics don't typically accommodate a lot of challenge from social scientists.

On the other hand, I have a more nuanced view today than I did in 1988. Thinking about the work in PEMD, I remember how much courage it took to pursue certain kinds of studies (like that of the nuclear triad). But would we really have wanted to tackle an evaluation of, say, the short- and long-term effects of Roe versus Wade (the Supreme Court decision that legalized abortion)? The ideological nature of the abortion debate and the consequent doubt that anyone could dispassionately examine the data from such an evaluation, make it unlikely that I would have recommended doing it. Still, if we had been asked by the Congress, we would certainly have made an effort to find an entry into the issue that could have illuminated some part of the debate. But we were not asked, and that is the essential point here: Even in a democracy, some questions simply will not be posed. This may be because the legislature is intimidated by the executive, or because opinion is too evenly divided to make winning a political victory possible, or because a policy is so deeply entrenched that re-examining it seems like a waste of time and money, or because evaluating outcomes risks uncovering the shakiness of assumptions in some policies or programs and embarrassing the officials who believed in them, or for whatever reason.

But if this is true, then the extreme case – that is, the ability to do evaluations on the most highly sensitive of subjects – is not the best measure of the applicability of an evaluation function. After all, most evaluation situations are necessarily imperfect to some degree, given the unequal balance of power between researchers and politicians. So a better measure of applicability might be whether there is a clear recognition in government of the need to improve institutions, along with the will to do it. The goal of "better value for money," for example, is certainly a worthwhile accountability purpose, and has for many years inspired the evaluation of public policies and programs throughout the world. However, the more profound knowledge and accountability purposes of evaluation, with their essential characteristic of following wherever the evidence leads, are likely to be dangerous and difficult under authoritarian regimes. Further, when regimes consider themselves to be above the law, they feel no special compulsion to be accountable to the people.

In democracies, however, all evaluative things are possible, although they may not always materialize optimally. In France, for

example, the Cour des Comptes (the French national audit agency) has done some excellent evaluations under the leadership of Pierre Joxe. I recall, in particular, a study presenting strong supporting data on the growing disparities between policy and practice in the management of France's highways (Cour des Comptes, 1992). Another impressive piece of work was an evaluation (performed under France's National Commission on Evaluation) of the RMI welfare program. RMI stands for "Revenu Minimum de l'Insertion," or the minimum income believed necessary to facilitate the integration of disadvantaged people into the society. This was a complex study seeking to measure not only what proportion of the real need had been addressed by the program (involving problems of identifying hidden populations, and of accounting for different concepts of need), but also to what degree improvements in integration had occurred, as viewed by the program's clients (Commission Nationale d'Evaluation, 1992).

Although the capacity and courage certainly exists in France for performing high-quality evaluations, there is a political problem that restricts the number and funding of evaluations likely to be done. That is the weakness of France's legislature. Power, largely centralized in Paris, has always resided preponderantly in the executive. Vigorous parliamentary oversight of government initiatives, never a very realistic possibility, was diminished even further by the French constitutional changes of 1958. However, the Cour des Comptes, with its considerable independence and its power to plan its own evaluations, can usually be counted on to do the kinds of studies needed to serve democratic government in France.

In Switzerland, a similar capacity exists, but again a political problem slows the development of the evaluation function: the extraordinarily lengthy process that precedes the implementation of government initiatives. Because of the division of powers between the individual cantons and the Swiss Confederation, debate may take many years to be sufficiently resolved to permit legislation. As a result, when a policy or program is finally agreed to and put in practice, it may be almost impossible, politically, to pose serious questions about actual outcomes and effectiveness.

In the United Kingdom, evaluators not only have a strong evaluation capability, but also political institutions that have supported and facilitated the development of national models to assess government services and performance. Mawhood (1997) described the large-scale efforts of the "new public management" to examine programs and services by "setting specific output measures, performance indicators and targets," using these to evaluate achievements, and then following up, year after year. In Australia, a reasonably workable system has been established that links evaluation to the budget process (something we tried for many years to do in the US, without success). And Canada has developed a multi-pronged evaluation function which has much in common with that of the US.

Finally, the World Bank (under the evaluative leadership, first, of Mervyn Wiener, and later of Yves Rovani and Robert Picciotto) has promulgated a policy of evaluation for institutional improvement among all its nation-members, which makes very good sense as an evaluative common denominator in democratic and non-democratic governments alike. Also, the European Commission, the International Atomic Energy Agency (whose evaluation function used to be headed by David Kay), the International Monetary Fund, and many international foundations have adopted sound principles in evaluation design and practice, and have produced some thoughtful and courageous studies.

It seems reasonable to believe, then, that the evaluative work we did in PEMD – in support of

knowledge, oversight, agency development, and public information – is largely generalizable to other democratic countries, when there is a national appetite to learn about what the government is doing, and when evaluators can count on adequate independence and protection for their work. In non-democratic societies, however, the evaluative menu is more likely to be restricted to "value for money" studies and institutional improvement, rather than knowledge and accountability evaluations that question the true bases for policies and programs. Still, the important point is that at least some useful evaluations can be done almost anywhere in the world. Indeed, the Vancouver conference gives tangible evidence that this has already begun to happen.

A Final Thought on Change and the Political Climate for Evaluation

In the United States, we continue to live under the same carefully balanced, 200+-year-old government which evaluation serves in so many ways. Quite a few things have changed in our thinking, but that balance still prevails. Other things that are important to evaluation have not changed at all, and this is as true internationally as it is in America. Evaluators still face closed administrations, and the more difficult it is to deal with agency walls and secrecy, the more important it is to do so, because people still need to know about the inner workings of their government. The war among different public sectors rages on (in varying degrees and distributions, of course), but the sectors also work together when the focus of their battle shifts, and this furnishes opportunities for evaluators (as we saw in our evaluation of drinking-age laws). Ineffective policies and programs continue to be implemented, poorly tested assumptions didn't die with the War on Poverty, and wasteful spending is always with us, so the playing field for evaluation widens each day.

Yet evaluation is a fragile reed to send up against all those giant oaks – against entire agencies sometimes, as happened in our nuclear triad study – and evaluators need to be ingenious, lucky, and much better protected than they currently are if they are to survive in any government. Alas, we make a lot of enemies, although we try hard not to, and our bosses sometimes prefer being "part of the team" to defending the independence of evaluators and their often unwelcome and inconvenient findings.

Nevertheless, for most agencies, for legislatures, and for citizens everywhere, evaluation is a pretty good bargain, accomplishing a remarkable amount with just a few people. We help to keep a healthy balance of power across sectors and agencies. We improve the government's products and services. We hold down expenditures (in Fiscal Year 1992, for example, GAO's work saved taxpayers more than $36 billion). And, with the aid of the press, we tell the public the results of government initiatives.

It goes without saying that evaluators don't do all that alone. In democratic societies, we can count on institutions, like a determined legislature, service-oriented agencies, and a knowledgeable, persevering, imperturbable press. But we count even more on a watchful population that, in and of itself, creates the right climate for evaluation.

Some have expressed concern today about public distrust of government. This is seen as a very bad thing in France and Germany as well as in the United Kingdom and the US. Polls have shown, for example, that the American electorate is disenchanted with politics and politicians, and that citizen confidence in government is low. But there is nothing new about this, as we realize from reading Dickens' *American Notes* of 1842: "One great blemish,"

he wrote, "in the popular mind of America, and the prolific parent of an innumerable brood of evils, is Universal Distrust. Yet the American citizen plumes himself upon this spirit, even when he is sufficiently dispassionate to perceive the ruin it works, and will often adduce it . . . as an instance of the great sagacity and acuteness of the people, and their superior shrewdness and independence."

What Dickens discounted, however, in his irritation with "universal distrust" is that it was precisely this distrust on which the American framers relied to control the excesses of "unseen rulers in distant places." And it was this same distrust which brought to the British their Magna Carta and 1689 Bill of Rights, to the French, their Declaration of the Rights of Man, and to the US, its government of checks and balances. It is distrust, once again, that generates the deepest constituency for evaluation. After all, a trusting population is not likely to ask searching questions about cozy arrangements, wasted resources, or data-free policies in government.

Put another way, public distrust is, by its function and modalities, a positive, not a negative element in a democratic society. In that larger sense, the search for political balance and open government makes evaluators of us all.

References

Bolling, R. (1978). In *Proceedings of a Three-day Workshop on Congressional Oversight*, US House of Representatives, pp. 29–30.

Bradshaw, K. & Pring, D. (1981). *Parliament and Congress*. London: Quartet Books, pp. 479, 358.

Brock, B. (1975). In Chelimsky, E., Program evaluation and appropriate governmental change. *Annals, American Academy of Political Science*, 466 (March 1983), 106.

Bronowski, J. & Mazlish, B. (1993). *The Western Intellectual Tradition*, Barnes and Noble, p. 390.

Chelimsky, E. (1977a). The need for better data. *Evaluation Quarterly*, 1(3), 439–440.

Chelimsky, E. (1977b). *Proceedings of a Symposium on the Use of Evaluation by Federal Agencies*, Vols. I and II, MITRE Corp., VA.

Chelimsky, E. (1978). Differing perspectives of evaluation. *New Directions for Program Evaluation*, No. 2, 1–18.

Chelimsky, E. (1985). Budget cuts, data and evaluation. *Society*, 22(3), 67.

Chelimsky, E. (1996). Thoughts for a new evaluation society. *Evaluation*, 3(1), 97–109.

Chelimsky, E. & Shadish, W. R., Jr. (1997). *Evaluation for the 21st Century*. Sage Publications, pp. 10–18.

Chen, N. R. & Galenson, W. (1969). *The Chinese Economy Under Communism 1969*. Aldine Publishing Co, pp. 159–161.

Commission Nationale d'Evaluation (1992). *RMI: Le Pari de l'Insertion* (2 volumes). Paris: La Documentation Francaise.

Cour des Comptes (1992). La Politique Routiere et Autoroutiere, Evaluation de la Gestion du Reseau National (Report to the President of the Republic, May 1992).

Ellis, J. J. (2002). *Founding Brothers: The revolutionary generation*. Random House, Vintage Books, pp. 7, 9, 15, 16.

Fascell, D. B. (1992a). *Congressional Record*, July 21, p. E2179.

Fascell, D. B. (1992b). *Congressional Record*, September 29, pp. H9861–9864.

Hamilton, A. (1788). The Federalist, No. 84: On a bill of rights and freedom of the press." *Selected Federalist Papers*, Dover Publications, 2001, pp. 199–200.

Hicks, J., Mowry, G., & Burke, R. (1964). *The Federal Union*. Houghton Mifflin, p. 314.

House Committee on Public Works and Transportation (1986). Hearing Before the Subcommittee on Oversight, The National Minimum Drinking Age Law, USGPO, pp. 1–40.

Jefferson, T. (1946). *Thomas Jefferson on Democracy* (ed. S. K. Padover). Mentor Books, New American Library, pp. 44, 89, 90, 92–97.

Jewell, M. E. & Patterson, S. C. (1966). *The Legislative Process in the United States*. Random House, p. 131.

Madison, J. (1788). The Federalist No. 51: The structure of the government must furnish the proper checks and balances between the different departments. *Selected Federalist Papers*. Dover Publications, 2001, pp. 120–122.

Mawhood, C. (1997). Performance measurement in the United Kingdom, 1985–1995. In E. Chelimsky & W. R. Shadish, Jr. (eds) *Evaluation for the 21st Century*. Sage, pp. 134–144.

Moynihan, D. P. (1969). *Maximum Feasible Misunderstanding*. The Free Press, MacMillan Company, p. 170.

Patton, M. Q. (1996). *A world larger than formative and summative. Evaluation Practice*, 17(2), 132–142.

Perry, W. (1993). Evaluation of the U.S. strategic nuclear triad. Hearing Before the Committee on Governmental Affairs, U.S. Senate, June 10, USGPO, p. 16.

Rhoads, S. E. (1978). Economists and policy analysis. *Public Administration Review*, March/April, p. 114.

Rossi, P. H. & Freeman, H. E. (1985). *Evaluation: A systematic approach*, third edition. Sage Publications, pp. 321–356.

Senate Committee on Governmental Affairs (1993). Hearing on the Evaluation of the U.S. Strategic Nuclear Triad Testimony of Eleanor Chelimsky, USGPO, pp. 39–48.

Supreme Court of the United States (1987). The State of South Dakota Versus The Honorable Elizabeth H. Dole, Secretary, U.S. Department of Transportation, March 16.

USGAO (1982). A Profile of Federal Program Evaluation Activities (GAO/IPE, Special Study I, September).

USGAO (1983). The Evaluation Synthesis (GAO/PEMD-10.1.2, Revised March 1992).

USGAO (1986a). Water Quality: An Evaluation Method for the Construction Grants Program Vol. I, Methodology, Vol. II, Case Studies (GAO/PEMD-87-4B).

USGAO (1986b). Post-Hospital Care: Efforts to Evaluate Medicare Prospective Payment Effects Are Insufficient (GAO/PEMD-86-10).

USGAO (1987a). DOD Simulations: Improved Assessment Procedures Would Increase the Credibility of Results (GAO/PEMD-88-3).

USGAO (1987b). Federal Evaluation: Fewer Units, Reduced Resources, Different Studies from 1980 (GAO/PEMD-87-9).

USGAO (1987c). Drinking Age Laws: An Evaluation Synthesis of Their Impact on Highway Safety (GAO/PEMD-87-10).

USGAO (1988a). Children's Programs: A Comparative Evaluation Frame work and Five Illustrations (GAO/PEMD-88-28BR).

USGAO (1988b). Program Evaluation Issues (GAO/OCG-89-8TR).

USGAO (1989). *GAO Management News*, 16(36).

USGAO (1990). Improving Program Evaluation in the Executive Branch: What OMB Could Do (GAO/PEMD-90-19).

USGAO (1991). Child Support Enforcement: A Framework for Evaluating Costs, Benefits and Effects (GAO/PEMD-91-6).

USGAO (1992a). U.S. Strategic Triad: Vulnerability of Strategic Ballistic Missile Nuclear Submarines (C-GAO/PEMD-92-1).

USGAO (1992b). U.S. Strategic Triad: ICBM Vulnerability (C-GAO/PEMD-92-2).

USGAO (1992c). U.S. Strategic Triad: A Comparison of ICBMs and SLBMs (C-GAO/PEMD-92-3).

USGAO (1992d). U.S. Strategic Triad; Modernizing Strategic Bombers and Their Missiles (C-GAO/PEMD-92-4).

USGAO (1992e). U.S. Strategic Triad: Strategic Relocatable Targets (C-GAO/PEMD-92-5).

USGAO (1992f). U.S. Strategic Triad: Costs and Uncertainties of Proposed Upgrades (C-GAO/PEMD-92-6).

USGAO (1992g). U.S. Strategic Triad: Current Status, Modernization Plans and Doctrine of British and French Nuclear Forces (C-GAO/PEMD-92-7).

USGAO (1992h). U.S. Strategic Triad: Final Report and Recommendations (C-GAO/PEMD-92-8).

USGAO (1992i). Program Evaluation Issues (GAO/OCG-93-6TR).

USGAO (1993). Public Health Service: Evaluation Set-Aside Has Not Realized Its Potential to Inform the Congress (GAO/PEMD-93-13).

Wholey, J. S. (1997). Trends in performance measurement: challenges for evaluators. In E. Chelimsky & W. R. Shadish, Jr. (eds) *Evaluation for the 21st Century*. Sage Publications, p. 128.