Remarks made at the
Environmental Evaluators' Network Forum
George Washington University
Washington, D.C.
June 8, 2010

NAVIGATING EVALUATIVE COMPLEXITY IN THE AGE OF OBAMA

Eleanor Chelimsky
Former Assistant Comptroller General, USGAO
Former President, American Evaluation Association

Good morning, everyone. It's such a delight to be here today to see this organization thriving, when I remember how hard it was, just a few years ago, to get environmental evaluation off the ground. My subject this morning is the problem of tackling complexity in evaluation, of trying to find some magic thread, like the one Ariadne gave Theseus, to get us through this Minotaur's maze of surrounding issues that it seems we have to confront. Now, these issues -- like the history of a cultural, social, economic or environmental problem, or the politics and policies of a particular period, or the battling theories of an intellectual climate, or the spillover of a subject area into bordering fields – these issues are complex, but they're hardly new. In fact, they've been with us since the first agricultural evaluations, but the truth is, we haven't paid a lot of attention to them. Perhaps we just didn't see their relevance to our work; perhaps we were a little bit mesmerized by those methodological tunnels we love to dig; or perhaps we simply hadn't grasped the power of these issues to affect our credibility.

So evaluation has always required consideration of the factors surrounding its subject matter. And because the methodological choices for an evaluation spring precisely from the careful analysis of these factors, evaluators need to recognize their relevance and integrate them into the blood and bones of the work in progress.

That's what i wanted to talk to you about today: the possibility of finding a reasonably practical, feasible way to navigate these complexities. We need some sort of checklist to make sure we keep track of the key external factors likely to affect an evaluation, and then use them and their interactions to shape both the evaluation question and the methodology that flows from it. It's a good time to do this, I think, first because there's an increasing awareness of the systemic way in which these factors interact to influence an evaluation, and also, perhaps, a greater willingness on our part to try to find ways to deal with that. This is true in a number of evaluation fields: for example, Atul Gawande's new book, "The Checklist Manifesto," in which

he devises a 19-item checklist that appears deceptively simple, but is structured so as to encompass the inter-related factors of a patient's health status as they apply to a medical intervention, with the purpose of maintaining proper sequence, avoiding error, and continuously integrating the medical needs of the patient and the dynamic requirements of surgery as it's carried out in hospital operating rooms. (Gawande, 2010)

Of course, Gawande's use of checklists is hardly the first in evaluation. Most of you here are familiar with Don Campbell's invaluable 1963 listing of nine threats to internal validity, which we've all used over the years to critique our methodologies. The checklist I want to tell you about today is my own: it emerges from a data base of about 300 evaluations, and 14 years of experience running the GAO's Program Evaluation and Methodology Division (we called ourselves PEMD). Almost from the beginning, I began looking for ways to steer through some of these complexities, once it became clear how important they were to our success or failure with the Congress. In PEMD, we found that the right locus for this work was always the very beginning of the evaluation: that is, the evaluation planning stage, when evaluators still have time to think freely and openly about their subject and are not yet bowed down by the demands of methodology. It was there that it turned out to be easiest to look at a great many things (going far beyond methods and data) that could then be integrated into the entire evaluation and give us our best chance at a strong and useful study.

We divided our thinking into two parts:

(1) The general subject of the proposed evaluation; and

(2) The specific evaluation question that had been posed to us, along with the kind of evaluation design that might be appropriate for answering it.

We applied this two-part strategy, starting in 1982, to all our major evaluations, based both on experience, and a concept derived from Max Weber, called Entzauberung. Weber thought that the most important thing an analyst could do for government was to de-mystify the prevailing myths to be found there. As a consequence, an evaluation should be judged by its successful provision of the best possible information in the public interest. Now, Entzauberung doesn't in any way discard the idea of use as a criterion for evaluative success, but it does say that an evaluation is valuable if it produces strong information on subjects that are important for the public to know, even when use seems unlikely: for example, when there's political unwillingness to look objectively at evaluation findings.

Implementing this concept in PEMD not only made us more concerned about the issues surrounding the evaluation's general subject area, but also pushed us to include all the perspectives relevant to the work, and to avoid any premature rigidity, either about the evaluation question itself, or about the likely best methods for answering it.

So here's the checklist. Bear in mind that, in PEMD, our evaluation designs always remained at least somewhat iterative, so as to keep up with any new issues that might surface during the course of the study, and also that the degree of complexity varied from evaluation to evaluation, depending on the subject matter, on the amount of general evaluative experience in the area, and on the number of key factors involved.

So, as i said, the first part of our work focused on:

The General Subject Area of the Evaluation

We typically undertook four kinds of reviews, depending on the subject. These were:

- A review of the history of the field;
- A review of the present-day political environment for the evaluation;
- An analysis of subject-area peripheries: that is, the places or borders where there was overlap with other subject areas; and
- A review of the lessons and experience of past evaluative work in the field.

Although I'm going to talk about these reviews separately, for clarity's sake, it goes without saying that, as elements in a framework of moving parts, they all interact with each other continuously. The first element, then, was:

The Historical Review

. . . . . .

Here we looked at the first six issues of the checklist: (a) the evolution of the subject being addressed by the evaluation; (b)the history of prior interventions for dealing with it;(c) the theories underlying those interventions, along with their controversies; (d) past and current scientific or technological applications in the area; (e) the development over time of the federal/state/local partnership for addressing the issue; and finally, (f) the status of current thinking about the subject.

Now, the utility of this historical review is that it familiarizes methodologists with the thinking and passions of the past, it provides background for understanding prior program experience, and for us in PEMD, it regularly turned up basic theoretical conflicts – economic, social, scientific, technological -- in almost every subject area, which then, of course, needed to be addressed in the evaluation design.

For example, the disagreements among criminal justice researchers about the purpose of prison – whether it's to deter criminals, punish them, warehouse them, or rehabilitate them – critically shaped data collection in our work on the results of probation and parole programs in state prisons. Again, our methodologies in studies looking at the causes for high rates of infant mortality, the unpopularity of food stamps among the elderly, and juvenile drug abuse, were all deeply influenced by the arguments among theorists, and between theorists and practitioners, about etiology and treatment. In some cases, the historical review forced us to change the evaluation question we were setting out to answer; in other cases, it compelled us to add new data collections to those already planned.

I think we're likely to see this same issue of conflicting theories reflected in the interventions policymakers adopt for dealing with environmental problems: for example, whether to use education to try to change people's attitudes and behavior vis-à-vis the environment, or whether " to accept people as they are," and focus instead on technological fixes, economic incentives, regulation, or other approaches that target people's circumstances rather than their mindsets. As Amitai Etzioni used to say, "Human beings are not so easy to change, after all." Etzioni, 1972)

On the other hand, after forty-plus years of saying you can't educate people to stop smoking no matter what the risks to their health, we've seen a significant decline in tobacco use. It took a lot more time than we originally thought, and we added a number of other policy fixes as well (like forcing people to go outdoors to smoke), so that the attribution question is still unanswered. We can't say that education alone changed smoking habits. But we know that something – or some set of things – did, and so, education, as a policy tool, is still on the table: alive and well. This means that when we look at education as a way to improve the environment via public

behavior, our evaluation design would have to feature an adequate timespan as well as a methodology capable of ruling out rival hypotheses for any changes we found.

Our second area of focus was:

A Review of the Current Political Environment for the Evaluation

Here we added five more issues to our checklist: (a) first, the known legislative, executive and judicial branch positions in the subject area; (b) the general political climate (especially the current degree of partisanship and ideological fervor); (c) the stances of the political parties regarding the specific subject being evaluated; (d) public opinion regarding both the subject area and the intervention proposed to address it, as well as current economic, social or cultural trends likely to affect public support; and finally, (e) views expressed by populations of particular interest to the evaluation: stakeholders, experts in the field, program or policy staff, beneficiaries, practitioners generally, and public interest groups.

For us, an important part of this review concerned a clear understanding of the degree of political passion we should expect to face throughout the evaluation. As legislative branch evaluators, we were an instrument of congressional oversight, just as executive branch evaluators serve an administration's obligations for public accountability. When passions are high, it not only becomes difficult for evaluators in either branch to protect their positions as legitimate actors in the government structure of checks and balances, it can become nearly impossible simply to get the job done. In PEMD, we encountered obstacles like the unexpected disappearance of funds for a program already under evaluation; the sudden classification of data that we needed to compare against an already-collected baseline; implementation delays that forced the sacrifice of a strong methodology in favor of a weaker, but more flexible

one; and the development of an atmosphere in which partisans from BOTH sides would try to discredit the evaluation findings. In this kind of political climate, evaluators need to pay careful attention to their credibility, to any appearance of advocacy, and especially to the inadvertent exclusion of relevant voices from the debate.

For example, in the work we did on chemical warfare, which was begun as an assessment of weapons effectiveness, we found, in looking at statements made by various defense experts, that the Department of Defense had presented to the Congress only those views favorable to the establishment of a U.S. chemical warfare capability. Based on this, we changed our evaluation question from one relating to effectiveness, to one relating to the current knowledge base. That is, before getting to the issue of effectiveness, we needed to determine the entire spectrum of expert views on the subject, along with what was known about the past experience of chemical warfare, what the broad areas of agreement and disagreement were, and what data existed to support the different positions. This moved us to a synthesis methodology, rather than the cause-and-effect design we'd been planning, plus an intense effort to document every step of our work, as testament to evaluation quality in the overheated political climate.

Unfortunately, evaluations and their designs are never perfect: they all have their strengths and weaknesses, their warts and their noble efforts to get at the truth. But evaluators can use their understanding of the current political climate, whatever it is, to help them explain their work systematically and persuasively. We need to do this because, as Fred Mosteller used to say, "it's always impressive to discover the sudden methodological expertise of our political critics in the wake of an unpopular finding." (Hoaglin, Light, McPeek, Mosteller and Stoto, 1982)

Of course, given the polarized politics of the year 2010, evaluators are not likely to escape unscathed. In this kind of climate, they should prepare to defend their methodologies, lower

their rhetoric to pianissimo, and be ready to adopt tail-end data collections, as we often did in PEMD, to enhance the political acceptability of a finding. But the truth of the matter is that politics nearly always trumps evaluation and there's not a lot we can do about it beyond sticking to our findings amidst the heat. As Robert Solow once told the Joint Economic Committee, "if a man comes into your office and tells you he's Napoleon Bonaparte, you can nod and smile, if you want to, but don't get drawn into a discussion of cavalry tactics at the Battle of Austerlitz."

Actually, evaluation does trump politics sometimes. In our work on chemical warfare, for example, we never expected any use whatever of our findings because they had been so hotly disputed, over the years, by the Department of Defense. But in fact, the contrary occurred: the House Committee on Foreign Affairs held numerous hearings on the studies, they were read carefully by Members of Congress, the program was zero-funded based on our findings, and they eventually served the State Department in negotiating the U.S.-Soviet Bilateral Chemical Weapons Agreement. So, Entzauberung, yes, but USE as well. (Fascell, 1990)

The third general-subject review we made was:

An Analysis of Peripheries, or the Spillover of the General Subject into Other Areas

Here we brought four new issues to the checklist: (a) first, explicit or implicit interactions between the subject area and other related systems or fields of knowledge, especially conflicts in policies across two areas; (b)second, whether those interactions were important to the proposed evaluation; (c) whether they were defined (or undefined) by bureaucratic boundaries; and (d) whether there were potential data-sets stemming from those interactions that might be

available for use. In addition, we also examined related areas of expertise, and, when relevant, the overlap of subject and function among levels of government.

For some evaluations, this analysis was quite limited because whatever overlap existed had only trivial importance for the proposed study. But for others, it was significant. In one study, this analysis showed us the possibility of merging data bases from both health care and law enforcement agencies to better understand the incidence of accidental shootings nationally. Again, in a study of high-school dropouts, the peripheral analysis pointed up the utility of combining education and police data to better clarify the size of the problem. In a third example, we found major disagreements in policy between medical and police approaches to drug abuse in teenagers that forced multiple revisions of our data collection plans.. Finally, for an evaluation of the Runaway and Homeless Youth Program, the analysis made clear that, to reach an objective judgment about the program, we'd have to integrate the views of a variety of different groups: the youths themselves, their families, the program's practitioners and managers, social agency caseworkers involved in foster care, welfare and homelessness, city and county police, as well as juvenile justice officials, in every program site included in the evaluation. This meant that an appropriate methodology needed first, to solve the problem of synthesizing opinions among groups of people with differing expertise, differing perspectives, and differing agendas; and second, to assure reliability and validity in the interview and survey instruments.

Now, for environmental evaluation, this area of spillover has always been important: many environmental issues are of peripheral but important concern to different agencies of the federal government. For example, if you had to answer the question, "is cap-and-trade the most appropriate policy tool to reduce carbon emissions?", then at least three agencies (EPA, Energy, and the Treasury Department) would need to be involved, and different kinds of

expertise – economic, scientific, regulatory and financial – would be required, along with the usual evaluative skills. And of course, when policies conflict across borders – environmental protection versus economic growth, for example – this again calls for that difficult synthesis of voices, especially including those of the least powerful.

Our fourth area of focus was:

A Review of Past Evaluations in the Subject Area

This, of course, is the traditional literature review. Here we added five issues to the checklist, asking, for each evaluation reviewed: (a) what was the evaluation question and what overall design was used? (b) what comparisons were made, what data were collected? (c) what program challenges had to be overcome? (d) what were the major strengths and weaknesses of the methodology, and what efforts were made to compensate for the weaknesses? And finally, (e) what findings were produced, what controversy was experienced, and what use, if any, was made of the findings.

Of course, the purpose of this review was not so much to judge the methodological quality of these earlier evaluations as to learn what the evaluators' experience had been, as a basis for planning our own evaluation. So we focused on the candor with which evaluators reported errors, weaknesses, or mistaken assumptions, and looked carefully at the general credibility of the work, in order to gauge how much confidence we could place in what they were reporting. In short, the effort here was not to use these earlier evaluations to determine acquired knowledge in an area, as in a synthesis or meta-analysis methodology or a Cochrane review, but simply to become familiar with the evaluation work that had been done in the past, and to understand how much we could lean on the reported experience.

In PEMD, we usually published this review of the evaluation literature as a chapter or appendix of our final report and it helped us in four ways: it guided us and our readers with respect to past experience, it expanded our thinking about possible methodological approaches, it showed us real-world mistakes to avoid, and it set historically-based, realistic expectations for the new evaluation.

With these reviews completed -- the historical, political, and peripheral analyses, along with the critique of the evaluation literature -- we could now assess:

The Specific Evaluation Question Posed

Here we added another four issues: (a) whether the question was bona fide; (b) for what purpose the answer was needed; (c) whether the question was sufficiently specific and objective for an evaluation to be performed that could satisfactorily answer it; and (d) whether obvious obstacles stood in the way of legislative or executive branch use. Our options, in PEMD, realistically speaking, were either to answer the question that had been asked, or change it just enough for it: (1) to make sense in terms of what our reviews had shown; (2) ensure appropriate methodology and feasibility of execution; and (3) still produce information that was valuable to the sponsor. This meant we also had to determine the sponsor's basic information need, and to understand when, where, and in what immediate political circumstance the question had arisen.

In PEMD, with sponsor accord, we often changed the question based on the four-part review we'd done. For example, in areas where few evaluations had been conducted, we tried to begin our work answering descriptive rather than cause-and-effect questions, but only if the

likely information yield would be compatible with the sponsor's needs. In other words, rather than move immediately to a question like "what effects did the program have?" We would begin by asking what the program WAS, in precise detail, how it had been implemented, how it differed in various sites, and whether it was well accepted by practitioners, beneficiaries and stakeholders. This could then be used as the first stage of a full-bore cause-and-effect evaluation. Again, if the historical review had shown past problems with relationships across levels of government, as was often the case in education, criminal justice or environmental programs, we would try to begin with a knowledge-base question, using a synthesis methodology to gather perspectives on the issue. Finally, when we answered cause-and-effect questions in a politically charged arena, we always set aside resources to devote to data collection, if and when unexpected issues arose.

Now this last strategy was the direct result of the trauma caused by the Income Maintenance Experiments of the 70s, which used elegant randomized controlled designs, but ended by failing to address a public policy issue of the greatest importance to the Congress: the effects of the program on family stability. The evaluators were looking at whether giving cash transfers to poor people would act as a work disincentive, but they'd been confronted by an unanticipated and significant rise in divorce rates among program beneficiaries, while the evaluation was under way. And since the evaluation design had not considered family stability as an issue, the evaluators were not collecting data on it, and found out about it "only incidentally," as Lee Cronbach reported. (Cronbach, 1982, , and Fienberg, Singer and Tanur, 1985).

Of course, unexpected issues can arise in any evaluation. My strategy was to unearth them, if possible, at the evaluation planning stage with the checklist-reviews, but if that didn't succeed, to keep resources in reserve to document them as they emerged later on.

With regard to evaluation questions in the environmental area, I would expect a heavy emphasis on descriptive and normative studies, because so much that evaluators need to know is as yet undescribed, unclassified, uncounted. Another issue is timing. In human services evaluation, for example, we worry about getting as much as three years to answer a question. But in environmental evaluation, it can take decades for the environmental effects of some substances to become known and perhaps decades again before we can evaluate the effects of efforts to mitigate them. Finally, lengthy data records are normally part of the evaluator's toolkit. But even lengthier series of observations will be required if we want to answer cause-and-effect questions using quasi-experimental methods like the interrupted time-series design.

In short, the framing of the evaluation question is absolutely crucial in determining what kind of methodology is most appropriate for answering it, but we can't frame the question properly unless we've made our way through a forest of complexities. We tried to do this in PEMD with our checklist of reviews and i believe it served us well. Although our checklist contained a few more items than Gawande's, it was an entirely feasible exercise, it got smaller as our expertise increased, and it provided us over time with a knowledge tool, a productivity-enhancer, and an effective early-warning system.

Conclusion

In conclusion, let me make three observations about this checklist and its place in evaluation.

First, it's an effort to grapple in an orderly way with the interactions of the complex systems that surround evaluation. It builds on pragmatism. It says, with William James, that evaluators can navigate a lot of complexity in history, politics and a host of other matters in order to anchor a new evaluation in a better understanding of reality. But it also says, with Edmund Burke, that if

we want to do useful things in government, we need to base our efforts on deep experience of the past. As Faulkner put it, "The past isn't dead; it isn't even past."

Second, the checklist serves as a guide. It helps us sort through the key factors in the non-methodological world that really matter for the evaluation design. If it turns out that changes in the issues surrounding the evaluation develop as the work proceeds, then the checklist- reviews document the original thinking, and pave the way for the design changes or technical fixes that need to be applied. We can do this so long as our design-thinking remains iterative, not rigid.

Third, we should NOT count on the checklist to improve the normal inelegance or messiness of the evaluation process: for example, the program's unexamined assumptions, our own muddy path from cause to effect, the unmeasurable differences in service delivery from site to site, our inability to hold things in place while political priorities, administrations, budgets and policy debates are all changing around us (Rivlin, 1974), and just the plain, honest, muddling-through that characterizes any search for truth in government. These are the day-to-day realities of our work, and checklists shouldn't be expected to change them.

On the other hand, these reviews do some remarkable things for us. They explain our evaluative perspective and our point of departure to our readers. They help us understand vulnerabilities in our planned design. They may even turn up examples of how earlier evaluators have compensated for them. And they bring us the evidence we need to re-negotiate the evaluation question with our sponsor, if that has to be done. All in all, they not only help to shape an evaluation, they also enhance its overall credibility. But in a more profound way, these checklist-reviews can also avoid or correct an evaluator's unconscious biases of memory; they strengthen the important features of appropriateness and iterativeness in an evaluation by the simple fact of their use in making design choices; and they counter our tendency toward

over-specialization in methodology by concentrating on interactions with other subject areas. As Charles Beard said, "the science of any subject is not at its center, but at its periphery, where it impinges upon other fields." (Beard, 1945)

From the particular perspective of the environmental evaluators in this audience, I think it's clear we've come a long way since the Age of Reagan in the 1980s, when the very existence of environmental problems was often denied. Now it's the Age of Obama, and there's a new and important opportunity to bring together the evaluative experience from a number of fields over the last thirty years, to provide strong and useful new information. That is, as always, we'll be answering the question, "What works,?" But we'll be answering it better, despite some difficult and entirely predictable methodological problems in moving from cause to effect. And again, as always, we'll be exploding a number of myths along the way, for which some policymakers may not thank us. But if the environment remains a high priority for this Administration, if we're principled, pragmatic and pluralistic in our approach, if our designs and methods are credible and we defend them with courage, at least some of what we say will be heard and some of our findings will make it into policy. It's going to be an exciting time for evaluators. Thank you all so much..

REFERENCES

BEARD, Charles, "The Economic Basis of Politics," Knopf, 1945, p.5.

CRONBACH, Lee, "Designing Evaluations of Educational and Social Programs," Jossey-Bass, 1982, p.360.

ETZIONI, Amitai, "Human Beings Are Not Very Easy to Change After All," Stanford Review, June 3, 1972, p.45

FASCELL, Dante, Chair, House Committee on Foreign Affairs, Letter to Eleanor Chelimsky, June 14, 1990.

FIENBERG, S., SINGER, B., and TANUR, J., "A Celebration of Statistics: Large-Scale Social Experimentation in the United States" (Chapter 12), Springer-Verlag, New York, 1985, p.295.

GAWANDE, Atul, "The Checklist Manifesto," Metropolitan Publishing, 2010

HOAGLIN, D., LIGHT, R., MCPEEK, B., MOSTELLER, F., and STOTO, M., "Data for Decisions," Abt Books, 1982, p.57.

RIVLIN, Alice, "How Can Experiments Be More Useful?" American Economic Review, 1974 (64), pp.346-354.