Chelimsky, E. (2012). Valuing, evaluation methods, and the politicization of the evaluation process. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New Directions for Evaluation, 133,* 77–83.

7

Valuing, Evaluation Methods, and the Politicization of the Evaluation Process

Eleanor Chelimsky

Abstract

The author argues that valuing is often about methodology, or questions about how to measure the value of a public program or policy, including how we measure the factual underpinnings of programs and how we synthesize information about issues relevant to public programs and policies. The organizational context is discussed as an important determinant of what methods are used, and that these decisions have become increasingly influenced by a single narrative—a narrative that sees increasing numbers of government programs and policies embodying a single idea, or positing a simple, one-on-one cause-and-effect relationship, both of which are established, not by evidence, but rather by suppressing existing evidence that is inconvenient to the particular idea or relationship being advanced. The implications of this single narrative on methodology are discussed and ways forward described. © Wiley Periodicals, Inc., and the American Evaluation Association.

As Raymond Carver might ask, what do we talk about when we talk about valuing? To me, it seems that valuing, like Carver's version of loving, is complex, diffuse, emotionally laden, and loaded with euphemisms, but that it boils down in the end to some pretty prosaic questions (Carver, 1988). For evaluators, one question might be, how do we measure the value of a public program or policy? Or else, how do we measure its factual underpinnings, the truth of assertions made to rationalize it?

NEW DIRECTIONS FOR EVALUATION, no. 133, Spring 2012 © Wiley Periodicals, Inc., and the American Evaluation Association. Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ev.20008

Or even, how do we synthesize all the knowledge available on a given issue? In short, what we are talking about when we talk about valuing is methodology, and we evaluators have spent a lot of time and effort trying to get it right.

Stronger methods allow us more confidence in the value judgments we make about a particular intervention, and weaker ones (which may be the only alternatives feasible in a particular design situation) force us to pepper those judgments with needed caveats. Our big problem is that we cannot always choose the strongest methodology per se. The type of evaluation question posed—seeking to assess, say, (a) the merit or worth of some initiative (which would make it an accountability question), or (b) the state of existing research underlying an initiative and its subject matter (which would make it a knowledge question), or (c) the current or changing size of some population or public problem (which would make it a descriptive question)—will largely determine which methods are appropriate for a particular evaluation, based also, however, on the historical, political, and evaluative context of the subject matter. This means, of course, not only that choosing methods for valuing is a complicated process, but also that those methods may be used for fact finding, or truth telling, as well as for establishing merit or worth.

Agency Involvement With Particular Methods

Some evaluative methods have benefited from various periods of vogue in different government agencies. For example, cost/benefit analysis was an important component of the Department of Defense's Planning, Programming and Budgeting System in the sixties; randomized controlled designs were favored by the Office of Economic Opportunity and other agencies for social program demonstrations in the sixties and seventies (as well as more recently, by the Department of Education); and the case-study method was the rock on which the Government Accountability Office built its evaluations over a period of decades.

Unfortunately, these methods, like all evaluation methods, suffer from warts and flaws that make them inappropriate for some evaluation questions (Chelimsky, 2007). For example, the randomized controlled design, which can be very strong for establishing the effectiveness of an intervention in one place (internal validity), does not normally bring evidence, without the use of additional methods, as to whether that intervention will be effective somewhere else (external validity). Cost/benefit studies are usually pretty good at predicting either costs or benefits, given an adequate basis in past evaluations, but there have been stubborn problems in getting agreement on definitions of costs and benefits. In addition, these studies have suffered from variability in their results when costs or benefits are hard to quantify, and when costs are short-term and benefits long-term, or vice versa (Chelimsky,

1977). As for the case study, which brings detailed information to illuminate a specific instance, it can be generalized only at some risk to the evaluator. Even with a careful plan to link a set of case studies together, it may be difficult to generalize because of the infinite variety of local circumstances.

Today, in the summer of 2011, it appears that some agencies still lean on treasured methods. This may be due to habit (the seductions of familiarity with one or another methodological tool), or to an unwillingness to take on the uncertainties involved in matching methods more appropriately to questions, or even, as some suggest, to a general decline in our societal ability to entertain complexity. Still, in some agencies at least, there now seems to be a growing understanding of the strengths and weaknesses implicit in the various evaluation methods for answering a given question, and greater willingness to try a mixed-methods approach that uses the strengths of one (or several) methods to shore up the weaknesses of others. Some of these agencies have benefited, perhaps, from the American Evaluation Association (AEA) publication, "Roadmap for a More Effective Government" (American Evaluation Association, 2010). This can be seen in recent communications on evaluation by the Office of Management and Budget; in the State Department's issuance, last May, of a new policy on evaluation; and the National Institutes of Health's current effort to adapt the Roadmap's evaluation approach to translational research. And there are similar endeavors in the planning stages in other agencies and in the Congress. So there does seem to be some progress in moving toward a more fully explicated use of methods for valuing an intervention.

The Advent of the Single Narrative

On the other hand, the interventions themselves appear to be changing in the opposite direction: that is, they are becoming less, rather than more sophisticated, and more, rather than less political. We are seeing increasing numbers of government programs and policies that embody a single idea, or posit a simple, one-on-one cause-and-effect relationship, both of which are established, not by dint of a preponderance of evidence, but rather by suppressing existing evidence that is inconvenient to the particular idea or relationship being advanced. This is the single narrative, which often results in programs or treatments that feature quite profound distortions of fact and of reality. And because we are often called to evaluate these initiatives, their genesis—and especially the suppression of evidence involved in developing their rationales—has important ramifications for all of our methods. This is to say that if the program or policy is based on invalid premises or assumptions, evaluators are likely to be faced either with a "weak-thrust, weak-effect" situation, or one in which, even with the best possible choice of methods, their evaluations will be answering the wrong questions.

It's true that in the past, evaluators have often complained about government interventions, in particular, the fact that policies and programs were not typically initiated "directly and clearly as an output of research" (Abt, 1977). And of course, they were not. Sometimes they were the product of administrative judgment and experience rather than of research or policy analysis of a public problem. Sometimes they emerged in all their splendid vagueness from the hard-won political consensus required to pass a bill in the Congress (the Bureau of Indian Affairs, for example, once had to develop a legislatively mandated program whose goal was to "improve Indian civilization") (Hemmes, 1977). Sometimes the eccentric idea of an administration official achieved agency implementation with little justification in theory or evidence. However, in a democratic society, it is, in fact, elected officials and agency policymakers, not evaluators, who are charged with developing governmental initiatives. Evaluators have dealt with this situation by pointing out intervention inadequacies in their final reports and by maneuvering to change the evaluation questions, evolve specificity and precision out of vagueness, and ground their findings and recommendations in reality. The main point here, however, is that, in the past, the conceptual weaknesses of programs and policies usually—although not always—came from lack of sophistication, not political purpose.

The Single Narrative and Its Implications for Methodology

Today, however, we are seeing a remarkable expansion of the politically inspired single narrative across a variety of subject areas, with interventions deliberately structured to ask only those questions that will lead to the right, or desired answer. This expansion is occurring not only in the public sector but in the federally regulated private sector as well. Also, the single narrative, which used to operate mostly at the pre-evaluation stage of a program or policy (by constricting the body of knowledge already existing about a public problem, by presenting only one set of facts and views, and by developing an intervention that reifies that truncated version of reality), now increasingly features an additional component that takes place during the actual implementation of the evaluation. As Ernest House has reported, "drug companies have found ways of producing the results they want, including manipulation of treatment (dosages), selection of sample, control of data, and calculated publication" (that is, the selective reporting of favorable findings and the suppression of negative results) (House, 2008). Two points are important here. First, the interferences in the evaluation process reported by House could bring improper compromises and conflicts of interest for the evaluator. And second, these cases show the power of the two political components: the single narrative and the introduction of bias into the evaluation treatment can be effectively deployed at the very heart of evaluation's strongest methodology: the randomized, controlled experimental design.

Current Expansion of Efforts to Direct and Control Evaluations

One can speculate on the reasons for the current expansion. First, we are clearly at the ideological end of the regular democratic political cycle that moves from incrementalism to ideology and back (Chelimsky, 1996), and many single-narrative programs have been hatched during these periods of ideological fervor (e.g., the Reagan Administration's "Just Say No" program, which reduced adolescent drug problems in America to a series of stereotypes) (Chelimsky, 1996). Second, the combined single-narrative and implementation-control effort no longer needs to rely on improvisation, but benefits today from an effective model for manipulating evidence that was largely unknown in the past. This is the process developed by the Department of Defense in the 1980s and 1990s for presenting only one side of a case (in a budget request, say, for funding a new program or technology) and for using classification and selective reporting to suppress any data that might cast doubt on that presentation (U.S. Government Accountability Office/Program Evaluation and Methodology Division, 1988).

Today, the single narrative is alive and well at the Department of Defense (for example, in the continuing saga favoring piloted vehicles over drones, big technology over small, and bombers over nuclear submarines, despite continuing experience and evidence to the contrary) (Beschloss, 2011). However, we also find the single narrative spreading to health care (not only in the shape of drug companies that refuse to release unfavorable data, but also in the formulation of guidelines for clinical trials and conflicts of interest within "the expert committees reviewing clinical trials and selecting what constitutes 'best evidence'") (Groopman, 2009). We see the narrative again in education (with respect both to unfounded assumptions about charter schools, and the notion that bad teachers are the single cause of poor student outcomes) (Ravitch, 2010). We find it in energy (with the Nuclear Regulatory Commission's assurances of system safety that are presented without discussion of the Commission's loosening or weakening of the very safety standards through which indications of lessening safety would emerge) (Donn, 2011); in the environment (with the continuing insistence, often heard in the Congress, that climate change is a hoax) (Warner, 2011); and in regulation (e.g., the dismantling of the Glass-Steagall Act, based on the assertions of special banking interests and the suppression of evidence about its long-term public benefits) (Nocera, 2011).

In my judgment, these two political components—the single narrative (with its ability to direct and control the evaluation process) and the manipulation of treatment procedures (which nullifies strong methodologies and exposes evaluators to potential conflicts of interest)—present major problems for the evaluative enterprise. At a time when ideological warriors can be expected, as in the past, to cut both public programs and the evaluation systems that inform on their results, in order to reduce the size of government

(Chelimsky, 1985), it may be that the best we can do in the face of the hurricane is to wait for a change in political cycle dynamics toward a more prudent, Jeffersonian incrementalism. But there are, perhaps, a few small things we might do now that could help avoid similar distortions of both the evaluation process and government accountability in the future.

Four Suggestions

1. Given the attention that the AEA's "Roadmap" has been receiving both in federal agencies and in the Congress, it may be useful to amend that document to provide measures for dealing with the evaluation and accountability problems that occur when evaluations are either politically skewed at the start, or improperly manipulated during their implementation.
2. Evaluators should be urged to use their final report, as they have done in the past, to cast light on shadowy areas of program or policy construction and rationale. This can be a successful tactic when the suppression of evidence is egregious, or when there is a powerful constituency that is willing to help in the debate.
3. Standing review panels could be appointed in neutral places like universities to examine cases of interference with evaluative independence during the course of a study, in order to assess damage to evaluation quality in the particular study, and to deter further political efforts in this direction.
4. Finally, the AEA might consider opening an inquiry on this double problem under the aegis of its Ethics Committee. The effort could seek to give guidance to both policy makers and evaluators by examining the role evaluators should play in dealing with political manipulations that threaten the evaluative capability to produce new knowledge about government programs, or to measure their merit and worth.

We could, of course, simply wait and hope for things to improve, but single narratives are pernicious because of the falsehoods and the constricted ways of thinking they embody, and because they have far-reaching effects that are not limited to evaluation. As Hannah Arendt once put it, "Freedom of opinion is a farce unless factual information is guaranteed and the facts themselves are not in dispute" (Arendt, 1967). Indeed, the problem with rigging the evaluation process through the single narrative is precisely that: We do not get all the facts, and the facts we are given may not be facts at all. This affects us in at least three ways: in our ability to think clearly about public problems, in our ability to perform meaningful evaluations, and in our ability to use those evaluations to improve accountability and transparency in government.

References

Abt, C. C. (1977). *Perspectives on the costs and benefits of applied social research* (p. 4). Cambridge, MA: Abt Books.

American Evaluation Association. (2010). *An evaluation roadmap for a more effective government*. Retrieved from http://www.eval.org/EPTF/aea10.roadmap.101910.pdf

Arendt, H. (1967, February 24). Reflections on truth and politics. *The New Yorker,* p. 52.

Beschloss, M. (2011, April 24). Missing the target. *New York Times Book Review*, p. 14.

Carver, R. (1988). *Where I'm calling from*. New York, NY: Vintage Books.

Chelimsky, E. (1977). Perspectives on the costs and benefits of applied social research. In C. C. Abt (Ed.), *Perspectives on the costs and benefits of applied social researc*h (pp. 179–195). Cambridge, MA: Abt Books.

Chelimsky, E. (1985). Budget cuts, data and evaluation. *Society, 22*(3), 65–73.

Chelimsky, E. (1996). *Distinguished public policy lectures*. Evanston, IL: Center for Urban Affairs and Policy Research, Northwestern University.

Chelimsky, E. (2007). Factors influencing the choice of methods in federal evaluation practice. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology*. *New Directions for Evaluation, 113*, 13–34.

Donn, J. (2011, June 20). NRC loosens, ignores safety rules to keep nuclear reactors operating. *Plain Dealer,* p. 2.

Groopman, J. (2009, November 5). Diagnosis: What doctors are missing. *The New York Review of Books,* p. 26.

Hemmes, R. A. (1977). Evaluation at the Bureau of Indian Affairs. In E. Chelimsky (Ed.), *Proceedings of a symposium on the use of evaluation by federal agencies* (Vol. I, pp. 26–28). Bedford, MA: MITRE.

House, E. R. (2008). Blowback: Consequences of evaluation for evaluation. *The American Journal of Evaluation, 29*(4), 416–418.

Nocera, J. (2011, June 18). The banking miracle. *New York Times,* p. 14.

Ravitch, D. (2010, November 11). The myth of charter schools. *The New York Review of Books,* pp. 22–24.

U.S. Government Accountability Office/Program Evaluation and Methodology Division. (1988). *Weapons testing: Quality of DOD operational testing and reporting* (No. 88–32BR). Washington, DC.

Warner, J. (2011, February 27). Fact-free science. *New York Times Magazine,* p. 11.

ELEANOR CHELIMSKY is an independent consultant for evaluation policy, practice, and methodology and was for 14 years the director of program evaluation and methodology at the U.S. Government Accountability Office. She previously evaluated programs and policies at NATO, and then at the MITRE Corporation.