Evidence, Argument, and Persuasion
in the Policy Process

The main thesis of this chapter is that an adequate theory of policy development requires that attention be paid to ideas, theories, and arguments as well as to technology, economics, and politics. Analysis and arguments conceptualize, and thus transform, the institutions and processes of policy-making. At the same time, institutional and political factors influence the rate and quality of conceptual innovation and determine which among the available proposals will be selected for actual use. Thus, the relationship between policy and its intellectual superstructure, or meta-policy, is a dialectic one.

For the student of policy development it is as much of a problem to explain the continuity of the policy framework as it is to explain how particular programs and institutional arrangements change within that framework. The distinction between core and periphery has been introduced with this particular problem in mind.

An adequate theory of policy development must also take into consideration that an increasingly significant part of the growth of public policy is the consequence of previous policies and their interconnections and overlaps, rather than the result of deliberate choice. The emergence of unplanned policies is too important a phenomenon to be dismissed as the random resultant of political, economic, or bureaucratic forces pulling in different directions. In many cases, unplanned change is best explained in terms of endogenous processes taking place in a relatively autonomous policy space.

Finally, the link between conceptual and policy development is provided by an evolutionary model according to which changes in current policy may be analyzed as the outcome of a dual process of conceptual variation and subsequent selection from the pool of available policy variants. The policy community is the locus of conceptual innovation, while the political arena is the locus of selection. The dialectic relationship between policy and meta-policy finds concrete expression in the interactions between a policy community and the corresponding political arena.

EIGHT

Evaluation and Accountability

The debate through which criteria of evaluation and standards of accountability are established is an essential part of the process of policy development. In fact, the dialogue among legislators, policymakers, and the electorate whereby policy is formed in a system of government by discussion is always, directly or indirectly, about evaluative issues. Analysts have contributed to this debate in a number of ways, but especially through the new subdiscipline of evaluation research.

Evaluation research is a large and expanding area of policy analysis devoted to collecting, testing, and interpreting information about the implementation and effectiveness of existing policies and public programs. As I noted in a previous chapter, the importance achieved by evaluation research within the field of professional policy analysis shows that analysts have finally come to realize that the effective delivery of public services requires more than the design of some theoretically optimal program. Even more important is to learn how the program is actually implemented, who benefits and who loses from it, whether the program is accomplishing what was intended, and if not, how it can be improved or discontinued.

Many evaluators seem to assume that these are purely empirical determinations, involving neither value choices nor personal opinions. In fact, values and opinions count a great deal in evaluation not only because of the ambiguity of the outcomes of practice—the difficulty of assigning specific causes to particular effects, of measuring outputs, and assessing unintended consequences, of distinguishing between flawed conceptions and failures of implementation—but even more because of inescapable disagreements about the kind of evaluative criteria that are meaningful, fair, or politically acceptable in a given situation. Such ambiguities and disagreements can never be fully resolved by improved measurement and testing techniques, but can be represented and clarified by debate and mutual persuasion. In order to understand the role of argument in evaluation it is essential to distinguish between standard setting and standard using.

In chapter 2 I drew a parallel distinction between norm setting and norm using—between defining the criteria of what constitutes a policy problem and searching for solutions that satisfy those criteria. The distinction between setting and using standards is equally important in evaluation. At the standard-setting stage, or when a reform of evaluative standards is being considered, it is open to anyone to put forward a proposal as to what the standards should be and to use persuasion in order to influence others to accept the proposal. Objective analysis—for instance, to clarify the consequences of adopting a particular set of criteria, or to point out logical inconsistencies among different standards—can provide useful inputs to the debate, but cannot play a decisive role at this stage. Empirical determinations that show whether certain standards are in fact attained become relevant only after agreement about the appropriate evaluative standards is reached. Because reaching agreement on standards is so difficult in the presence of different values and interests, much evaluation analysis is really concerned with the merits of various standard-setting proposals rather than with the application of particular standards of merit to a given program.

MULTIPLE EVALUATION

Professional evaluation is only a small part of the general process of criticism and appraisal of public policies to which all politically active members of a democratic community contribute in different but equally useful ways. Arguments about standard-setting proposals play an even bigger role in policy criticism than in professional program evaluation. Policies and policy instruments are constantly assessed, ex ante and ex post, from the diverse critical perspectives of legislators, judges, policymakers, program managers, implementing bureaucrats, interest groups, independent experts, the media, and private citizens.

These perspectives are different both because evaluative criteria vary with the role and position of the evaluator and because different evaluators tend to focus their attention on different aspects of the policy-making process. General standards of performance like legality, legitimacy, economy, effectiveness, efficiency, or responsiveness to public needs are characteristically related to the distinct roles of judges, politicians, budget officers, public accountants, and consumers of public services or their political representatives. Moreover, some criteria, such as efficiency and effectiveness, apply primarily to the outputs or outcomes of public policy, other criteria (for instance, economy) apply to the inputs, and others still (legality, legitimacy) to the process that transforms inputs into outputs.

This multiplicity of evaluative standards and critical perspectives reflects the complexity of policy-making in a pluralistic society. Experience shows that debate among advocates of different criteria is often useful in reaching agreement and permits a more sophisticated understanding of public policy than is possible from a single perspective. Even professional evaluators now recognize that their work becomes relevant only in the broader context of competing criteria and evidence presented by various actors and interest groups. The new slogan is "multiple evaluation." This phrase acknowledges the legitimacy of different criteria and perspectives, but also suggests the need to reach a level of understanding that is more than the sum of the separate evaluations. As I said earlier, the purpose of multiple

evaluation is not to combine all the partial criteria into one general criterion of good policy, but to contribute to a shared understanding of the various critical perspectives and of their different functions in the process of public deliberation.

Multiple evaluation starts with two basic questions: "Evaluation by whom?" and "Evaluation of what?" The first question emphasizes the importance of accounting for the presence of different evaluative *roles*, while the second question directs our attention to the three basic *modes* of evaluation—inputs evaluation, outcomes evaluation, and process evaluation.

EVALUATIVE ROLES

Policy or program evaluation serves a wide variety of uses and users. That different criteria are used by people in different roles is not a bad thing as such. It simply reflects the various needs, interests, and concerns of different actors and stakeholders. So long as the judgments expressed from the perspective of one particular role are not presented or misinterpreted as judgments relevant to or speaking for all possible roles, we have a healthy state of multiple or pluralistic evaluation.

Difficulties begin to arise when this neat partitioning of roles and the criticism voiced from them is not possible. Unfortunately, such breakdowns seem to be more the rule than the exception in practice. Perhaps the most common problem occurs when the conclusions of an evaluation done for use in a particular role are assumed to be equally relevant from the perspective of other roles with different evaluative criteria. Because roles and criteria are mismatched, the conclusions of the evaluation are almost inevitably found wanting.

Such difficulties arise, for example, when academic ecologists are convened as the sole reviewers of the environmental impact assessment mandated in the United States for federal projects. From the perspective of a project manager or administrative law judge, the relevant evaluative criteria might be the timeliness of the assessment, the likelihood that potentially serious impacts have been noted, and, perhaps, whether practical development

alternatives are suggested. From the ecologists' point of view, however, such criteria are at best vaguely comprehended and given only secondary consideration. Their evaluation would likely be based on such criteria as the adequacy of sampling design, the use of appropriate theory, and the accurate characterization and quantification of uncertainties. As a result, the ecologists may accept or reject attempted impact assessments for reasons that are largely irrelevant to the people who will eventually have to resolve the practical problems of environmental management.[1]

Examples of mismatched criteria abound in professional evaluations of public programs. Many program evaluations have a narrow managerial focus, being concerned with goal achievement and administrative control rather than with the responsiveness of the program to the divergent values of different individuals and groups. Such a narrow perspective neglects a more structural analysis of changes in societal values and of the ability of bureaucracy to adapt to such changes. This may be the most important information from the point of view of top policymakers.

In turn, bureaucrats often feel that the stress placed by many evaluation studies on effectiveness and efficiency is in conflict with such basic values as employee participation, personal development, and high morale. Others question how economic rationality should be balanced against professional standards. Again, evaluation done to learn about a program's operations and its effects—information that is important for allocating resources and drafting new guidelines—may be unsatisfactory as a means of controlling the implementing agency.

Is there a cure for such common tendencies to confound evaluative roles or to mismatch criteria? Probably not. At a minimum, however, efforts to build a critical capacity for judging particular programs or entire policies should explicitly recognize that multiple roles exist, each with a legitimate claim to set eval-

1. William C. Clark and Giandomenico Majone, "The Critical Appraisal of Scientific Inquiries with Policy Implications," *Science, Technology, and Human Values* 10, no. 52 (Summer 1985):8.

uative criteria. Further, such efforts should appreciate the complex pulls and pushes that the resulting diversity of evaluative modes exerts on the evaluation itself.

EVALUATIVE MODES

Roles are not the only factors that need to be distinguished in making sense of the criteria by which programs and policies can be critically appraised. Analysts have also found it useful to distinguish three general modes of evaluation. In the *outcome* mode, evaluation focuses on the outputs or outcomes of a particular activity. In the *input* mode, the emphasis is on the resources, skills, and people engaged in the activity. Finally, in the *process* mode, attention shifts to the methods used to transform political, economic, and other inputs into outputs/outcomes. Procedural rules that govern participation in and administration of the program are also relevant in this context. We next consider each of these modes separately, all the while keeping in mind that they are usually mingled in practical efforts to evaluate programs and policies.

Evaluation by outcomes or results is commonly viewed as the obvious way to assess the value of any purposive activity. Goals or benchmarks are defined, results are produced, and the two are compared. In the case of an educational program, for example, one would appraise the difference between pre- and post-tests, or between the experimental and the control group, on a number of different criteria. In health programs, the outcomes are changes in incidence and prevalence rates; in manpower programs, the outcomes are employment rates, and so on.

This mode of evaluation has a strong intuitive appeal. Indeed, one may well wonder why any other form of evaluation is at all needed. What common sense overlooks is that outcomes evaluation can be successfully performed only under rather stringent conditions. One obvious condition is that it must be possible to measure with reasonable precision the level and quality of the desired output or performance. If the indicator of performance is expressed by the distance between goals and outcomes, goals have to be clearly defined, outcomes must be unambiguously measurable, and the measuring instrument should be reliable.

These conditions are not often satisfied in practice, even approximately. When they are not, other modes of evaluation must be used. Evaluation by inputs focuses on the quantity and quality of the resources available to perform a certain task: number and technical quality of the staff, available information, level of funding, political support, and so on. These are indirect indicators of performance, at best. Unless a definite relationship between inputs and outputs or outcomes—a well-defined production function in the language of the economist—can be assumed, input variables are a poor proxy for what we are really interested in knowing, namely, how effective is a given program, or how good is a particular policy?

But in some situations, input variables are all the information the evaluator has to work with—for example, when the problem is to estimate the likely results of a new project or to assess the feasibility of a new program. Moreover, for purposes of control, input variables are often strategically more important than outputs. The detailed rules of public accounting that severely restrict the freedom of public managers to substitute one input for another in response to changing circumstances and new opportunities are a historically important example of evaluation and control by inputs.

Measurability is not the only, or the most serious, problem of outcomes evaluation. The main problem is that in many situations this type of evaluation gives policymakers, program managers, and interested citizens very little information upon which to act. Simply knowing that outcomes are satisfactory or unsatisfactory does not tell decision makers and critics very much about what to do. Where outcomes are evaluated without some reasonably accurate and coherent definition of the program, and without knowledge of the manner in which it is implemented, the results seldom provide a direction for action because the

decision maker lacks information about what produced the observed outcomes. Pure outcomes evaluation is the "black box" approach to evaluation.[2]

Perhaps the best-known example of this approach are the standardized achievement tests routinely administered in American public schools. Regarded by many evaluators as the epitome of all that is most objective and "scientific" in educational evaluation, standardized tests are of little use to teachers and parents and do not tell school officials what to do to improve the educational experience of students. In order to improve schools, officials need information about what actually happens in the classroom—course content, grading procedures, teaching methods, teacher-student interactions—and standardized tests do not provide such information.[3]

Course content, grading procedures, teaching methods, and teacher-student interactions are examples of process variables. In many respects, process evaluation is the most subtle and informative mode of evaluation—it provides information that input and output measures are almost sure to miss. Even in commercial activities where outcomes can be easily quantified, prudent managers try to avoid too narrow a focus on results. They do so in the knowledge that the best outcome measures never capture more than a small fraction of the total range of performance that is important to the organization. For example, sales volume is an unambiguous and robust output measure, but it tends to focus too narrowly the attention of salespeople on maximizing sales in the short run, with the result that they ignore other functions that have a large effect on future sales. People on straight commission have no incentive to arrange stocks, take inventory, or train new salespeople who become their competitors, and their supervisor cannot affect their salary by taking into account the other, unmeasured goals.

Hence, as William G. Ouchi discovered in a study of seventy-eight retail department store companies, commission payment

2. Michael Quinn Patton, *Utilization-Focused Evaluation* (Beverly Hills, Calif.: Sage, 1978), 155–58.

3. Ibid., 155–57.

tends to be found in sales areas (like cosmetics, major appliances, and furniture) that require a relatively high degree of expertise and where therefore salespeople develop something like professional norms, or where knowledgeable and active clients can replace professional norms as another source of evaluation and control. Some companies even prohibit the maintenance of sales volume records for individual salespeople.[4]

EVALUATION AND CONTROL

Among the many possible purposes that evaluation may serve, directly or indirectly, control is certainly one of the most important. In fact, the relationship between evaluation and control is extremely close. On the one hand, in order to control any activity or organization, it is necessary to monitor and assess its performance with reference to a set of standards. On the other hand, evaluation modes and criteria have operational consequences for organizational and individual performance, since if people know that certain dimensions of performance are highly rated by the evaluators, they will tend to change their behavior accordingly.

Peter Blau, in his classic study of a public employment agency, provides instructive examples of the close relationship between performance and evaluative criteria. Thus, when the number of interviews completed by a subordinate was the only evidence the supervisor had for evaluating him, "the interviewer's interest in a good rating demanded that he maximize the number of interviews and therefore prohibited spending much time on locating jobs for clients. This rudimentary statistical record interfered with the agency's objective of finding jobs for clients in a period of job scarcity."[5] Even the more comprehensive system of monitoring introduced later produced serious displacements of organization goals:

4. "The Relationship between Organizational Structure and Organizational Control," *Administrative Science Quarterly* 22 (March 1977): 95–113.

5. *The Dynamics of Bureaucracy* (Chicago: University of Chicago Press, 1955, 38.

> An instrument intended to further the achievement of organizational objectives, statistical records constrained interviewers to think of maximizing the indices as their major goal, sometimes at the expense of these very objectives. They avoided operations that would take up time without helping them to improve their record, such as interviewing clients for whom application forms had to be made out, and wasted their own and the public's time on activities intended only to raise the figures on their record. Their concentration on this goal, since it was important for their ratings, made them unresponsive to requests from clients that would interfere with its attainment. Preoccupation with productivity also affected the interpersonal relations among interviewers, and this constituted the most serious dysfunction of statistical reports.[6]

Similarly, there is evidence that compensating teachers on the basis of their output, as measured by student test score gains, creates incentives for teachers to concentrate their time on students in the middle of the test score distribution, neglecting those at the top who would advance well on their own and those at the botton whose test scores would not respond to small additional amounts of teacher time. Also, where the compensation of teachers depends on the number of students who acquire a set of narrowly defined skills (as, for example, under the payment-by-results plans used in England in the middle of the nineteenth century to compensate elementary school teachers), there is a tendency to narrow the curriculum to exclude all non-tested subjects—including many that are generally perceived to be important but are difficult to test.[7]

As these examples suggest, a careful analysis of the relevant production activity is essential for the choice of an appropriate method of evaluation and control, since knowledge of the activity provides the best clues to the responses that a particular mode of evaluation will elicit. More precisely, two parameters are crucial for determining the conditions under which different modes

6. Ibid., 46.

7. Richard J. Murnane and David K. Cohen, "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive," *Harvard Educational Review* 56, no. 1 (Feb. 1986): 1–17.

of evaluation/control are appropriate: measurability of the outcomes and knowledge of the process that generates the outcomes.[8] If, for the sake of simplicity, we dichotomize these parameters, we obtain the accompanying table:

		KNOWLEDGE OF PROCESS: Complete	KNOWLEDGE OF PROCESS: Incomplete
MEASURABILITY OF OUTCOMES	High	Evaluation by process or by outcome	Evaluation by outcome
	Low	Evaluation by process	Evaluation by input

The four cells represent the situations in which the different modes of evaluation may be applied. The best situation is obviously when measurability of outcomes is high and the transformation process is completely known. Here the choice between process or outcomes evaluation depends on cost and economic convenience. Production of standardized goods using a well-understood technology is a familiar example. At the other extreme we find an activity like teaching whose outcome—education—is difficult to measure and where process is so idiosyncratic that it is impossible to characterize effective teaching as the consistent use of standardized techniques. Hence, teachers are evaluated mostly by input criteria like educational credentials and years of teaching experience, and this method of evaluation is reflected in uniform salary scales.

Activities that fall between these two extreme cases are usually evaluated by a mixture of the three basic methods. In for-profit firms, for example, managers are rewarded, at least in part, by results—sales and profits—but internal staff activities like research and development, accounting, or personnel management are evaluated by a combination of input and process criteria. On the other hand, traditional methods of evaluation and control of public programs rely heavily on inputs, but some analysts

8. Ouchi, "Organizational Structure and Organizational Control," 97–99.

think that recent advances in the measurement of outcomes in areas like health care and personal social services may eventually induce a shift in policy concern from inputs to outcomes. If true, this development would have important consequences for public accountability.

ACCOUNTABILITY

Our discussion of different roles and modes of evaluation, and of the relationship between evaluation and control, bears directly on the problem of public accountability. In a democracy managers and producers of public services are expected to be accountable for their performance to those who consume their services and to those who pay for them, or to their political representatives. Accountability cannot be enforced without adequate standards of performance, but these, as we have seen, are often difficult to define. Traditional input criteria like budgetary rules, administrative guidelines, or staffing ratios may be sufficient for purposes of oversight, but cannot satisfy the growing demand to state accountability in terms of outputs and to reward those who produce more efficiently.

The advocates of accountability by results argue that public managers should be free to choose the way they allocate their resources as long as they achieve specific targets. This procedure would free them from cumbersome and ineffective input controls and provide powerful incentives to improve results, just as private managers are free to vary production methods but are rewarded according to sales and profits.

Like outcomes evaluation, the idea of accountability by results has a strong intuitive appeal. It sounds simple and reasonable; implementing it, however, is hard. The congenital reluctance of liberal democracies to grant too much discretion to public administrators is only part of the problem. Another and more immediate reason for the difficulty of implementing the idea is that most public programs have vague and diverse goals, and as a consequence agreement about relevant criteria of success is hard to achieve.

The exact nature of the difficulty is not always clearly understood. For example, some analysts claim that it is possible today to develop the objective measures of performance that are needed to implement the concept of accountability by results. A. J. Culyer, a British economist, has recently written that "in health care, education and the personal social services it is now feasible to measure and monitor outcome. For example, we can measure health. We can measure dependency of the young and the old. We can measure handicap. We can measure deprivation. We can measure attainment of educational and social skills. We can associate changes in these measures with the forms of care we provide."[9] The availability of such measures, Culyer argues, represents a major, indeed a revolutionary change in the way we manage and finance social services and the way we expect the professionals who provide them to be accountable to society as a whole.

But what does measurability mean in the present context? With sufficient ingenuity it is always possible to devise scalar measures of particular dimensions of performance. The goals of most public programs, however, are multidimensional. Schools, for example, are supposed to improve cognitive skills, but also to socialize young children and teach them democratic values. Hence, multiple measures are needed to reflect multiple objectives and to avoid distorting performance.

Even if one assumes that the level of achievement of each separate objective could be measured unambiguously and objectively, one would still have to solve the problem of aggregation, that is, of reaching a consensus about the weights to be assigned to the various objectives. Until a consensus is reached, there is no overall measure of performance. Disagreement about the weights is to be expected, however, since the issue of weights is at its core a debate about how the activity should be organized and whose interests matter the most.

Moreover, we know that performance in the public sector can seldom be expressed by means of output measures alone. In

9. *The Withering of the Welfare State? Whither the Welfare State?* (Vancouver: Department of Economics, University of British Columbia, 1986), 25.

general, a combination of input, process, and output criteria is needed, and different groups will also disagree about the weights to be assigned to the different elements of the combination. In education, for example, teachers' unions favor input criteria like years of teaching experience, size of classes, and hours of work, while parents are more concerned with various dimensions of outcome and process.

Because the issue of weights is so divisive, most policy debates about health, education, or social services avoid the problem. Instead of openly debating what the weights should be, the tendency is to delegate decisions about resource allocation to the service producers. But such delegation is not consistent with the idea of accountability by results, for the different decisions about weights of individual producers mean that they are each trying to produce a somewhat different mix of outputs.[10] The need, then, is less to develop "objective" measures of outcomes (though improved measurement would certainly help) than to begin and sustain a wide-ranging dialogue about the meaning and implication of different sets of weights among producers and users of public services.

The difficulties of evaluation by results loom even larger in debates about the accountability of government to the legislature and the electorate. In the past it may have been possible to agree on a few stable standards—maintaining law and order and a stable currency at home, peace abroad—for evaluating the activities of government. But with the great expansion of these activities and the lengthening of the time span on which informed judgments of performance should be based, the record of any government is much less clear and the evaluative criteria are correspondingly more controversial.[11]

Under such conditions, the confidence that accountability is supposed to generate can hardly depend on the relation between some measurable outcome and a predetermined standard of success. Rather, it depends on methods of evaluation capable of

10. Murnane and Cohen, "Merit Pay and the Evaluation Problem," 5.

11. Geoffrey Vickers, *The Art of Judgment* (London: Chapman and Hall, 1965), 149.

providing more information than a simple judgment of success or failure. Exclusive reliance on measures of short-term results often leads to the conclusion that most public policies are ineffective. The apparently ubiquitous phenomenon of "little effect" concerns professional evaluators. As Carol Weiss writes, one of the major obstacles to putting evaluation results to use is precisely their dismaying tendency to show that the program has had little effect.[12] Organizations do not fare better: "Measured against the Olympic heights of the goal, most organizations score the same—very low effectiveness. The differences among organizations are of little significance."[13] Programs and organizations, like scientific theories, seem to be born to be refuted, and evaluation, as usually conceived and practiced, can play no crucial role in their development.

The phenomenon of "little effect" becomes less surprising once we recognize that evaluation exclusively in terms of short-term results is bound to be inconclusive under normal circumstances. First, to get on with their work, evaluators must assume that their models and measuring techniques are unproblematic, or at least less problematic than the working hypotheses incorporated in the program they evaluate. In fact, a conclusion of little or no effect could be interpreted as a failure of the evaluator's model just as well as a failure of the program.

Second, any reasonably accurate and coherent description of a particular program must take into consideration the policy framework in which the program is embedded. In particular, its position with respect to the policy core is likely to influence in a significant way how the program is implemented, the level of support it enjoys, and the meaning it has for different policy actors. Even the most sophisticated measures of outputs/outcomes are almost sure to miss these aspects of the program and thus may underestimate its effects.

Finally, in evaluating efforts to significantly change the be-

12. "Evaluation Research in the Political Context," in E. L. Struening and M. Guttentag, eds., *Handbook of Evaluation Research*, vol. 2 (London: Sage, 1975), 13–25.

13. Amitai Etzioni, "Two Approaches to Organizational Analysis: A Critique and a Suggestion," *Administrative Science Quarterly* 5, no. 2 (1960): 258.

havior of large numbers of people, a limited time frame is inappropriate because it neglects both the severity of the initial administrative problems and the possibility of learning by doing. For example, in the case of compensatory education under Title I of the U.S. Elementary and Secondary Education Act, evaluation studies conducted a few years after passage of the 1965 legislation produced widespread evidence that disadvantaged students were showing no improvement in basic learning skills. Yet studies conducted after a decade of implementation revealed significant improvements in the administration of the program and a number of substantial improvements in educational performances. The new findings suggest that there was a pattern of learning by program administrators and their congressional supporters as they identified obstacles and then devised various strategies to deal with them.[14]

To sum up, the greatest problems of public accountability and policy evaluation are associated with the choice of criteria by which to measure success. Experts and citizens alike must face the inevitable conflict between crude but intuitively appealing criteria on the one hand, and more refined but also more controversial criteria on the other. This conflict may not have been so serious once. As Geoffrey Vickers has observed, there have been times in the not-so-distant past when popular expectations were relatively clear, realistic and verifiable—the maintenance of law and order, protection against foreign aggression, a stable currency, a stable level of taxation, relief of extreme poverty. Today we expect much more from our government, but we do not know precisely how *any* government could fulfill our expectations.

At the same time, change—largely self-induced, as we saw in the last chapter—has become so rapid that the past becomes an ever less reliable guide to the future. Thus, policy outcomes become increasingly elusive both because we are less certain about the limits of the possible in public policy and because we

14. Paul A. Sabatier, "What Can We Learn from Implementation Research?", in Franz-Xavier Kaufmann, Giandomenico Majone, and Vincent Ostrom, eds., *Guidance, Control, and Evaluation in the Public Sector* (Berlin: Walter de Gruyter, 1986), 313–26.

suspect that the most important results may not yet have had time to appear. In these conditions the mere comparison of immediate results with expectations is likely to be uninformative as well as inconclusive.

According to social psychologists, learning is the dominant form in which rationality exhibits itself in situations of great cognitive complexity. This suggests that the rationality of public policy-making depends more on improving the learning capacity of the various organs of public deliberation than on maximizing achievement of particular goals.

It is not the task of analysts to resolve fundamental disagreements about evaluative criteria and standards of accountability; only the political process can do that. However, analysts can contribute to societal learning by refining the standards of appraisal of public programs and by encouraging a more sophisticated understanding of public policies than is possible from a single perspective. The need today is less to develop "objective" measures of outcomes—the traditional aim of evaluation research—than to facilitate a wide-ranging dialogue among advocates of different criteria.