The Role of Experience in Formulating Theories of Evaluation Practice

ELEANOR CHELIMSKY

ABSTRACT

Although it is widely believed that evaluation theory is practice-based, it is rare to see any systematic examination of the fit between the two. This address, given at AEA's most recent meeting in San Diego, uses a data base of about 300 published evaluations to compare some of the lessons learned in recent practice against some of our more hallowed theoretical precepts.

It is a great pleasure to be asked to give this paper at this particular AEA conference which focuses more on theories of practice than we have done in the past. Lewin wrote 60 years ago that without theory, we might as well be blind because we would lack "that element which alone is able to organize facts and give direction to research." Lewin said he had in mind theory "which is empirical and not speculative," and whose hypotheses are closely related to the data and experience brought by practice. (Lewin, 1936, p. 4)

In this paper, I want to vary my usual agonized reporting from the trenches and speak to you, in Lewin's terms, about how and whether some of the theoretical formulations we make about practice do, in fact, relate to recent experience in the field. Naturally, I cannot talk about all experience or all theory, only about my own experience and those aspects of theory that have been important to me.

But I should point out that my discussion today derives from at least two views of practice: that of the Executive Branch evaluator (who works for agencies with vertical management structures where most political power is concentrated at the top), and that of the Legislative Branch evaluator (who works for individual committees in a horizontal parliamentary structure, where power is both effective at the committee level, and also shared across committees). I said, "at least two views" because within those two, others are hidden. For example, I have been an "inside" evaluator, working developmentally to build evaluative capability at NATO, the Law Enforcement Assistance Administration, the United States Gen-

Eleanor Chelimsky • 4711 Village Drive, Fairfax, VA 22030.

eral Accounting Office (GAO), and the World Bank; and I have been an “outside” evaluator conducting knowledge and accountability evaluations, first at the Mitre Corporation and then at the GAO. There I worked for the Congress as head of GAO’s Program Evaluation and Methodology Division (PEMD), and ended up producing nearly 300 evaluations. It is the latter body of work on which I will draw principally here, so I will be speaking more from knowledge and accountability perspectives than from a developmental one.

In short, the experience of practice that serves as my data base here is not exhaustive, but it contrasts favorably with other data bases, and it is certainly considerable enough to be compared, in Kuhnian fashion, against the theories and principles of practice that we count on to organize and guide our work (Kuhn, 1962, p. 77). So which, then, are those theories and principles?

When I moved to the GAO from the Mitre Corporation in 1980, the theoretical issue that most concerned me was use: that is, how to ensure that our evaluation findings would be used by the Congress. That there was a problem with use in the Executive Branch had become clear to me as a result of a symposium I had convened in 1976 at Mitre. There, representatives of nine Federal agencies had told evaluators not only that much evaluation was ignored, but also—in rich and painful detail—why it was ignored and what needed to be done about it (Chelimsky, 1977). But beyond the question of use, I did not give a lot of thought to theory. This was because I fully—and erroneously—expected that the panoply of principles which had guided my Executive Branch work would continue to serve in a Legislative role. Also, bombarded as I was with preoccupations like organizing and setting up administrative procedures for a new evaluation unit, hiring evaluators, and getting started on some evaluations, I believed, with Scriven (1991, p. 360) that theory is something of a luxury to an evaluator. Of course, I had not yet grasped how desperately I would need principles and theory to explain—to GAO managers, to my own staff, and to the Congress—my rationales for the organization, activities and research choices I would make at PEMD. My focus was on translating evaluation practice as I knew it into legislative reality, and thus showing, despite an army of naysayers, that it really was possible to perform evaluations, on a routine basis, that were generally accepted as credible, and that would be used by their congressional sponsors.

In other words, I expected to plug the old ways of doing things into the new legislative setting, while at the same time trying to learn a little more about how to facilitate use. So I set out to discover when, how, and why the findings of any kind of study were used by members of Congress, with the idea of shaping research procedures at PEMD to optimize use (Chelimsky, 1981). But I soon found that it was hard to study use in isolation because it was deeply embedded in three other issues that did not appear to have been much examined in our theoretical literature. These issues were: the general political environment, the history of the topic being studied, and the political, social and economic values embodied in that topic. Now, these issues may have seemed less significant in an Executive Branch evaluation climate (which most of our theoretical literature appears to reflect) because an administration’s political commitments are usually well known, and it can be counted on to come down, more or less clearly and at least for a short time, on one side of a policy question. But evaluation for the Congress can never avoid these issues. The political environment there is immediate, and it has a great deal to say about which topics will be studied, and which evaluations will be completed, released and used; history continues to matter in the thinking on both sides of the aisle, which are always represented in every congressional committee; and most of the past value clashes that cause legislation to be vague and ambiguous do not fade gently into the
night. As Congressman Richard Bolling once said, Legislative strategy doesn't consist of reporting out the best possible bill, but rather of reporting out a bill that will pass.

In PEMD, as these insights gained support from on-going work, we developed a kind of theoretical framework for our practice that eventually governed a great many of our decisions: for example, whether to perform an evaluation at all; whether to answer, or try to change a policy question; and how to choose methods that were not only feasible and appropriate to the questions posed, but that also took value-clashes and likely controversy into account. We deliberately left this framework iterative throughout the life of PEMD, and we kept changing it as new experience was gradually folded in.

Today that experience is complete, and it is the framework we used that I want to tell you about. Although time does not permit a full discussion, I have organized these remarks around what I think are the two elements that most need to be contrasted with current theory:

First: the general relationship between evaluation and politics; and
Second: the relationship between the evaluation process and program or policy history, embodied values and past research.

I. THE GENERAL RELATIONSHIP BETWEEN EVALUATION AND POLITICS

In my judgment, two problems exist in the way we have traditionally thought about the fit of evaluation into the real world, and in particular, into the world of politics which conditions the policies or programs we want our work to influence. The first problem is that we have not thought a whole lot about it. The second is that, when we have thought about it, we have examined evaluation and politics separately, rather than probe the relation between them. And we have looked at politics as peripheral or contextual to evaluation, not viscerally connected to it.

Yet many evaluations have been deeply affected by the way politics works in a democratic, pluralistic society, and especially by the continually changing nature of the political environment. Whether evaluation occurs in the Legislative or Executive Branch, practicing evaluators learn quickly that their study sponsor may not be there for the duration of the evaluation, and this can have a pretty chilling effect on use. A legislative sponsor may be defeated in the next election and replaced by someone with very different views on the evaluation topic. And the department official who asked for the evaluation may leave at any time for greener pastures, to be replaced by someone who may have little or no interest in evaluation generally, or who represents a different shade of opinion within administration politics.

We have recently seen examples of this with welfare studies undertaken for Health and Human Services officials who lost their battles within the administration and resigned. In PEMD, our experience with the problem dates from the very beginning of our existence when we discovered how rare were members of Congress willing to ask serious policy questions about Defense Department programs. This remained true no matter which party—Republican or Democrat—was in control. What it meant for us was that the loss of a particularly strong sponsor of a defense evaluation, like Senator David Pryor or Congressman Dante Fascell could spell not only problems for use, but worse: the elimination or devitalization of an important, expensive study.
Changes that go beyond individuals to the general political stance of government also affect evaluation in major ways. The entire climate for evaluation can be altered by presidential elections (like those of Lyndon Johnson or Ronald Reagan, for example), or legislative elections (like that of 1994 which changed the balance of power between the parties in the Congress). In the Johnson era, evaluators studied the results of expanded government programs seeking to help the disadvantaged. Under Reagan, we studied the outcomes of cutting or devolving those same programs. Today, with Clinton and Gingrich, although we did escape having to evaluate orphan asylums, we will be studying the dismantling of assistance to families with dependent children (AFDC), and also the initial reductive changes in middle class programs like Medicare. These are 180-degree turns in policy making, but evaluation must be able to accommodate them.

In PEMD, we went from evaluating categorical grant programs at the federal level to looking at state-level block grants, and it was no mean trick to do that because the Reagan Administration was systematically decimating the data bases of programs that had been devolved, eliminated or consolidated (Chelimsky, 1985). In an evaluation we did of the Comprehensive Education and Training Administration Program (CETA), our work began in pre-Reagan 1980 as an effort to look at the effectiveness of the program in training disadvantaged adults, but by the time we were done, the Reagan revolution was in full swing and CETA was tottering amid allegations of fraud and abuse, distrust of activist government, and a general rejection of the idea that training—and especially government training—could actually help the disadvantaged (USGAO/IPE, June 1982).

Well, our findings were mildly favorable. We presented evidence that women, and people with extremely poor earnings histories, had profited from the program, and that early participants were better off, on average, after CETA than before. But because the findings were somewhat favorable, they ran up against the prevailing “destroy-CETA” sentiment, and found few takers in the administration, little interest on Capitol Hill, and no discussion whatever in the press. This happened despite methodological strength in the study which should normally have assured it an audience. The truth is that had the study been published just a few months earlier, it would at least have served during the debate on CETA, and might have mitigated some of the rhetoric, even if it was too late to influence the program’s demise, and the substitution of a public/private training program promoted by Senator Dan Quayle.

We had a similar experience in evaluating home health care, which was still a new concept in 1980 (USGAO/IPE, December 1982). Here we had found a serious potential for cost escalation in the proposed new program, an issue that Democrats in the Congress, and even Senator Orrin Hatch, though Republican, just did not want to hear about. Had our evaluation been published a few months later, in the new Reagan climate, Republicans and many Democrats as well would have looked much more carefully at the findings. The point here is that political volatility has effects on evaluation practice that can prevent a study from starting, delay its execution, and render its findings moot, depending on the place in the evaluation process at which the political changes occur.

Another way in which politics drives evaluation arises not from dynamism in the political environment, but from its contentiousness, and from the fog of war that surrounds most important controversial topics. That is, in a world of highly sophisticated and continuous jockeying for political advantage, advocacy abounds. Not only do policy makers have their own political agendas, they are also besieged by pressure groups, vested interests and lobbyists, all with their war stories about “success” or “failure,” and all trying, with money, power, and data, to move policies and programs in specific directions.
There is not much new about this, except perhaps in degree and in analytical sophistication. Jefferson was already talking in the year 1816 about the propensities of lawmakers, functionaries and others to “command at will the liberty and property of their constituents. There is no safe deposit for these,” he said, “but with the people themselves, nor can they be safe with them without information” (Jefferson, 1939, p. 89). The need in a political environment is not for still another voice to be raised in advocacy, but rather for information to be offered for public use that is sound, honest, and without bias toward any cause. Policy makers in the Congress expect evaluators to play precisely such a role and provide precisely this kind of information.

In PEMD, right from the start, we received requests from members of Congress asking us to “make sense,” as they put it, of conflicting research findings, or to examine the evidence behind statements that claimed to be based on research. The traffic in this kind of question grew so heavy that we developed a special category of studies which were essentially methodological reviews, but whose larger purpose responded to the public-interest need for distinguishing good information from bad. One such request asked us to determine whether Secretary William Bennett’s statements about the “failure” of bilingual education were true, and whether they were in fact based on research evidence as he alleged (USGAO, March 1987). Another request was for us to critique the methodology used by Harvard’s Physician Task Force on Hunger in arriving at its list of “the 150 counties with the worst hunger in America” (USGAO/PEMD, March 1986). A third request wanted us to explain how it could be that four different studies, estimating the annual U.S. production of hazardous waste, used different samples (one about a quarter the size of another) and different approaches, yet all arrived at the same mystical total of 260 million metric tons (USGAO/PEMD, February 1987). We were asked to examine the validity of different claims for the number of homeless that ranged from 250,000 to 3 million (USGAO/PEMD, August 1988), and different estimates for the number of illegal aliens ranging from 1 million to 12 million (USGAO/IPE, September 1982). The requests went on and on, but the point here is that political advocacy dictated an overwhelming majority of the estimates, methods and conclusions we were asked to review.

What implications are there for theory if we move toward a view of politics as central to evaluation practice?

Let me discuss four of them here.

The first implication is a need to broaden our thinking about politics and policy making. It is true that we have come a long way from imagining as we once did that evaluation is an inflexible process that must be implanted in some uninvestigated political environment where unwilling bureaucrats are maneuvered into letting the evaluation proceed untrammeled. Still, even today, we see politics as merely the “context” of an evaluation, as something that “intrudes” upon good practice, rather than the engine driving it (Shadish, Cook & Leviton, 1991, p. 86). While this may be better than looking at a political landscape simply as a spot on which to drop an evaluative egg, it is only a little better because it does not explain the nature of the political beast, or account for its many effects on the evaluation enterprise, or recognize the policy making diversity which necessarily flows from flux and change in politics. For example, we’re told that policies happen as the result of “gradual accretion” (Shadish, Cook & Leviton, 1991, p. 192); that the nature of program change is “slow and incremental” (Shadish, Cook & Leviton, 1991, p. 188); that no single authority can radically change a program
(Shadish, Cook & Leviton, 1991, p. 39); and that evaluators should not expect “go/no go decisions” to be based on their findings (Cronbach et al., 1980, pp. 4 and 62).

Although these precepts may describe one type of policy making at a particular time and place, none of them matches our experience. The Congress makes go/no go decisions often, in any political environment, on the basis of evaluation findings or anything else they may happen to have at hand. Single authorities can and do change programs radically when the political climate makes it safe to do so. As for gradual accretion and incrementalism in policy making, they fit better with more stable political climates than they do with unstable ones. In short, political dynamism means that no single model of decision making fits all environments, and that we have to accommodate both incrementalism and rapid radical change in our thinking.

A second implication for evaluation practice arising from the nature of politics is a need for credibility. If evaluation is to fulfill its purpose of furthering the advance of knowledge in the public interest, if study findings are to survive political change, obtain a thoughtful hearing in the public forum, and successfully counter at least some of the myths and mystifications disseminated by advocates, then evaluation must itself be perceived as credible, without advocacy. Put another way, if evaluation is not impartial, then it has no special place or prestige in public debates. Yet we have recently seen attempts to rationalize advocacy by evaluators, and this idea has some roots in theory.

Cronbach, for example, believes an evaluator should not study programs “with whose basic aims he is not in sympathy,” nor even “undertake to serve an agency,” unless he or she agrees with its general mission (Cronbach et al., 1980, pp. 208-211). House advises evaluators to “advance the interests of the least privileged members of society” (Cronbach et al., 1980, p. 209). Weiss points out that it is dangerous to foster the idea that social programs do not work, since that would give “aid and comfort to the barbarians” (Shadish, Cook & Leviton, 1991, p. 420); and Greene (Chelimsky & Shadish, 1997, p. 471) states that “evaluation should advocate for the interests of program participants.”

Our experience in PEMD was that advocacy of any kind destroys the evaluators’ credibility and has no place in evaluation. Our continuing battles on the Hill and with executive agencies about whether we had an ax to grind in making certain statements show the political importance for the evaluator, of being able to demonstrate impartiality if study credibility is to be preserved, and for the advocate, to demonstrate bias if the study is to be successfully attacked. Had we tried to work only on subjects that appealed to us, or shied away from telling the truth about social programs, or advanced the interests of any group at all, we would have been making a priori rather than evaluative judgments about what is good or bad for society, introduced demonstrable biases into our work, and given our political opponents—that is, those who were unhappy with our findings—wonderful new ammunition with which to attack us.

It is true, of course, that every evaluator has values, beliefs and feelings, but that is hardly a reason to recommend partisanship. Instead, one way to correct for an evaluator’s bias is to counter it with the opposite bias in selecting the evaluation team. Another way is to bring in a group of advisors, carefully chosen for this purpose. We did that for many of our defense studies, and, in particular, for the analysis of Secretary Bennett’s statements on bilingual education that I mentioned earlier. But in my judgment, the most important thing we did to correct for bias in our studies was to think about it all the time and especially at particular points in the evaluation process: for example, during the literature review, by trying to understand value clashes in the program and critiquing past research for advocacy as well as methodology; dur-
ing construction of the evaluation design, by making use of methods specifically combined to increase objectivity; and during the writing of the final report, by emphasizing careful, specific and precise language. All of these strategies are, of course, procedural, practical ways of trying to maintain evenhandedness: they worked quite well for us, and the game was clearly worth the candle. In Robert Stake’s words, “the evaluator should not impose one ethical view on a program in a political system characterized by value pluralism” (Shadish, Cook & Leviton, 1991, p. 49).

A third implication that the political engine has for evaluation is the need for timeliness. When the political climate changes, presenting findings too late can be almost the same as not presenting them at all. Yet some theorists believe that “timeliness is an overrated concern” (Cronbach et al., 1980, p. 63), and Weiss has written that:

> Being practical and timely and keeping the study within feasible boundaries may be unimportant or even counter-productive. If the research is not completed in time for this year’s budget cycle, it is probably no great loss. The same issues, if they are important, will come up again and again (Shadish, Cook & Leviton, 1991, p. 197).

But in PEMD, working for the Congress, we learned to be very respectful of timeliness with regard to legislative milestones like program authorization, reauthorization, or appropriations. Even in the Executive Branch, timeliness is important if the evaluation was mandated by the Congress, if the administrative cycle is closely tied to congressional hearings or other legislative events, or if the agency has come under attack. In periods of political uncertainty, timeliness can mean the difference between getting and not getting a hearing, between use of the findings and non-use, between increasing the evaluators’ credibility and weakening it. Untimely evaluators hurt their long-term reputations not only because they have failed to live up to their engagements, but also because of an iron-clad conviction in policy circles that delay in a report signals either incompetence, or a deliberate effort to cover something up. This is one reason why advocates opposed to an evaluation will usually begin by trying to slow it down.

Moreover, although it is true that important issues do come up over and over again, the deduction that timeliness therefore does not matter applies only to Executive Branch evaluation in politically stable environments. As things change, and especially when they move toward reform ideologies, untimely evaluators in both branches risk discovering that the debate is no longer the same, the policy questions they have addressed are irrelevant, and the audience for their findings has gone up in smoke.

The fourth and last implication of the general relationship between politics and evaluation is the need for flexibility, for a certain tentativeness during much of the evaluation performance. This strengthens the evaluators’ hand, because they know that things are not set in concrete, and that they have the power to confront unexpected controversy by changing the evaluation design to make it more robust, by adding consultant help to increase expertise and impartiality, or even by joining a whole new component to an evaluation, as we once did at the end of a study of enterprise zones, to convince a disbelieving sponsor (Congressman Jack Kemp) that our negative findings were valid (USGAO/PEMD, December 1988).

Such tentativeness also applies to the need for care in making absolute judgments, which are especially vulnerable to different political optics in changing times. Medicare, for example, was for decades qualified as a great success because access by the elderly to health care was the most important program goal and the program had reached 96 percent of the elderly. But after 1994, with entitlements on everyone’s hit list, the most important goal was no longer...
access, but cost containment, and Medicare became an instant failure because of its steady cost increases and numerous inefficiencies.

As a result, we learned in PEMD to emphasize observations that are as specific, precise and validated as possible, not only because they are apt to be more accurate and useful, but also because they are more difficult to challenge as political values change over time. We reduced our quota of blanket qualifiers, such as "successful" or "unsuccessful," "good," or "bad," and, when we did use them, we tried to be very clear about their specific meaning.

To sum up, then, I have been arguing up to this point for a sea-change in the way evaluators look at politics. I am suggesting that evaluators need to better understand the climate of political dynamism and contentiousness that surrounds them, and to understand as well that politics is central in establishing the evaluator's non-partisan role, in shaping the evaluations themselves, and in determining the use that will be made of them. Because of this centrality, I am also suggesting the need to develop some strategies for coping with political change—strategies like timeliness, flexibility and attention to long-term credibility, among others—that often run counter to current theory about evaluation and politics.

Now let me move on to the second element in my discussion.

II. THE RELATIONSHIP BETWEEN THE EVALUATION PROCESS AND PROGRAM OR POLICY HISTORY, UNDERLYING VALUES, AND PAST RESEARCH

A policy or program in America is usually a witch's brew of ideas, often in conflict, that were current at the time the legislation was passed, and that have been modified over a subsequent span of years. The history of these ideas and their current status are important because they clarify the disagreements among competing values that often constitute the crux of an evaluation. The problem for public policy is that no clash of values—no matter how well camouflaged by legislative language—ever seems to die, but goes on forever, concealed within the recesses of historical debate. As Marris and Rein put it:

Since every society is informed by a great variety of ideals and interests competing for expression, it compromises them all and can fully satisfy none... (T)his fundamental incompatibility reappears at every level of discussion. Any policy implies the reasons by which it could be refuted. In appealing to the values which justify it, it must disparage others which are also valid, and whatever balance it strikes enjoys only a grudging and provisional acquiescence. (1973, p. 236)

In the AFDC and Medicare programs, for example, it is clear today that acquiescence was indeed grudging and provisional, even though it took 60 years in the one case and 30 in the other for that to be obvious. At PEMD, we became familiar with a kind of infrastructure of values in almost every program we evaluated. But in one program, that of the Administration on Aging, we found a clash of values that was overt, without any kind of acquiescence whatever.

The Administration on Aging (AOA) is an agency with the grandiloquent and expanding objectives appropriate to the Johnson-era consensus on Government's social role. But it had only small and rapidly declining funds to serve a large and rapidly increasing elderly population. The irrationality was evident: How can expanding services be provided to a growing number of people, with resources that are not only tiny but shrinking? Yet what was irrational
from a common-sense viewpoint made political sense as a reflection of the 1980's environment. On the one hand, the declining funds mirrored the desire of Reagan-era conservatives to move government out of the service-delivery business. On the other, lofty objectives and multiplication of services, at least on paper, mirrored the desire of liberals to claim that they were doing something for the elderly. Even though the funding cuts were real while the inflated objectives were rhetorical, the political balance was a meaningful one and could not be disturbed over the short term. This was important for us to understand because it reduced the likely usefulness of proposing either decreased grandiloquence or increased funding—since neither was feasible in this particular environment—and because it was crucial in negotiating the right policy questions to answer (USGAO/PEMD, February 1992).

We encountered a different type of values-clash in our work on the Defense Department's Chemical Weapons Program. Here we found two entirely separate sets of literature on chemical warfare: one was classified, favorable to the program, "hawkish," and opinion- rather than data-driven. The other was in the public domain, opposed to the program, "dove-ish," but equally opinionated and data-deprived. The two sets of literature clearly indicated deep political conflicts, but there was no need for DOD to try to embody these conflicts within the program as the Congress had done with the AOA program. Instead, because of classification and DOD's resulting ability to present only favorable information to the Congress and the public, the agency could afford to simply ignore its critics. So in 1981, when we started our first evaluation on the topic, there was little realization in policy making circles that conflicting views even existed on chemical warfare, much less what the specific weapons issues were.

This situation had ramifications for our work. In particular, it led to two decisions: (1) that we should re-negotiate our evaluation questions with the requesting committee so as to answer only a single question about knowledge acquisition in the program, and (2) that our first product should not be an evaluation of weapons effectiveness as we had originally intended, but rather a critical synthesis of all the literature (USGAO/IPE, April 1983). This report, when it finally emerged after a long battle with DOD, had an electrifying effect on Members of Congress who were confronting certain facts for the first time. In addition, it brought our division congressional requests for eight more evaluations, over time; it opened up our first serious interchanges with thoughtful journalists; and it eventually resulted in the strongest possible instrumental use of our findings by the U.S. Government in its negotiations with the then Soviet Union: first for a bilateral chemical weapons agreement, and later for the international treaty that evolved from it (Fascell, 1990).

This experience taught us to pay a lot more attention to past research and its patterns of advocacy. Like most evaluators, we had all typically performed literature reviews, but now we extended them to include program history and values, past and current controversies, and political positions taken by earlier research. That is, in reviewing previous evaluations, we not only looked at their findings, their design strategies and measures, the data they used or developed, and their experience of success and failure, but also at the patterns of partisanship they betrayed. In short, we investigated the political underpinnings of both the program and its evaluations, especially the controversies of the past that remained relevant, the major stakeholder positions today, and the likelihood of new controversy awaiting our evaluation. This expanded literature review was the first step we took toward knowledge construction in every evaluation, but the larger need it fulfilled for us was to integrate conflicting values and past research into our core thinking about the evaluation we were setting out to perform.

This work had cascading effects on the rest of our evaluation process, especially the questions we would agree to answer, the methodological strength of our evaluation designs, the
way we would write the final report, the knowledge contribution we hoped to make, and the eventual policy use that might be achieved.

There are certain implications for theory that result from assimilating program history, value conflicts and controversy into the evaluation process. Let me discuss five such implications here.

1. The Need to Change the Way We View Stakeholders

The first implication for theory is that we need to change the way we view stakeholders, people who have a vested interest in, or are advocates for a policy or program. At PEMD, we had to recognize the power of stakeholders to influence our work right from the beginning. Although the size of the problem and the place of intervention in the evaluation process differed for different evaluations, few of our studies in even minimally controversial areas were entirely exempt from stakeholder involvement and pressure. This is perfectly normal in political environments, and the games played by stakeholders typically feature high risks, low blows, and territorial ferocity. Yet this is not how some theorists perceive stakeholders.

Reviewing Guba and Lincoln’s *Fourth Generation Evaluation*, for example, the reader comes away with the idea that political environments are bland, and that clashing stakeholders might conceivably agree to “conditions for a productive hermeneutic dialogue” (1989, pp. 191–204). Cronbach and his colleagues also offer the vision of a benign “policy-shaping community,” designed to minimize conflict in a system of mutual accommodation, and they believe that the main reason why evaluation results are challenged and discredited is not due to the actions of stakeholders, but “because no adequate critical process precedes their release” (Cronbach et al., 1980. pp. 100, 2, and 131).

Weiss (Shadish, Cook & Leviton, 1991, p. 186), on the other hand, does present a much more realistic picture of stakeholder forces in operation, and if you go back to Suchman (1967), you can get a good account, although somewhat understated by today’s standards, of the Byzantine strategies and elaborately calculated tactics that serve stakeholder interests (Suchman, 1967, pp. 143-144). Nonetheless, it seems to me that nowhere in our literature is the scalding nature of some advocacy battles depicted as we have experienced them, and adequately conveyed to practicing evaluators.

Perhaps some of this downplaying (or ignorance) of how stakeholders operate derives from current notions that achieving objectivity is impossible, that bias is a universal characteristic, and that one bias is not worse than another. Even the evaluator is guilty of advocacy in favor of his or her evaluation, according to Stake (Chelimsky & Shadish, 1997, p. 471). So some may imagine that since everyone is an advocate and all advocacy is equal, then—in an ideal world where everyone is willing to relinquish his or her particular bias—it should be possible to sit down together amicably and negotiate.

Our experience in PEMD was that little or no amicable negotiation ever took place with stakeholders concerned that an evaluation could hurt their interests. Negotiations? Yes. Amicability? Never. Further, the idea that all bias is equal seems singularly unpersuasive. Even if no one is bias-free, including the evaluator, there is still a huge difference between a stakeholder with an agenda and an evaluator trying to be as impartial as humanly possible. Scriven (Shadish, Cook & Leviton, 1991, p. 89) has it right: “Crude measurements are not as good as refined measurements, but they beat the hell out of the judgments of those with vested interests.”
The bottom line is that evaluators should not expect any gifts from stakeholders, but there are many nuances in our relationships with them, depending on the kind of evaluation that is being done and the kind of stakeholder involved. Certainly, in developmental evaluation, in the kind of work that builds evaluation capability in agencies, in the relationships described by Patton (1997) and Fetterman (Chelimsky & Shadish, 1997, pp. 381–395) or even—in a different way—by Wholey (Chelimsky & Shadish, 1997, pp. 124–133), evaluators and stakeholders often work together toward a common goal. But even though the relationship is collegial and sympathetic, evaluators and stakeholders are still very different: the evaluator has no vested interest in the program, or at least much less interest than the stakeholder has, and the evaluator’s purpose is the public good, which may or may not be congruent with the stakeholder’s good.

But even in accountability evaluations of the sort I have been describing, where evaluators and stakeholders may be at war with each other, the good news is that if evaluators can predict stakeholder attacks soon enough and well enough to plan for them adequately, and if they have the courage to stick to their guns, there is an excellent chance they will win their argument, as we learned in PEMD from our ten years of work on chemical warfare. We certainly had major skirmishes with stakeholders (such as medical device manufacturers, agency managers, highway and insurance lobbies, military technology corporations and their media, pharmaceutical or drug companies, scientists, the beer lobby, liberal or conservative think-tanks), and those skirmishes went on and on, but in the majority of cases, the work we did won out, sometimes even immediately, but sometimes not.

In a previous AEA address (Chelimsky, 1995), I mentioned a belated victory we had won with respect to cataract surgery against an important group of stakeholders in the Medicare Program—the ophthalmologists (USGAO/PEMD, April 1993). Today, another of our evaluations, which I thought had been definitively shot down by the National Rifle Association (NRA), has achieved delayed but very real instrumental use of its findings and recommendations. Six years ago, we estimated that “about 1 in every 3 deaths from accidental firearm discharges could be prevented by a firearms safety device,” and, in particular, we strongly recommended child-proof safety locks. The work featured an interesting combination of data from police and health care sources, but because the NRA opposed its recommendations (and indeed, had opposed the entire study right from the start), it went largely unnoticed, abandoned even by its sponsor, Senator Howard Metzenbaum. But a few weeks ago, as you may have seen in the press, 9 major gun manufacturers, under threats and pressure from the administration, moved at last—“voluntarily”—to provide childproof locks with all their handguns. In this case, the victory was not due to our persistence or superior strategy (we were much too dejected for that), but rather to a change in the political climate featuring some small weakening of NRA’s power and influence in the Congress. This, along with our recommendations and the solid new estimates we provided, gave the administration, six years later, a stronger basis for effectively pressuring gun manufacturers than had existed before we did our study.

2. The Need to Re-think How We Use Goals and Objectives

The second implication I want to mention of trying to integrate program history, values, and past research into our evaluation process, is the need to re-think how we use goals and objectives. In accountability evaluations, it was normal some years ago to use them for assessing success or failure, but our theory today tends to reject them as serious benchmarks for measuring program achievements. Cronbach (1980), for example, tells us that it is “unwise for
evaluation to focus on whether a project has ‘attained its goals’” (p. 5); Scriven (1973, 1976, 1980) has long desired to see “goal-free evaluation”; and Stake emphasizes “allowing evaluation to emerge from observing the program.” (Shadish, Cook & Leviton, 1991, p. 273) In Scriven’s case, the concern is to distance the evaluator from what he calls “massive” bias built into goals and objectives by stakeholder managers (Cronbach et al., 1980, p. 132). But in other cases, ignoring goals and objectives is part of a loss of interest in assessing accountability, and a loss of interest, as well, in the methods required for that purpose. Cronbach is quite direct about this. He says he is “uneasy” about the relation between accountability and evaluation, and feels that “evaluation is better used to understand events and processes” (Cronbach et al., 1980, pp. 133–134), which is, of course, a purpose that calls for different methods than those typically used for studying accountability (e.g., process rather than outcome evaluation). Stake is also wary of accountability, and the case study method he favors certainly accommodates accountability less well than other methods. In short, both of these stances with respect to goals, objectives, and accountability are reinforced by the methodological choices these evaluators have made.

In PEMD, we had no choice about whether or not to assess accountability, or use the methods appropriate for it, because policy makers in the Congress simply insisted on it as an essential part of the evaluator’s role. And in the process of fulfilling these expectations we discovered that goals and objectives mattered greatly, although not necessarily as benchmarks, unless they were the carefully developed, operationally defined, measurement-friendly type to be found in Wholey’s universe of the Government Performance and Results Act (GPRA) (Chelimsky & Shadish, 1997, pp. 124–133). Instead, we learned to use even biased, vague or conflicting goals as indicators of a program’s past history and clashes of values. A careful deconstruction of goals and objectives, combined with a more general study of what the program debates and past research findings have been, can help evaluators ask the right questions about a program and quickly tap into the key elements of a problem. The multiple objectives of the Women’s, Infants’ and Children’s Program, or of CETA, for example, were instructive in showing us the overload of good, but sometimes antithetical, intentions in the programs, the inaccessibility of some of those objectives to analysis, and hence the right places for us to focus our research.

As for accountability evaluation, because it involves assessing effectiveness or merit, it remains a fundamental part of the research effort to improve policy making, it continues to be expected of evaluators world-wide, and we found in PEMD not only that it can and should be done, but that there are many strong methods for doing it that do not necessarily involve the use of goals and objectives as benchmarks. When these are too unspecified, too “pie-in-the-sky,” or too biased, previous research can often help. In a number of studies on drug use, for example, we relied on evidence of past effects and used those as comparisons, rather than the overoptimistic program goals. Still, even when they cannot serve as benchmarks, goals and objectives remain useful indicators of past program accommodations and value clashes, they alert evaluators to likely controversy, and force them, early in the evaluation process, to think about how to strengthen their study against attack.

3. The Need for Inclusiveness

A third implication of assimilating history, values and past research into the evaluation process is the need for a certain inclusiveness. This is to ensure that the new evaluation will be complete and balanced enough both to achieve credibility in the current political environ-
will also pass muster in the next. One aspect of such inclusiveness involves awareness not only of highly-publicized stakeholder positions with regard to a policy or program, but also of stakeholder positions that are less well-known because they are not represented by an organized lobby. This is often the case for program beneficiaries such as children, patients, poor people, immigrants, or the severely handicapped. Their views are important for an understanding of the program, but they are often overlooked by evaluators.

Paying attention to what the beneficiaries of a program think about it is a hallmark of a credible study, and has nothing to do with advocating for those beneficiaries. Instead, it does three things for an evaluation. First, it aids impartiality and even-handedness by counter-balancing management biases with beneficiary biases. Both sets of views need to be understood and validated, not championed. Second, because so much research has failed to deal with beneficiary views in the past, an evaluation that does so may well produce critical new information. Finally, since beneficiaries know more, from personal experience, about the qualities and inadequacies of a program than anyone else, learning what they know greatly improves the sensitivity of an evaluation. For example, one of the first things we did in our study of the Americans With Disabilities Act was to survey disabled people about barriers important to them, which they had encountered before the act was passed. We used their responses in constructing both the evaluation design and the survey instruments, recognizing—based on what we had learned from them—the need to ask probing questions of business owners and operators not just about observable barriers, but also about invisible ones, such as whether a blind person with a guide dog might be refused entry to a cafe or restaurant (USGAO/PEMD, May 1993).

A second aspect of inclusiveness is cost. It is almost a truism in politics that when conservatives attack a policy or program, even though the real reform effort may be directed at something else—its values, for example—the initial battles will usually be about its affordability, its rising consumption of taxpayer funds—in a word, its costs. So, a critical implication of the effort to integrate values within the evaluation process is the need to study costs. Yet if you look at our evaluation theory, there is a staggering silence there about why costs are important, how and when they should figure in our work, or—with exceptions for the work of Rossi and Freeman, (1993) and Wholey, et al. (1971)—what methods should be used, with what specific applications, what pitfalls and booby traps. In a field known for its intense concentration on methods, the area of costs constitutes a singular exception.

In PEMD, the study of costs became a standard element in our work. We examined them, and their changes over time, as a matter of course in most evaluations, whether or not a cost question had been asked. We also learned to cost out any recommendations we made, whether for expenditures we considered necessary or for cost savings we hoped to achieve. In controversial topics like the strategic nuclear triad and its weapons systems, cost comparisons constituted an entire component of our evaluation, and turned out to furnish one of its most potent and persuasive arguments against buying any more B-2 bombers (USGAO/PEMD, September 1992). In short, including cost issues in the evaluation process improves perceptions of study credibility, helps evaluations to resist political attack, and increases the likelihood that their findings will be used over time.

A third aspect of inclusiveness involves the need to broaden our repertoire, despite warnings from theorists about doing this or that kind of study. Campbell for example, tells us we should avoid looking at entrenched, currently operational programs (Shadish, Cook & Leviton, 1991, p. 152); Suchman (1967) says the same thing; and Cronbach agrees that such programs are "immune to serious evaluation" (Shadish, Cook & Leviton, 1991, p. 352). We are
cautioned by Cronbach not to worry overmuch about internal validity. As Shadish, Cook and Leviton (1991) note, Cronbach also questions the wisdom of trying “to assuage the last shred of plausible doubt about causation” on the grounds that policy makers pay little attention to it (p. 344).

Our experience in PEMD was that if evaluation is to be done credibly in a political environment, then we need to extend our horizons, not shrink them. Besides, methods development in recent years allows evaluators to answer many kinds of questions, including accountability questions, and to study all kinds of policy initiatives, including big, entrenched, operating programs. We were continually involved, in PEMD, with supposedly “immune” programs like AFDC, Medicare, the Construction Grants Program, or the Strategic Nuclear Triad, which spanned a multitude of giant operating programs across two of the Armed Services.

Further, although it is quite true that policy makers pay relatively little attention to the degree of proof brought by evaluators to establish causation, it is also true that they pay no attention to any methodological issue. However, policy makers have competent advisors who let them know about the methodological credibility of studies whose findings are of interest to them. At PEMD, this came home to us in the form of sometimes arcane methodological inquiries and criticisms from congressional staff and consultants, if not from the policy makers themselves. There is no lack of methodological expertise in the research circles surrounding policy makers, and queries are not rare about the extent to which findings of causation are properly supported. Indeed, evaluators, like Confucius, are fortunate people: when we make a mistake, someone is sure to notice it.

So even in a stable political environment, evaluators get queries about their methods. But when there is a change in the climate, or the topic is controversial, and a full-bore political attack is under way against an evaluation’s methodology, it is quite astounding how much attention policy makers pay and how quickly they pick up on small flaws in causation, or statements about the causes of an observed finding which flow from an inappropriate methodology. As Mosteller and his colleagues have noted, “statisticians are perennially impressed by the sudden methodological expertise of critics immediately following an unpopular finding” (Hoaglin, Light, McPeek, Mosteller & Stoto, 1982, p. 57).

Moreover, even in stable political climates, the GAO was criticized by the Congress for years and years because auditors made general causal statements based on a single case study. This criticism, from the GAO’s House Authorizing Committee, was not motivated by the usual political concern about findings rather than methodology, but proceeded from an analysis of many GAO studies over time. Thus, whether in stable or unstable times, it is unsafe to assume that the quality of causal methodology is unimportant, or that policy makers will not pay attention to it.

Inclusiveness applies not only to methods, questions and programmatic challenges, but also to subject matter. In PEMD, because our sponsors required it and because evaluation methods allowed it, we went beyond the usual evaluation subject matter to include areas like defense, transportation, energy, the environment and others. Overall, we found that most evaluation methods were robust, and easily transferable to a number of areas where they had not been seen before. In addition, for the same reasons, and using the same methods, we learned to answer prospective as well as retrospective questions. In short, assuring that evaluations are credible over time in a political environment means including all the major program values within the evaluation process, and paying particular attention to the views of program benefi-
ciaries, the size and growth-trends of program costs, and opportunities to expand the form and content of evaluation itself.

4. The Need to Re-think Criteria for Deciding When to Do An Evaluation

The fourth implication I wanted to discuss involves the need to rethink our criteria for deciding to conduct an evaluation. In PEMD, especially in the beginning, we had a tendency to agree enthusiastically to the Congress’ requests for evaluations, but we soon found out that it was necessary to think through carefully what should cause us not to do an evaluation, if we wanted to survive as a credible evaluation shop. We began by using Cronbach’s resource allocation criteria: that is, “prior uncertainty about a question; costs of information; anticipated information yield; and leverage of the information on subsequent thinking and action” (Cronbach et al, 1980, p.8). But although these criteria are excellent as far as they go, they are mostly aimed at knowledge acquisition, they presume a rational use of information by policy makers, and they do not take much account of the political forces affecting the success of an evaluation.

Based on our early experience, I ended up in 1982 with a set of 12 questions for making preliminary resource allocation decisions (See Appendix I). I posed these questions to staff about a month after we had received a new congressional request, so as to respond quickly to the requester. If the answers to the questions were troubling, we would try, together with the would be sponsor, to revamp the study to our mutual satisfaction. But if that failed, it happened that we might have to refuse an evaluation request for a variety of reasons, some infused by political considerations inherent in the work. In general, I did not turn down evaluation requests because the subject was controversial, or because the program’s objectives were unmeasurable, or because they required special expertise from our staff, or because policy use might be impeded by program or other advocates. But I did refuse them when: (1) it appeared they could not be done convincingly within reasonable funding constraints; (2) there was little or no knowledge base to build on and the area was new to us; (3) original data could not be collected and extant data were not available; (4) they could not be finished in time to be useful; (5) they would add little new knowledge, either about the program or about evaluation practice; (6) there was no policy fix available for implementing potential recommendations; (7) the public interest did not supersede all other considerations (as in the case, say, of major problems in a program that were unknown to the Congress, the press and the public); and (8) the policy questions posed to us were too numerous, or too broad, or too trivial, or too biased, and could not be changed.

We learned to present our reasons for saying “no” to the Congress very carefully, and sometimes our conclusions were accepted, sometimes not. In one case, I was called “recalcitrant” in a letter to GAO from Senator Ted Kennedy because I had refused a request to estimate the future impacts of a new provision in the immigration legislation without any historical data on which to base the estimates. In another case, the 10-page letter I had written to a House requester, explaining the reasons for our refusal, was faxed all over the country by that requester, without my knowledge, and accompanied by the query, “Is she right?” The first I heard of this was some months later when I received a letter from the requester, enclosed in a huge packet of mail, saying that, to his surprise, most of his correspondents seemed to agree with me, and so, where should we go from here?

It will surprise no one to learn that, in both cases, we were eventually obliged to do the studies (USGAO/PEMD, November 1989; and USGAO/PEMD, August 1988). But the refus-
als turned out to have been extremely worthwhile because they bought us what we had not been able to achieve without them—a serious reconsideration of the requests, major changes in the questions to be answered, appropriate timelines, and agreement for intensive bipartisan committee participation, in the one case, and extensive Executive Branch consultation, in the other. In short, incorporating program values and past research issues into our evaluation process brought us different, and much expanded resource allocation criteria than we would have used if we had thought of our work uniquely as a methodological operation for producing new knowledge.

5. The Need to Re-think Our Views on the Use of Evaluation

The final implication I want to discuss is the need to rethink the idea of use. As I said earlier, changeable political environments mean that both policies and styles of policy makers vary, that evaluators must do their work in both stable and unstable climates, and that no political environment is ever permanent. This puts new emphasis on the kinds of things I have just discussed, such as sensitivity to the harbingers of change, awareness of past and present controversies about a program, knowledge of program stakeholders and their positions, along with neutrality and a set of strategies for coping. We discovered that all of these new emphases, combined with special attention to communications and dissemination of findings, considerably enhanced the use of our work. Indeed, direct, instrumental use of our findings happened with great regularity, and I have already recounted the spectacular experience of use we had in our chemical warfare work, despite 10 years of opposition to it by the Defense Department. A few other examples should also help to illustrate the range and diversity of this type of use.

One evaluation we did of the AFDC program, for instance, brought about a change in the law which allowed working mothers leaving the program to receive Medicaid for their children over longer periods of time than was possible before our study (USGAO/PEMD, July 1985). Our work on down-sizing methods used by the Government—that is, attrition, furlough, and reduction-in-force, or RIF—showed that RIFs were often more costly than attrition or furlough, because of hidden pension costs. This finding was cited by many Executive Branch agencies as the reason for reducing their use of RIFs (USGAO/PEMD, February 1985). Another evaluation led to doubled congressional funding for the Runaway and Homeless Youth Program, an effective and well-managed federal effort whose appropriations the Reagan Administration had proposed cutting in half (USGAO/IPE, September 1983). A study we did of employee stock ownership plans was responsible for a saving of 1.9 billion dollars in tax expenditures by the Treasury (USGAO/PEMD, October, 1987). Another evaluation showed that an increase in the drinking age (from 18 to 21) significantly reduces traffic fatalities (USGAO/PEMD, March 1987). This finding countered beer-lobby propaganda that changing the law would have no effect at all on fatalities. So one use of this study was by 16 state legislatures which voted to increase the drinking age in their states. A second use was by the Supreme Court which cited our evidence to set aside objections about the legislation’s effectiveness. A third use was by the National Highway Traffic Safety Administration which borrowed our methodology for studies of its own, one of which estimated that our work had been responsible for saving about 1000 lives in the course of one year.

These are, of course, just examples among others of instrumental use of our findings, and they show clearly that this use was not restricted to the sponsoring entity (that is, the congressional committee or the senator) but extended to the Congress as a whole, to Executive Branch
Agency policy makers, the evaluated programs’ managers, to state legislators, to the Supreme Court, and to Executive Branch researchers. Further, regular secondary users of our work were the press, which disseminated our new findings and used our past work for support in investigative reporting; television, which aired our congressional testimony and reported on our battles with stakeholders; and the research community, which cited our findings, re-used our original data, and included our studies in their syntheses.

Now, this is not at all what most evaluation theorists lead us to expect. On the contrary, instrumental use of findings is supposed to be a rare event. We read, for example, that “few clear incidents of such use are documented,” and that if it should occur, “the slow, incremental nature of policy change implies that instrumental use is also slow and incremental (Shadish, Cook & Leviton, 1991, pp. 449 and 53).”

Our experience in PEMD was that instrumental use is not rare but frequent, and that it is inextricably tied to the political environment in which the policy questions originate. This means that non-use and slow use are similarly frequent, not because policy change is always slow and incremental but rather because its tempo varies unpredictably, because the evaluators’ timing may be off, or because a study’s findings may be unacceptable to particular forces in a particular political environment.

We in PEMD eventually grasped that, no matter how credible the evaluators, some problems of use would always be there. This is simply a fact of life in accountability or knowledge evaluations, because evaluators must often undertake studies that offend powerful people, because methodological risks are present in any study, and because, on any particular day, so many subjects compete for attention that even very good evaluations need luck if they are not to get lost in the shuffle. Even though the framework for practice that we developed in PEMD had originated in a desire to maximize use, and even though most of our studies were well and appropriately used, we came to realize over time that too much emphasis on use makes evaluators dependent on the favor and good graces of users, pushes them toward accommodations with stakeholders, and puts a premium on the immediate acceptability of findings. My guess is that the much greater risk to our field is not lack of use for the right reasons, but rather a declining capability or willingness to question conventional wisdom, which is our most important task and the best justification for our work.

III. IN CONCLUSION

I have argued for an evaluation practice that recognizes its sources in politics but keeps its independence; that uses the history and underlying values of a policy or program to maintain that independence; and that makes some changes in strategy within the evaluation process to help it survive and flourish within its political environment. Among those changes are the nine I have just discussed:

1. Being more acutely aware of political dynamism and contentiousness;
2. Identifying evaluation’s role as purveyor of honest information in a sea of advocacy;
3. Stressing timeliness and flexibility;
4. Making rare use of absolute judgments;
5. Emphasizing impartiality in doing the work, and credibility in its perception by others;
6. Using goals and objectives as indicators of program history, embodied values, and past political accommodations;

7. Fostering inclusiveness with regard to stakeholder viewpoint, subject matter, program costs, type of policy question, backward or forward focus, and methods choice;

8. Expanding the criteria for deciding to conduct an evaluation by emphasizing political as well as methodological feasibility; and

9. Justifying evaluation less by the use that is made of its findings (although use is important for improving policy making) than by its success as a provider of information in the public interest.

I have also taken this logic, these changes, and a set of experiences in conducting evaluations, compared them to relevant aspects of current theory, and found many points of disagreement. This should not be too surprising because, in developing theory, each step builds on the one before it—whether positively or negatively—and each step exists as a result of the one that preceded it. As we shift or place the latest building blocks, it is easy to forget how dependent we are on those who have gone before, how much they have created that we use without even thinking about it, and to imagine that we know more now than they did. But as T. S. Eliot remarked about dead poets, “They are that which we know.” Critics of the earlier steps are not wiser, only later.

It is also true that our field is still quite young, that there are different kinds of evaluations needing different kinds of precepts, and that the experience I have described here is recent. Still, the lack of congruence does generate some concern about how much other relevant experience (in developmental areas, for example) may have been omitted from our generally accepted theoretical formulations, and whether evaluation is, in fact, a practice-driven field.

Further, because theory is a body of knowledge, a set of rules or laws at which we arrive by testing propositions, in most fields the rules or precepts are explicitly related to those propositions. But in our own field, where the materials of practice, represented by completed evaluations, are the propositions on which we build our theory, it is not clear which evaluations and which types of experience, and which tested propositions, are the explicit sources of our theory. This leads to a rather ad hoc situation in which evaluators must determine which precepts apply to their particular kind of evaluation. If it turns out, as in the case of PEMD, that there are just too many anomalies in trying to match practice to theory, then each time they are placed in new organizational circumstances, evaluators must develop an appropriate theoretical framework, de novo, to guide their practice.

We are making some advances. Evaluators now know where to get information about the theory that our best minds have produced, so that, like Beethoven and his parallel fifths, we can at least learn the rules before we break them. I am referring here to the pioneering and extremely useful text by Shadish, Cook and Leviton (1991), which now gives us a convenient basis for confronting current theory. But it does seem at least somewhat urgent for us to think of ways by which new ideas and experience can be more regularly folded into the general knowledge base and exposed to view. Perhaps a few more efforts like “The Foundations of Program Evaluation” can move us ahead in that direction.

This is not to suggest that better theory will solve our problems in doing evaluations. To paraphrase Samuel Johnson, evaluations are like watches: the worst is better than none, and the best never seems to go quite right. Moreover, our central task, telling the truth, is always going to be hard because, in a political environment, so many people have so many strong reasons not to want to tell it or hear it, and so many good weapons to ensure that it does not...
emerge. The burdens and opportunities of this task are irremediably ours, but we can make it easier for ourselves if we have experience-based theory to guide us, if we strategize well to get our voices heard, and if we accept momentary defeat with respect to use when other voices are just too loud.

Telling the truth, especially to power, is a critically important function in a democratic society. It is why many of us came to evaluation in the first place, and it is the most meaningful and charismatic part of our work. As Hannah Arendt wrote in a letter to Mary McCarthy, truth is "not the end of a thought-process, but the very condition for the possibility of thinking" (Segal, 1997, p. 14). If that is not something worth fighting for in these last thousand days of the century, then I cannot imagine what is.

APPENDIX 1

Twelve Questions for Building Evaluation Designs (Chelimsky, 1982)

1. What are the main characteristics of the policy or program to be evaluated? What is its history?
2. What is the background of values embodied in the program? What is the degree of controversy today? Who are the major stakeholders?
3. What is the policy question or questions we've been asked to answer?
4. What is the size and quality of the knowledge base for the program, how credible does the past research seem, what are the major unanswered evaluation questions about the program? Are there relevant extant data we can use, or do we have to collect original data?
5. What are the alternative methods that could be used to answer the question(s)? Does an evaluation seem feasible, a priori?
6. If the evaluation seems feasible, should it be done? (That is, will the evaluation tap into a key element of the particular problem addressed by the program? Does the likely knowledge gain seem significant? Does a policy fix exist, or is it the kind of issue about which little can be done? If a policy fix exists, is it politically and otherwise implementable? Or is it wise to do the evaluation for its own sake, because of its intrinsic importance, even if no policy fix is available or implementable, and even if opposition rather than policy use is likely to be the evaluation's immediate fate?)
7. If the evaluation results are different from what the sponsor hopes, will the findings still be useful to that sponsor? What other potential users are there? Who will oppose the evaluation and why?
8. What kind of evaluation design would be appropriate? What data uncertainties will need to be confronted? Is methodological innovation involved, and if so, how much triangulation of methods will be required to assure political credibility?
9. What will the study cost, approximately? How long will it take, approximately, to complete it? Do the answers to both these questions make sense in terms of the sponsor's needs?
10. What strategies will be needed to assure study legitimacy (e.g., a politically balanced advisory board, consultant help, special presentation or dissemination tactics)?
11. Are there likely knowledge gains for evaluation practice, to be derived from, say, the use of a new method, new measures, original data, or combinations of data from different sources?

12. Overall, does the evaluation seem to be worth doing? If not, would it be worth doing if the questions were modified?

REFERENCES


Fascell, D. B. (1990, June 14). Chairman, the House Committee on Foreign Affairs. Letter to the Author.


**Notes on USGAO Publications**

1. Listing is in chronological, not alphabetical order.
2. All GAO reports are published in Washington, D. C.
3. IPE (Institute for Program Evaluation) was the initial divisional abbreviation for PEMD (the Program Evaluation and Methodology Division) prior to January, 1984.