American Journal of Evaluation

A Clash of Cultures: Improving the "Fit" Between Evaluative Independence and the Political Requirements of a Democratic Society

Eleanor Chelimsky

The online version of this article can be found at:

Published by:

$SAGE

http://www.sagepublications.com

On behalf of:

American Evaluation Association

Additional services and information for *American Journal of Evaluation* can be found at:

Email Alerts: http://aje.sagepub.com/cgi/alerts

Subscriptions: http://aje.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations http://aje.sagepub.com/cgi/content/refs/29/4/400

A Clash of Cultures

Improving the "Fit" Between Evaluative Independence and the Political Requirements of a Democratic Society

Eleanor Chelimsky

Good morning, everyone. Today, I want to talk to you about cultural clashes, about what happens when evaluation meets politics. This is not something we've emphasized in any great specificity or detail at past American Evaluation Association conferences. Mostly, we've been content to point out that evaluation and politics do influence each other in broad and general ways and then go on to more urgent and weighty matters like the pros and cons of individual evaluation methods. I don't mean to suggest that our methodological debates have been unproductive, even though they remain largely unresolved. Indeed, they show the importance we attach to achieving credible outcomes in our work, and they help clarify a fairly complex craft. However, there is some danger that a disproportionate preoccupation with methodology, over which we have a lot of control, could distract us from the larger problem of evaluation's political context, over which we have very little control, but which hugely affects both our processes and our outcomes as a field.

The fact is that most of us, who spend important parts of our lives evaluating state or federal programs, may well think from time to time about the political milieu in which our work unfolds, but mostly in terms of unpredictable difficulties that confront us without warning as we wend our way through the evaluation process. Naturally, we see these difficulties as highly specific and one of a kind, but in fact, there's a *general* circumstance that's fundamental to our problems and that's this: Because our government's need for evaluation arises from its checks-and-balances structure—which, as you know, features separation of powers, legislative oversight, and accountability to the people as protectors for individual liberty—evaluators working within that structure must deal, not exceptionally but routinely and regularly, with political infringements on their independence that result directly from that structure.

The irony here is that "independent evaluation" is a hallowed instrument within the political ideal of democratic societies, whether we refer to *accountability* evaluations performed to implement legislative oversight and inform the public, or *knowledge* evaluations to produce new information intended to influence government policy, or *development* evaluations to support an agency's mission and protect its independence. Yet in all of these cases, we've learned that politics is more than likely to impinge on *our* independence, sometimes throughout a study, sometimes at a particular step in the evaluation process (say, the choice of a methodology in the design phase).

Of course, there's nothing very startling about this. In fact, what should surprise us would be the absence of pressure on evaluators to make an agency "look good," or the lack of effort by agency managers to try to manipulate the work of evaluators implementing legislative oversight. These are simply characteristics of our political environment, and they hold true for Democratic as well as Republican administrations.

Unfortunately, the training we receive assumes unthreatened evaluative independence, and it concentrates on methodology, not milieu. This puts heavy emphasis on the technical merits of one evaluation design versus another, often considered without reference either to

400

the origins of the political question posed or to the reigning political environment, and it equips us poorly to recognize and deal with even quite ordinary issues in the world of governmental and bureaucratic maneuvering. Yet these issues can critically affect evaluative independence, and through it, the quality of planning, data analysis and reporting, as well as the overall credibility of our work.

We continue to imagine the political sphere as somehow divorced from us, extraneous to the technical nature of evaluation, forming a sort of benign backdrop to our work. But I would argue that evaluators practicing in the public domain can never escape the world of politics, first, because it's precisely that world that gives our work legitimacy, authority, and the potential for real consequence, and second, because the governmental structure that triggers the need for evaluation in the first place *also* triggers the political maelstrom in which evaluative independence is so often tested.

Specifically, evaluators derive their legitimacy, and also their role and mandate, their need for technical competence and credibility, along with their right to independence, from the political notion of accountability in government: that is, the idea that governments are responsible to the people for their actions. Accountability also requires evaluators to publish and disseminate their findings. This means that evaluations in the public domain need to have two components: one that answers a particular policy, program, or knowledge question and another that informs the public of what's been learned. But because this latter duty is not part of evaluation per se, and is instead a necessity of its political milieu, it hasn't received a lot of attention from evaluators. This is a gap we will need to fill as we go forward.

So today, I wanted to reflect a little bit about these political aspects of our work, about the kinds of clashes that occur on a regular basis between evaluative independence and the political culture it challenges, along with possible ways to predict, parry, or even avoid some of these clashes. To do that, I'll be drawing on my own experience in conducting both executive branch and legislative branch evaluations, part of which comes to me from running the Program Evaluation and Methodology Division (we called ourselves PEMD) within the Government Accountability Office, or GAO, over 14 years.

But before telling you how we tried to resolve some of the problems I've mentioned, I think it's important to examine how they've arisen and evolved, and to try to pinpoint where the major clashes of culture have tended to occur. This means we need to look a little more closely at how our governmental structure affects evaluation.

Government Structure and Evaluation

In this country, as everyone knows, we have a government built on purposeful tensions, whose functions are carefully split among different branches. By adopting such a structure, the framers of our Constitution hoped to keep too much power from accumulating in any one place or any single pair of hands. This is an enlightenment architecture born of distrust, based on past experience with a coercive central authority. But it's also the fruit of compromise between those framers most concerned with maintaining a powerful, effective government and those most concerned with preserving individual liberty and preventing corruption and abuse. That compromise is still with us after more than 200 years, and the same tension is reflected in evaluation, because we, too, must be concerned with promoting an effective government, while also looking out for the public interest (Ellis, 2002, pp. 7, 9, 15-16).

The framers, then, produced a divided and cross-divided—incredibly fragmented—governmental structure, featuring both *distribution* of powers between federal and state levels

of government and *separation* of powers at the federal level itself (among executive, legislative, and judicial branches). That is, this architecture provides *external* controls on centralized authority by partitioning power between state and national governments, and among the various federal branches. But the framers also instituted *internal* controls, such as the independence of individual branches, departments, and agencies, as a further check against centralization. Madison (1788/2001) wrote that "each department should have a will of its own" and wanted the interior structure of the government to be so contrived "that its constituent parts would, by their mutual relations, be the means of keeping each other in their proper places." So, and this is a very important point, we can see that agencies not only have an excuse but a political mandate to be independent.

Fortunately, the framers also understood that too much agency independence was dangerous to the liberties they were trying to protect. That is, with independence as a pretext, agencies could well decide to distort, censor, classify, or simply block the publication of information that might embarrass them, thereby creating bureaucratic walls around themselves that do, in fact, defend them from scrutiny but simultaneously close off openness in government and expose the agencies to public suspicion about what they might be doing. Patrick Henry (as cited in Suleiman, 2006) warned in 1787 that the liberties of a people can never be secure "when the transactions of their leaders may be concealed from them." Thus, the framers, having encouraged agency independence with one hand, established a check against it with the other, in the form of *legislative oversight*: that is, the particular authority to supervise the administration of government that has led to so many battles over the years and continues to do so as we meet here in Portland today.

Our legislature has oversight over both judicial and executive branches. For the executive branch, which is what concerns us most here, this authority is exerted through mechanisms like the appropriations and impeachment powers, approval of some nominations to executive office, and—of special interest to evaluators—the investigation of how well past legislation has been implemented: with what fidelity, integrity, efficiency, and outcomes. It's from this investigative authority that independent evaluations spring. Indeed, for legislative oversight to work, evaluations (along with audits and other types of analyses) are essential for answering the questions it raises.

In short, then, our government builds in democratic protections through tensions created by fragmented powers, checks and balances, agency independence, and legislative oversight.

But this structure depends for its authority on the support of a well-informed public: that is, an electorate that possesses the willingness and capability to debate, protest, and correct problems in government once they become known. Again, this idea—like the others on fragmentation, independence, and oversight—comes down to us from the framers (and from Montesquieu before them). Informing the public, then, is a basic component of accountability: The government's responsibility to the people for its actions naturally includes telling them what it's been doing. And accountability also requires evaluators to report their findings to the public. I mentioned earlier that we haven't always been very successful at this, but I think the fault is only partly ours.

First, performing evaluations within a political environment has sometimes been so arduous that we've had to channel most of our energy into producing sound and useful work in an inhospitable climate. Second, the primary responsibility for informing the public has traditionally fallen to the legislature. Evaluators have usually been satisfied simply to get their reports published and into the public record: That alone has been a heroic enterprise in some political circumstances.

Third, disseminating information has largely been left to the press: Journalists have had a consecrated role in this area since the time of Jefferson. Indeed, even after 8 years as the U.S. president, Jefferson's faith in the press was still intact. He wrote in 1816:

> The functionaries of every government have propensities to command at will the liberty and property of their constituents. There is no safe deposit for these but with the people themselves, nor can they be safe with them without information. Where the press is free and every man able to read, all is safe. (Jefferson, as cited in Padover, 1946, p. 9)

Today, things seem less clear. An enormous information industry has grown up since the time of the framers, and this industry, now beset with competition and striving to maintain profit margins, is more concerned with stories that sell papers or attract viewers than with informing citizens about their government. Although the press and other media—and especially some highly knowledgeable individual journalists—are still excellent transmitters of evaluation findings, this is more likely to be the case when the study subject has great popular appeal. But when the media view a study as tedious or highly complex, it's entirely possible the public may never hear of its findings.

Yet there's no doubt that evaluations of governmental activities exist to inform the public, no matter who commissioned the study, who asked the policy question, or which branch of government expects to use the findings. In other words, the ultimate client or user of our work is the public. Therefore, with only varying ability to rely on the press, evaluators, in the pubic interest, may need to think of new ways—and take on new duties—to ensure the appropriate dissemination of their findings.

Let me summarize, then. I'm arguing here that there are five characteristics of our government structure that together generate the need for independent evaluations. These are:

- a fragmented architecture,
- the independence of individual government units,
- the creation by agencies of bureaucratic walls to protect their information,
- the exercise of legislative oversight to investigate government activities, and
- the need to ensure that citizens are informed about the results of these investigations.

This framework of five issues sets up the political context from which we can infer the place, the form, and the function of evaluation in government. It sits at the heart of the tension between the need for governmental power and the need to preserve liberty and promote integrity in the exercise of that power. Yet to be meaningful in this political universe, evaluations must be competent and credible in their own right. So evaluators must somehow accommodate both the political culture within which they work and the *independence* from that culture that makes their work politically valuable. It's this tangle that we need to unravel if we want to understand and deal with:

The Regular Clashes That Arise Between Evaluative Independence and Our Political Environment

Evaluators focus, first and foremost, on getting their work right. Evaluations almost always involve scrupulous, iterative struggles to rule out bias: bias in the data, in the analysis, in the findings, in the presentation. Obviously, we have to be technically competent if our studies are to be methodologically sound. But this same competence is also needed for political persuasiveness, because flawed studies don't stand up well under hostile scrutiny. However, to

achieve technical competence, evaluators must have independence in their design choices and in their planning, performance, and reporting. So we need independence both to do good evaluation and to sustain political credibility. Yet in practice, independence is often threatened by the ongoing tensions of politics.

For example, independence can be jeopardized simply by the political origins of some questions, especially when it's hard to negotiate study issues with sponsors. Yet when we cannot refine and specify both the evaluation questions and the study design for answering them, the consequence may be not only loss of independence but also problems of credibility when the time comes to defend the evaluation's methodology, not to mention other difficulties involving infeasible, untimely, or overcostly studies. In some cases, study questions may hide a partisan or ideological purpose for the evaluation; this naturally brings expectations among sponsors for a particular set of findings, raising still another threat to independence. And in a more tactical way, political pressures—say, to meet a deadline—that intervene in the choice of the most appropriate design or prevent the collection of needed data, again strain independence, often with cascading effects on credibility and future use as well.

Furthermore, no matter how worthy the study being planned, or how exemplary the evaluation questions and design, the officials in any administration whose programs are being examined may be less than happy about the project. Fear of embarrassment can trigger roadblocks that both impede a study's performance and weaken its findings. That is, if the "targets" of an evaluation decide to protect themselves by hiding certain data or obscuring their decision processes, this not only affects the validity of the evaluation but also its independence, through the specter of advocacy that it raises. Missing or distorted data allow the perception that there's a relationship between somebody's presumptive interests and the direction of the evaluation finding. Yet it's precisely to avoid such bias that evaluators insist on independence in the first place and that legislators turn to evaluators within the oversight process.

On the other hand, it's also the case that data may be distorted from *within* the evaluation, by a politically unwitting evaluator. When we're not aware enough of the context of our work, we can lose sight of the need to deal with all of the stakeholder and other special interests (not forgetting the public's) that are involved in an evaluation. If some of the relevant voices are not as loud as others, they may not get a proper hearing within a study or may even be ignored and omitted entirely. This, of course, invites the same problem: the potential perception (or actuality) of advocacy in the findings.

Finally, clashes with our independence may occur during the reporting or publication of an evaluation. Here, sponsors and stakeholders can apply political pressure to make changes in report findings, or language, or presentation, or simply try to delay the study's appearance. This isn't just a problem for evaluative independence, however: It also severely affects the public's right to know.

My point here is that clashes with evaluative independence are facts of life in a political environment. Executive branch evaluators with favorable findings on an agency program should always expect intensive grilling by legislators who are opposed to the program, just as legislative evaluators should expect obstructionist behavior from agency officials. By recognizing that, we can better pinpoint at least some of the politically problematic places in every study; we can better prepare to counter the predictable attacks on our independence; we can better preserve our flexibility of choice; and overall, we can achieve a better fit between politics and evaluation.

Where then, specifically, can we expect these clashes to occur? And *how* do we prepare for them, at least in general terms, so as to head them off, perhaps, or prevent them from neutralizing our work? Evaluators won't always win, of course, in trying to maintain their independence against much more powerful forces. Still, we do have some effective

weapons to deploy in the fight, and among these are thoughtful, *realistic* planning, the capacity to negotiate, a reputation for technical credibility, and a ferocious unwillingness to be intimidated.

When I look at our experience in PEMD across nearly 300 evaluations over 14 years, my sense is that clashes occurred most commonly in just those places where you would have expected them to occur:

- at the design phase,
- during the conduct of the work, and
- at the end of the evaluation.

So first, then, let me examine some

Clashes Arising During an Evaluation's Design Phase

In PEMD, we regularly received requests to answer what can only be called "loaded" evaluation questions. Naturally, this happens because sponsors have a clear idea of what they want evaluators to find and don't mind telling them so. I remember one such question we received from a congressional committee: To what degree has the Secretary of Education distorted the evaluation findings on bilingual education? Note that we were not asked to say what the evaluation findings were, or to assess their soundness, or examine *whether* they'd been distorted. We were expected to simply assume the distortion and just say how big it was.

We negotiated with the committee, emphasizing the usefulness to *them* of being able to attach at least some credibility to our findings, and ended up with the "whether" question (i.e., whether the findings had, in fact, been distorted) and an excruciatingly transparent study design. We reviewed the body of bilingual evaluations, consulted with a politically well-balanced group of experts in the field (some of them suggested by the Department of Education itself), and compared the results with the department's statements on the subject. The point here is that the initial politicization of the question forced us to bend over backward in our attempt to use clearly described, simple methods that would not only *be* fair to the department but also *seem* fair, while simultaneously, of course, trying to reinforce our own reputation for credibility with the Congress and with the agency.

In another example, I recall a congressional request letter that asked 19 questions about the effects of giving bovine growth hormone to cows. Our answers to these complex and diverse questions were expected in 3 months, the letter said, and we were to use specific consultants, chosen by the committee and listed in the letter (complete with telephone numbers). On the very last page, at appendix, the 19 questions appeared again, but this time they were arrayed against 19 *answers* the committee wanted us to find. So again, negotiation, leading to a reasonable evaluation question: We would assess the effectiveness of the Food and Drug Administration's *review* of bovine growth hormone products and develop *our* findings in accordance with *our* study design, data, and analysis.

Unfortunately, negotiation didn't always work, and in some cases, I had to turn down study requests. I learned to present our reasons for saying "No" to the Congress very carefully, and sometimes our conclusions were accepted, sometimes not. On one occasion, Senator Kennedy wrote a letter calling me "recalcitrant" because I'd refused a request of his to estimate the future impacts of a new provision in the immigration legislation: that is, without any historical data on which to base the estimates. In another example, I explained, in a detailed 10-page letter to a House requester, the complicated basis for our refusal of his request. He

then faxed my letter, without my knowledge, to researchers all over the country, accompanied by the query, "Is she right?" The first I heard of this was months later, when I got a new letter from the requester, enclosed in an enormous 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?

None of you will be astounded to learn that, in both cases, we were eventually obliged to do the studies. But the refusals turned out to have been extremely worthwhile, because they brought us what we hadn't been able to achieve without them: a serious reconsideration of the request; major changes in the questions to be answered; appropriate timelines; and agreement on intensive *bipartisan* committee participation, in the one case, and extensive executive branch consultation, in the other.

I should mention that *stakeholders* also tried to influence us as we began certain studies, but because they were powerless—in the 1980s, that is—to affect our questions, they'd usually focus on our study design. In one case, we were lobbied for 6 months by an association of medical-device manufacturers. They wanted us to change our intended plan for examining the quality and volume of postmarketing surveillance information. What the Congress wanted us to do was to determine the problems, and especially the fatalities, that were attributable to particular medical devices. The alterations the manufacturers wanted to impose on our design would not only have doubled the time needed for the study but would also have totally precluded our ability to answer the legislative questions (from their point of view, of course, a consummation devoutly to be wished). So once again, we argued credibility—in this case, the credibility of methods that have *some* chance of answering the questions posed—and after a long and tedious exchange of letters that seemed to be going absolutely nowhere, the argument suddenly ended. Not in a draw, as I'd expected, but in the group's acceptance of our design. I think we convinced them, not that we were right, of course, but that we would stick to our guns.

A second place where politics and evaluative independence collided regularly in PEMD was located beyond the design phase. These were the

Clashes That Occurred During the Actual Conduct of Our Work

As evaluations moved toward data collection, for instance, we'd get letters, phone calls, or visits from "interested parties," who wanted to dissuade us from doing a study, or from doing it as we had planned, or from coming to conclusions that might be embarrassing to them. You could codify a whole repertoire of techniques here, going from apparently rational arguments for changing a study's focus "just slightly," through genteel and delicate intimations of trouble ahead, to the disappearance or classification of data essential to the evaluation, and finally, ad hominem attacks on the evaluators, sent to the Congress or the media.

In most cases, resistance was passive or diffused and hard to document, but one example of real harassment we encountered involved the National Cancer Institute (NCI). From the beginning of our work on cancer survival rates, the agency treated us with such antagonism that I added a physician and a lawyer to our team, made strenuous personal efforts at mollification, and *tripled* the customary layers of methodological review for both the evaluation design and the final report. Yet even though our conclusions were quite favorable to the agency, the steady stream of invective that pursued us during the evaluation continued after the report was published. The NCI director called it "politically motivated," "unfair," and "offensive," and of course, "methodologically flawed."

Nobody seemed to notice the strangeness of this situation until the magazine *Public Interest* picked up on it and wrote:

Such accusations are not unknown in the medical world. What's surprising in this case, however, is that virtually the entire national press *and* the NCI have misunderstood the purpose and conclusions of the GAO report. Although the report did acknowledge that the extent of improvement in patient survival in specific cancers is often not as great as reported, the chief finding was that patient survival has in fact *improved* in all but one of twelve types of cancer that the GAO surveyed. Further, the GAO report has been reviewed by a number of well-respected cancer researchers who all commended it for its balance, methodology and accuracy. How then to explain the reception given the report? ("Malignant Reporting," 1987, p. 150)

Well, I can tell you that the reception given the report, which simply continued the struggles we'd had just to carry out the evaluation, had more to do with structural relationships between agencies and the Congress than it had to do with anything in the report. The NCI director and his staff were so embattled, so focused on protecting their independence against what they saw as "another congressional intrusion" (otherwise known as oversight) that they were unable to review the report with any kind of detachment. In other words, our evaluation was seen not as a study in its own right but rather as the detested instrument through which the Congress implements its constitutional mandate to oversee the activities of federal agencies. Of course, you could also say that the NCI was exerting its *own* right to independence, in the best Madisonian tradition.

This experience shows us two things: first, the importance and legitimacy of evaluation's place in government, and second, the inevitability of structural pressures as it takes on its political role. At a more mundane level, of course, it also shows the value of thinking out ways of parrying political problems early on in an evaluation. The added interweaving of reviewers (specialists in medicine, law, and statistics) into the planning and conduct of the entire evaluation not only saved us from extinction in the cancer wars but maintained our credibility and our negotiating position in future battles.

Another way that executive branch managers resisted both legislative oversight and our evaluations was by simply making the data disappear. Between 1980 and 1994— that is, across the Carter, Reagan, Bush, and Clinton presidencies—we found that secrecy and classification of information were becoming prevalent in an increasing number of agencies. Yet it would be hard to find a more critical issue for evaluation than this one. After all, if evaluators ignore or can't get access to classified information needed for a study, their findings will likely be invalid. In PEMD, when we started our work on chemical warfare, we found that all the open literature had been written by "doves," all the secret material signed by "hawks." This meant that doing a synthesis of information in this area forced us to include classified data, with all the subsequent inhibitions on reporting, dissemination, and the public's right to know that classification automatically entails.

The irony is that the threat to validity posed by classification often has little to do with national security but rather with the dark side of agency independence. Agencies want the world to see what makes them seem virtuous; they hide under a blanket of secrecy whatever might cause problems, or highlight warts and foibles, or contradict a budget request. We found, for instance, in a review of test and evaluation of weapon systems at the Department of Defense (DOD) that once we'd examined all the classified information, it was clear that what DOD had released on an unclassified basis "resulted in a more favorable presentation to the Congress of test adequacy and system performance than the facts warranted" (GAO/PEMD, 1988b). I ended up testifying—at an open congressional hearing where classified information could not be introduced, and with an irate DOD program manager sitting next to me—that yes, the unclassified information available *was* favorable to the new weapon system for which funding was being sought, but only because the unfavorable data had been suppressed and because older, still serviceable, alternative weapon systems had been presented as obsolete.

Classification, then, by its selective release of data, not only threatens legislative oversight and public accountability but also puts evaluators in a terrible position. We can tell the truth and go to jail for revealing classified information; we can tell only truths that are palatable to the agency; or we can take so narrow an approach on such unimportant policy issues that nobody cares, not even the agency.

Of course, no one contests the idea that some information, such as the design of advanced weapon systems, should be classified. But we're now seeing a metastasis of secrecy, far beyond intelligence and the military, into unrelated domains like environmental impacts, or hospital error rates, or drug side effects, or student test scores, where no national security interest can possibly be invoked. And we're seeing extensive reclassification of materials that had already been declassified. This is a critical issue both for evaluation and for government, because it precludes the examination of all the facts, it inserts involuntary advocacy for agency programs and policies into "independent evaluations," it makes a sham of executive accountability to the legislature, and it drastically distorts the public's knowledge of what the government is doing.

Now let me speak to the kinds of

Clashes That Are Typical to the Final Phases of an Evaluation

When sponsors are unhappy with study conclusions, this can be a stressful situation for evaluators, whether in the legislative or executive branch. I often think of David Kay, who used to be head of evaluation for the International Atomic Energy Agency (IAEA) in Vienna, and how he must have felt when he had to tell President Bush that no weapons of mass destruction had been found in Iraq. We all know, or know of, evaluators who've been fired for bringing in the "wrong" answer and of evaluation reports that were changed to fit somebody's political agenda or were quashed outright.

In PEMD, we had many disappointed sponsors, but I'm happy to say that in most cases, we were surprisingly well treated. Senator Helms, for example, adopted our unwelcome advice on the Women, Infants, and Children's (WIC) Program. Before we began our study, he had announced that he would seek major cuts in the program's budget. But after our favorable findings on infant birth weight, he kept the funding level. In another instance, two congressional sponsors who were enthusiastic proponents of Enterprise Zones found our negative conclusions hard to accept but accepted them. For both of these studies, we had anticipated political problems at the design phase and multiplied our validation efforts. But in the Enterprise Zones case, I think the survey of employers that we tacked on at the end of the evaluation—to corroborate (or invalidate) in transparent fashion, the findings from an interrupted time-series design—greatly improved the persuasiveness of a politically unwanted conclusion (for WIC, see GAO/PEMD, 1984; for Enterprise Zones, see GAO/PEMD, 1988a).

Only one congressional sponsor ever asked us to actually change our findings. The year was 1989, and he'd requested that we examine whether or not American businesses were sustaining an increased burden of paperwork. Our study found no real change in burden, only changes in accounting procedures that made the burden *seem* bigger. So our sponsor calmly suggested that we just stop short at the raw data, without correcting for the altered procedures, so he could make the case for business mistreatment at the hands of government. Obviously, we refused, invoking a truly egregious credibility problem, and after a fair amount of negotiation, we were able to publish our report unamended.

As I've just said, the awareness of sponsor discontent, imminent or actual, weighs heavily on evaluators, at least in part because it bodes ill for the use of the findings. But the problem

doesn't end there. Unhappy sponsors don't normally alert the press, or hold hearings, or otherwise inform the public, so when conflict occurs, accountability to the public suffers, and evaluators must then give special attention to disseminating their findings (depending, of course, on their importance) to the public at large or to specific segments of that public.

Stakeholders may also bring political pressures to bear on evaluators once the findings are published, and although the pressures are usually predictable and can be planned for, their effectiveness varies. The gun lobby, for example, is powerful, efficient, and greatly respected on the Hill. They really did us in, successfully stalling a hearing we were supposed to have on an evaluation that showed that many deaths and injuries caused by firearms in accidental shootings could be prevented. The study was important not only for its findings but also for its methodological interest, merging, as it did, data that had never been combined before, across police and medical information systems. But the sponsoring Senate Committee, which had originally deemed the study a priority, just melted away like mist as the lobby increased its pressure. So the evaluation, finished in 1991, *was* published but had to wait for a new administration to use its findings some 3 years later.

On the other hand, the beer lobby was much less successful. Our requesting committee in the House and especially its chairman, Congressman Oberstar, stood up to their lobbyists with regard to our work examining the effects of drinking-age laws on traffic fatalities among youths under 21 years of age. *This* hearing was held as planned. I did get a barrage of lobby-oriented questions, but they were easily answered, and the study went on to receive remarkable press coverage and eventual use in a Supreme Court decision.

We learned from this experience that the capability and willingness of a legislative committee to ask the right questions, to support credible evaluations, and to stand up to crushing stakeholder pressure are absolutely essential elements in any equation involving evaluative effectiveness.

What can we conclude, then, about how to deal with the conflicts inherent in preserving evaluative independence while also being effective within the political structure that drives and defines our work? I have five suggestions to make. Although they derive largely from experience in working with the Congress, I think they're applicable in most situations where "independent evaluation" is the ideal as well as the theory but where reality generates systematic attacks on that independence in practice. Here's what we did in PEMD.

First Suggestion: Expand the Design Phase

We created a new component at the beginning of the design phase in major evaluations to include an analysis of program histories and values, past and current political controversies, and probable stakeholder positions vis-à-vis the projected work. This did a number of things for us:

- It integrated our political and methodological thinking with respect to opposing values concealed within study questions.
- It allowed us to predict, at least for some studies, *whether* there would likely be clashes strong enough to trigger political intervention in our work.
- It showed us which *parts* of the evaluation were politically or methodologically weak, and hence vulnerable to attack, so that we could strengthen them.
- And above all, it concentrated our minds on always being able to defend an evaluation's *political*, as well as its technical, credibility.

We began using this expanded design-phase model in 1982, and it developed steadily as our understanding deepened with continuing practice. We learned to look much more closely at the issues lying behind the requests we received. For example, at one point, we were asked to evaluate the effectiveness of the Food Stamp Program. But although we'd confirmed in our extended design phase that participation rates *were* very low (which is what had raised the question of program effectiveness in the first place), we also realized that we first had to understand *from participants* why the rates were so low. It seemed reasonable to wonder whether there might not be large numbers of poor people who found food stamps demeaning (which raises the question of the appropriateness of the program's goals and implementation, rather than its effectiveness) or whether potential participants were just too poor to buy food stamps (raising the question of the "fit" of the solution to the problem). So, in accord with our House sponsor, we decided to focus first on a detailed examination, by population group, of both participation and nonparticipation in the program.

Second Suggestion: Include Public Groups in Evaluations, When Relevant

In PEMD, we found that the inclusion of public groups needed to be considered

- when those groups have knowledge that informs study questions;
- when there's stakeholder conflict within a program, and the risk exists that some views may be drowned out by others; or
- when the program itself is situated at the heart of an ideological conflict within the larger society (as in food stamps, for example, or programs seeking to equalize educational or employment opportunity).

Many evaluation questions, after all, ask about the effects of programs on the behavior or well-being of particular population groups, and, of course, these groups possess information of the greatest importance for evaluators. Yet in 1980, when I arrived at GAO, it was common to see evaluations of service programs—programs serving the aged, for example—that never interviewed a single elderly person receiving the service. The reason I was usually given for this gap in data collection was the evaluators' concern that by involving public groups, they could be perceived as advocates for them and thus lose some of their credibility in a political environment. Still, they had no difficulty with the idea of interviewing program managers and other agency officials, so that the omission of the elderly clearly led to what Joseph Stiglitz (as cited in Nordhaus, 2004) called "asymmetric information" in the evaluation findings. This is a serious problem for both technical and political credibility, as serious, in my judgment, as any resulting from the denial of data by agencies. Although the sources of the problem are different, as well as the motivations and the direction—one is endogenous, the other exogenous—nonetheless, the exclusion of beneficiary data afflicts an evaluation in much the same way as does the exclusion of data through classification.

In PEMD, then, we involved the public, first and foremost, to improve the validity of a study. Some of the groups we queried were agricultural workers, discharged surgery patients, disabled people, directors of corporations, runaway or homeless youths (as well as their parents), and program participants generally. We included these groups as sources of information, not as designers or conductors of the evaluations themselves. We never allowed self-selection, because of its potential damaging effects on credibility, and we retained control of all our study decisions. The endeavor was innovative for its time (early 80s), and it was

certainly cautious, to avoid both the fact and the appearance of advocacy. But overall, we found that the inclusion of public groups in an evaluation was no barrier to its credibility. On the contrary, credibility increased because of the breadth and balance these new voices brought, countering in many cases, the voices of other, more politically powerful stakeholders. Their inclusion also brought us notable technical improvements, information that was often unique, and some important ongoing questions about the best ways to combine data from these multiple, disparate sources.

Third Suggestion: Lean Heavily on Negotiation

We learned to prepare for, develop, and count on a capability for argument, because in politics, most things can be talked through, even if they don't initially appear subject to compromise. As I said earlier, we were often successful in altering biased evaluation questions, but there were other problems as well that we needed to negotiate, things like vagueness about the issues, unclear feasibility with respect to the work, and sometimes, extreme uncertainty about dissemination and use of the final product.

Negotiation was also important with executive branch agencies. Sometimes this was about access to data, but more often it involved agency reaction to our study designs and draft reports, which we always submitted to them for their review. Here PEMD staff members were frequently put to the test because we experienced more than a little outright hostility, and an us-against-them attitude that was more related to our representation of the legislative oversight function than it was to the particular evaluation we happened to be conducting. We taught ourselves to prepare agency negotiations carefully; to keep in front of our minds what we needed to accomplish; to adhere to a planned sequence of bargaining positions; to expect efforts at intimidation; and to react to them, both by speaking out and by insisting on our *right* to speak out, especially in well-staged adversarial situations (40 agency staffers, for example, versus two PEMD evaluators).

Stopping negotiations flat, when necessary, is always an option, of course, but probably the most important thing for an evaluator to convey is an unwillingness to be intimidated, even when it's clear the outcome may not be a happy one. This is because a position taken *now*, on one evaluation, will factor in to future negotiations with the same groups. You not only have to worry about getting a successful agreement this time but also about preserving your options and strengthening your reputation for next time. So, standing up to intimidation and occasionally saying "no" can remarkably enhance your negotiating posture. But then again, so can strong arguments, tenacity, and a well-known history of winning.

Fourth Suggestion:
Never Stop Thinking About Credibility

The strongest defense for an evaluation that's in political trouble is its technical credibility, which, for me, has three components. First, the evaluation must be technically competent, defensible, and transparent enough to be understood, at least for the most part. Second, it must be objective: That is, in Matthew Arnold's terms (as cited in Evans, 2006), it needs to have "a reverence for the truth." And third, it must not only be but also *seem* objective and competent: That is, the reverence for truth and the methodological quality need to be evident to the reader of the evaluation report. So, by technical credibility, I mean methodological competence and objectivity in the evaluation, and the perception by others that both of these characteristics are present.

With regard to individual evaluations, technical credibility seems to derive from four things, all of which require evaluative independence to make them happen:

- the appropriateness of the methods chosen for answering the questions posed;
- the honesty with which we report our confidence in the data, evidence, and analysis presented;
- the seamlessness with which our findings flow from the data and don't go beyond them; and
- the absence of advocacy in the findings and recommendations, and in the presentation and language of the report.

Perhaps the most important use we made of our extended design phase was to plan a credibility defense of our design choices, especially when the subject was politically hot, or we were trying out a new method. To bolster findings, especially unpopular ones, we scheduled extra data collection, we brought in special expertise, we amplified our methodological reviews, we involved our Advisory Board and our Visiting Scholars, and we listened carefully to the opposing stakeholders. But the critical point I want to make here is that integration with the political culture that surrounds us has intrinsic limits. We have to be able to protect our independence and our credibility. Although the enlightenment spirit of balance and compromise is surely admirable, *our* problem is that all things can't be balanced and compromised in evaluation. Findings are findings, and data support is data support. The integrity of our work always needs to be defensible—and defended powerfully—in a political climate.

Indeed, if credibility is *not* an ongoing, day-and-night preoccupation with us, we risk losing it, in countless slow erosions from principle. It's like the damage termites do: You don't see it right away, but it makes its presence felt over time.

Fifth Suggestion: Develop a Dissemination Strategy

In PEMD, we came to realize that we were overdependent on a vigorous legislature and an alert press to help us fulfill the public information requirement of our work. As I mentioned earlier, it can be very hard to get the work out when findings are complex or obscure, and under other circumstances as well. But on the other hand, it seemed clear that we couldn't possibly take on the entire burden of informing the public. Even if that had been feasible in 1980, it wouldn't necessarily have been a good idea, because there's a fine line between dissemination and marketing, and evaluators who are perceived as selling their work lose credibility.

Still, we saw a few things we could do to help our findings get a hearing. We began by thinking a lot about how to write a technical report so that not only social scientists but also policy makers, the press, and the public might be drawn to reading it. Our big priority here was the writing itself. We wanted to be sure that if people were actually interested in our latest findings, we didn't deter them by a report that was self-indulgent, unclearly focused, and full of jargon.

A second thing we did was to prepare an individual dissemination strategy for each *major* evaluation. We plotted this only in general terms during the extended design phase but then finalized it as soon as the findings—especially politically displeasing findings—were in. This strategy could feature, singly or in combination, scholarly articles in journals; simplified statements of findings to relevant interest groups; or briefings to reporters, think tanks, and other organizations. We also looked carefully at our final reports to see if any targeted dissemination might be necessary, and we followed through on this for a number of reports. These included one to high school principals all over the United States about gaps we'd found in students' knowledge of federal aid availability. Another went to public health officials

detailing ways of educating dissimilar, high-risk populations about the prevention and transmission of HIV/AIDS.

This dissemination effort had a remarkable payoff for us: It helped to keep important findings alive and potent, rather than moribund on a shelf somewhere; also, as interest in them increased, we often *got* the congressional hearing, the press coverage, and/or the policy use we'd been denied in the first place.

Today, of course, we benefit from the Internet, and this extends our dissemination reach enormously. But this kind of dissemination mostly concerns those (other evaluators, for instance) who were already professionally interested in the work to begin with: That is, it doesn't necessarily increase the flow of information to the public, in the framers' sense. However, there seems no reason why we can't develop this tool to reach a truly vast audience and thereby greatly diminish our dependence on others to fulfill evaluation's role in public accountability. A great many avenues already exist (bloggers, the "new media," and so on), and they are multiplying in size, complexity, and interactivity almost as we speak. The capacity they show to nurture open debate on issues of public accountability is exactly what we need in evaluation.

Of course, it goes without saying that the general issue of dissemination I raise here relates specifically to accountability studies (i.e., evaluations conducted to inform the public). Their publication and dissemination must be protected along with the evaluators' ability to perform them independently. We've seen in the past that a decline in publishing and disseminating these studies can be a harbinger of decline in government accountability generally. Blackstone and Plowden (1988, pp. 193, 179-189), for example, commenting on the experience of the British Central Policy Review Staff (CPRS), noted that after Prime Minister Margaret Thatcher's "horrified public reaction" in 1982 to a CPRS evaluation of options for welfare spending, "CPRS reports were no longer being published because the prime minister preferred that they not be published." Only a few months thereafter, the CPRS itself was abolished by Mrs. Thatcher, and it was generally observed that "she got rid of it because there was no longer a place for a body which challenged the prevailing orthodoxy."

On the other hand, with respect to knowledge evaluations (i.e., evaluations intended to produce new information for policy use), the publication/dissemination problem may be less acute in that obstacles to getting the information out may be less intense. As for development evaluations (whose purpose of supporting an agency's mission and infrastructure often involves access to confidential matters), it may not be necessary to publish or disseminate the findings at all, if the work serves only the agency and will be used only by the agency. Indeed, publication here could impede use.

However, even in this latter type of evaluation, an accountability issue may arise when an agency dubs a study confidential or classifies it, even though its findings are directly relevant to the public interest. Unfortunately, because of the proprietary nature of some evaluation contracts, evaluators may be quite powerless to publish or disseminate their findings on their own, without client or agency permission. Here, we should try to work out stronger contractual arrangements that can both protect client needs for confidentiality and also recognize evaluator responsibilities to the public at large.

Those, then, are my five suggestions for helping evaluators improve the fit between politics and evaluative independence: Take more time at the beginning of a study to think about political issues, especially stakeholder balance; include public groups, but with prudence; negotiate seriously, no matter how unpromising the situation; prepare a strong credibility defense; and work harder to disseminate findings to the public.

Conclusions

Let me close now by reminding you of Kierkegaard's point that life can only be understood backwards (which is what evaluators do), but that it needs to be lived forwards (which is what politicians do). This is *not*, of course, the only difference in our two cultures, as I've been telling you today, but it certainly is a crucial one: For one thing, it explains why evaluators can be accurate and politicians can't. Still, the fact is that if we want to be successful working within our government structure, we need to conciliate better the different political and technical goals for evaluation.

This leads me to two final points, one about craft, the other about the larger goals and meaning of our work.

First, we need to think systematically about our milieu and how we fit, or don't fit, into it. "Systems integration" is what the engineers call that process, and in our case, we must adapt ourselves, within the evaluation framework, to the push and pull of politics, while at the same time maintaining the independence and credibility without which we cannot do our job. The problem is that politics and evaluation are both intrinsically connected and intrinsically separated. We need independence if we're to be competent, objective, and credible. However, independence can't always be preserved without a fight; this doesn't make us a lot of friends; and yet we need friends, partners, and teamwork to be effective in a political universe. Lewis Thomas (1982) wrote that "the great successes in evolution, the mutants who have, so to speak, made it, have done so by fitting in with, and sustaining, the rest of life." We evaluators "should go warily into the future, looking for ways to be useful, listening carefully for the signals . . . and having an eye out for partners."

Second, from the broader perspective of who we are and why our work is important, we need to recognize more explicitly the role we play in public accountability. The framers knew what they were doing. If you weaken legislative oversight—that is, if the Congress doesn't ask the right questions of executive agencies, if the agencies don't respond or respond untruthfully, if we evaluators don't implement oversight in a serious way, and if those of us working within agencies allow our voices to be muted or stifled—then we should expect some nontrivial consequences for accountability in our society.

When you consider the recent massive increases in outright disinformation coming from the executive branch; when you consider also the successful purchase of legislation by various commercial or corporate interests, and the equally successful purchase of newspaper articles by executive agencies; and when you consider the cumulative and uninhibited expansion of executive branch power, and the almost visible deterioration of the controls on that power that evaluation works to strengthen; then it becomes obvious how profound, how difficult, how risky evaluation is, and how essential it has become to our liberties.

Indeed, if you asked me what was the most important thing we've achieved during the 45 years or so that evaluation and politics have been partners, I'd say you need to look beyond the individual effects we've had on specific policies or programs over time, even though many of these effects have been substantive and consequential. In my judgment, the major accomplishment of evaluation has been the establishment and demonstration of a trustworthy, dependable tool in government for carrying out and preserving political accountability.

So we have to get it right. Because if we don't try, and don't succeed, and systematic evaluation of what the government is doing becomes a thing of the past, then *our* failure would affect not only evaluation itself but also our democracy and its political freedoms. When you come right down to it, we're like canaries in the mineshaft: Our presence means that public accountability is alive and well. But if we go, the nation will have lost a lot more than evaluation.

Thank you all very much.

References

Blackstone, T., & Plowden, W. (1988). *Inside the think tank: Advising the Cabinet 1971-1983*. London: William Heinemann.

Ellis, J. J. (2002). *Founding brothers: The revolutionary generation*. New York: Random House.

Evans, H. (2006, June 18). Eye on the times. *New York Times Book Review*, p. 16.

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

Malignant reporting. (1987). *Public Interest*, No. 88, pp. 149-151.

Nordhaus, W. D. (2004, January 15). The story of a bubble. *New York Review of Books*, p. 34.

Padover, S. V. (Ed.). (1946). *Thomas Jefferson on democracy*. New York: New American Library of World Literature, Mentor Books.

Suleiman, J. (2006, July 4). Freedom of Information Act turns 40 today. *D.C. Examiner*, p. 1.

Thomas, L. (1982, May 30). *Things unflattened by science*. Commencement Address at Williams College.

U.S. Government Accountability Office, Program Evaluation and Methodology. (1984). *WIC evaluations provide some favorable but no conclusive evidence on the effects expected for the special supplemental program for Women, Infants, and Children* (USGAO/PEMD Report 84-4). Washington, DC: Author.

U.S. Government Accountability Office, Program Evaluation and Methodology. (1988a). *Enterprise Zones: Lessons from the Maryland experience* (USGAO/PEMD Report 89-2). Washington, DC: Author.

U.S. Government Accountability Office, Program Evaluation and Methodology. (1988b). *Weapons testing: Quality of DOD operational testing and reporting* (USGAO/PEMD Report 88-32 BR). Washington, DC: Author.