DUKE ENVIRONMENTAL ECONOMICS WORKING PAPER SERIES
organized by the
NICHOLAS INSTITUTE FOR ENVIRONMENTAL POLICY SOLUTIONS

Incorporating Evaluation into the Regulatory Process

Lori S. Bennear*
Katherine Dickinson†

Working Paper EE 11-06
July 2011

* Nicholas School of the Environment, Sanford School of Public Policy, and Department of Economics, Duke University
† National Center for Atmospheric Research

Duke UNIVERSITY

Duke NICHOLAS INSTITUTE FOR ENVIRONMENTAL POLICY SOLUTIONS

Incorporating Evaluation into the Regulatory Process

Lori S. Bennear*
Katherine Dickinson

This paper was prepared for the workshop
"Crisis And The Challenges Of Regulatory Design"
Kenan Institute For Ethics, Duke University
June 2-3 2011

Date of Current Draft: June 27, 2011

Abstract: For the last two decades there have been substantial attention and resources devoted to increasing evaluation of government programs in an effort to promote evidenced-based and performance-based policies. However, federal efforts to promote evaluation through the Government Performance and Results Act and the Performance Assessment Rating Tool have had limited success. This paper seeks to evaluate the recent efforts at evaluation and provide guidance for how future efforts can be shaped. We provide a stylized model for evaluation in the regulatory process that is consistent with prior federal initiatives. We then examine four categories of barriers to implementation of this stylized model—cognitive barriers, social/cultural barriers, organizational barriers, and incentive barriers. We then present suggestions for how future evaluation efforts can be formulated to better overcome these barriers.

JEL Codes: H1, K00

Key Words: Government Performance, Evaluation, Regulation, Incentives, PART, GPRA, GPRAMA

* Bennear is an Assistant Professor of Environmental Economics and Policy at the Nicholas School of the Environment, Sanford School of Public Policy, and Department of Economics, Duke University. She can be reached at lori.bennear@duke.edu. Dickinson is a post-doctoral fellow at the National Center for Atmospheric Research. She can be reached at katiedickinson@gmail.com.

1 Introduction

In his inaugural address in January 2009, President Barack Obama said *"The question we ask today is not whether our government is too big or too small, but whether it works […]."* In the two years since this speech, all major political initiatives have been shaped by two competing tensions. First, there is tension over rising federal deficits, an issue that pre-dated the Obama administration, but has been made more salient by the recession, the subsequent stimulus spending, and looming costs of federal medical entitlement programs. Second, there is tension over failures in regulation that lead to significant disasters, most notably the financial crisis and the Deepwater Horizon oil spill. This has lead to a political climate in which a sizeable fraction of Americans believe government is both too big and not working. In this context, a focus on evaluation—determining "what works", promoting efficiency, and reducing waste—has become a political imperative.

Yet the desire to use evaluation to improve government performance, and the development of institutional mechanisms to promote evaluation, are hardly new. As the quotes in Figure 1 demonstrate, evaluation and performance-based governance have been discussed by every president, both Democrat and Republican, in the last two decades. Indeed, the core components of government evaluation, including collection of data on program outcomes, the development of reporting mechanisms, and the rigorous linkage of performance data to budgetary allocations, are roughly a century old. These principles were first institutionalized on the municipal level by the New York Bureau of Municipal Research as early as 1912 (Williams, 2003). In the 1960's the Johnson administration implemented the Planning Programming Budgeting System (PPBS), which provides a detailed set of guidelines for developing and evaluating policy alternatives (DonVito, 1969). This initiative was followed by the Management by Objectives initiative under the Nixon administration and the

Zero-Based Budgeting initiative in the Carter Administration (D. Moynihan, 2009).

The modern era of performance-based governance dates from the 1990s, when the Clinton Administration's efforts to "reinvent government" sought to institutionalize a performance-based governance system, which required agencies to set performance goals, regularly assess performance, and use the results of these assessments to improve programs. In 1993, Congress passed the Government Performance and Results Act (GPRA). The act was motivated by a desire to improve government performance, budgeting, and program oversight. It mandated that federal programs had to set performance-based goals, measure progress toward those goals, and publicly report this progress. While GPRA sets out a framework for performance-based planning and evaluation, the legislation required few specific requirements. For example, while GPRA requires agencies to report on progress towards their performance goals, it provides no guidance on what constitutes acceptable evidence of progress, nor stipulates any consequences for performance shortcomings. In particular, the GPRA legislation does not directly tie performance reports to the budgeting process.

Partially in response to the lack of budgetary "teeth" of GPRA, the George W. Bush Administration began a program to promote performance-based evaluation of government projects within the executive branch (Donald P. Moynihan, 2008). Drawing on prevalent techniques of evaluation in the corporate world, the Office of Management and Budget developed a Program Assessment Rating Tool (PART) to evaluate government programs and to link program performance to budget decisions.

There were four components of the PART: program purpose and design, strategic planning, program management, and program results and accountability. Each part contained a series of yes/no questions, with every question receiving a weight and the total score reflecting the weighted average of scores, where "yes" receives a point value of one and "no" a point value of zero. For example, in the program purpose or design section there were five questions

focused on the mission of the program, its clarity, and potential duplication with other programs. The strategic planning section contained eight questions related to short and long term goals, the existence of performance measures and baselines from which improvements in these measures could be judged, independent mechanisms of evaluating program progress, and the connection between such evaluations and the budgeting and strategic planning process. The management section contained seven questions regarding the collection of information within the program, management of contractors and program partners, and management of funds. The section on program results and accountability contained five questions on whether the program had attained its short-term goals and was making progress toward long-term goals. The instructions for PART included detailed guidance on what constituted a "yes" response, often with specific examples drawn from actual programs.[1] For programs that had well-defined performance metrics, the overall score resulted in a ranking of effective (85-100), moderately effective (70-84), adequate (50-69) and ineffective (49 or less). Programs without well-defined metrics received a rating of "results not demonstrated." Coverage was nearly universal; between 2002 and 2008, 98% of federal programs had been evaluated using the PART system.

Despite the purported desire by the White House and OMB to link performance to budgeting, there is little evidence that the PART ratings had much of an impact on subsequent budget allocations (Buell, 2011; Frisco & Stalebrink, 2008; Gilmour & Lewis, 2006a, 2006b; C. Heinrich, 2009). Heinrich (2009) examined correlations between the performance measures in PART and subsequent budget allocations for programs in the Department of Health and Human Services and finds no significant correlations. Buell and Tortorella (2011) conduct a similar study for environmental and natural resource programs and also find no correlation between performance measures and budget allocations.

[1] A complete list questions and guidance is available at http://www.expectmore.gov

Others have found some positive correlation between PART scores and OMB budget recommendations (Gilmour & Lewis, 2006a, 2006b). Even in these cases, PART data were not linked with budgets systematically and the use of PART data in budgeting apparently reflected political factors (Gilmour & Lewis, 2006a, 2006b). There is little evidence that PART data played a significant role in congressional budget deliberations (Frisco & Stalebrink, 2008).

Concern for institutionalizing performance reviews have continued in the Obama Administration. In January 2011, Congress passed, and President Obama signed, the Government Performance and Results Act Modernization Act (GPRAMA). GPRAMA shifts away from comprehensive evaluation at the program level toward establishment of a few key strategic performance goals at the agency level, which should be achievable in a 12-18 month timeframe. The legislation requires that agencies develop a set of outcome-based measures of performance for each goal and report annually on progress. Each agency also must hire a senior-level Performance Improvement Officer to oversee the strategic planning and reporting process.

With nearly a century of experience in the core elements of performance-based evaluation and two decades of significant investment in institutions to promote performance measurement and evaluation in the federal government, it is disheartening that recent public opinion polls show faith in government at all-time lows (Gallup, 2011). Furthermore, in a period marked by financial and environmental catastrophe, widely believed to be due to lax government oversight and regulation, the majority of the American public believes the government is too large and too intrusive (Gallup, 2010). Nonetheless, in response to the financial crisis a new federal regulatory agency has been created to help protect the public from financial chicanery, and in response to the Deepwater Horizon oil spill there has been significant restructuring of the Minerals Management Service into the Bureau of Ocean Energy Management, Regulation, and Enforcement. Significant new regulatory programs are expected in both agencies and ultimately both agencies will need to use the tools of

performance evaluation to develop successful regulations to achieve their missions and to communicate those achievements to taxpayers.

The goal of this paper is to provide guidance to these new regulatory agencies as they grapple with how to incorporate performance measurement and evaluation into the regulatory process. In Section 2, we begin by presenting a stylized model of evaluation in the regulatory process. This model contains the key features of evaluation that have been institutionalized in GPRA, PART, and GPRAMA over the last two decades. In Section 3 we discuss a set of four barriers to the implementation of the idealized model. Drawing on empirical evaluations of GPRA, PART and similar state programs, we argue that combinations of these barriers have frequently prevented performance-based institutions from having their desired effect of enhancing government performance and not-infrequently have led to perverse outcomes. In Section 4 we present several ideas for improving the use of evaluation in the regulatory process. Section 5 offers our conclusions and areas for future research.

2 A Stylized Model of Evaluation in the Policy Process

Historically, many regulatory agencies have been concerned largely with *processes*—developing regulations, writing permits, conducting inspections, issuing compliance orders and fines, encouraging participation in voluntary programs, and so forth (Bennear & Coglianese, 2004; GAO, 2000; Metzenbaum, 1998). While attention to process certainly has some connection to achieving desired policy outcomes, these relationships are imperfect and often poorly understood. Moreover, evaluation of agency activities typically has been based on whether the agency has complied with administrative processes rather than on the actual performance in attaining its mandate. This process-focused approach has led to "bean counting," whereby agency efficacy is determined by the number of inspections, the number of fines, the number of "beans" (GAO, 2000; Metzenbaum, 1998). In contrast, a performance-focused system

investigates the effect of an agency's activities on the ultimate policy outcomes of interest.

Scholars of public policy have often envisioned performance-focused policy formulation and implementation as a linear process involving three key steps: planning, action, and evaluation. This process is diagrammed in Figure 2.

The planning stage typically involves: 1) gathering information relevant to the decision; 2) integrating the information (e.g., deciding what information is valid, weighting aspects or attributes, making tradeoffs); and 3) assessing the information in order to determine the best course of action. The stylized model implies certain behaviors at each stage of this process. Information gathering should be thorough and efficient, covering all, or at least sufficient, relevant information sources, and ignoring irrelevant data. In the integration stage, decision makers should objectively judge the validity of information. They must then weight different kinds of information based on some stable set of social preferences, making tradeoffs between competing objectives (e.g., minimize costs, maximize safety). Finally, decision makers should assess different options and apply a consistent rule (e.g., expected social welfare maximization) to decide among different policies or approaches.

The action stage consists of implementing the policy that was identified in the planning stage. In practice, of course, this step is more complicated. Even when a decision maker knows what she should do or would like to do, a number of temporal, logistical, budgetary, administrative, and other constraints may prevent her from doing it. These points are discussed in more detail in the next section.

The next step in the policy cycle is evaluation. The purpose of this evaluation is to provide a measure of the policy's outcomes. Did the policy have the intended results? Were there any unforeseen or unintended consequences? In retrospect, how well did this policy perform? Ideally, evaluation of a particular policy or program should have several important characteristics.

First, the evaluation should measure the policy's impacts on a range of relevant outcomes. The most relevant outcomes have four characteristics: (1) they are aligned with policy's objectives and goals; (2) they approximate actual performance; (3) they are measurable at reasonable costs and with reasonable administrative effort; and (4) they are difficult to manipulate (GAO, 1997a, 1997b; C. J. Heinrich, 2002).

Second, the evaluation should attempt to separate the effects of the policy under study from other, simultaneously occurring events. This involves constructing some counterfactual account of what would have happened in the absence of the policy. This implies that data must be available on a "control group" or some reasonable model must be constructed to estimate this counterfactual.

Third, the evaluation should employ methods appropriate to the relevant questions and available data. Some questions are best answered through quantitative, econometric studies (e.g., What were the measurable impacts of this policy on variable X?), while others require more qualitative approaches (e.g., Why did this policy succeed or fail? What were stakeholders' attitudes toward this approach?).

Systematic, objective evaluations of this kind connect action and planning, providing a fundamental feedback loop that is essential to a performance-based system. Thus, evaluation marks the end of one cycle of policy design, but it can also mark the beginning of the next cycle, so long as the results of the evaluation shape understandings about the effectiveness of different policy options and inform future choices (Brewer and deLeon 1986, Bennear and Coglianese 2004).

The history of performance-based initiatives over the last two decades (described in the introduction) reflects the role of evaluation as presented in this linear model. From this perspective, the problem of ineffective governance can be solved by requiring systematic evaluation of performance (reflected in GPRA, PART, and GPRAMA) and potentially linking these evaluations to budget decisions (reflected in PART).

Not surprisingly, actual government decision making often deviates from the stylized performance-based paradigm. In the next section, we outline some of the possible reasons for this divergence. We provide a conceptual framework that identifies the cognitive, social, structural, and incentive barriers to achieving an optimal performance-based system. Since the purpose of this paper is to examine how government can improve the use of evaluations in policy design, we tend to focus on barriers to the use of evaluations in the learning process and on barriers to conducting evaluations after programs have been implemented.

3 Barriers to Evaluation in the Policy Process

Barriers to implementing the ideal policy model within government agencies can be grouped into four broad categories: cognitive barriers, social/cultural barriers, barriers in organizational structure, and incentive barriers. There is extensive literature drawing on psychology, economics, management, sociology, and political science on these four categories of barriers and we cannot do justice to all of these literatures in this summary review. Instead we highlight the key insights from each of these fields in order to illustrate how a given regulatory agency might be able to address these constraints on effective policy-making. Wherever possible we draw on empirical evidence of the impact of these barriers on the performance of prior evaluation initiatives, including GPRA and PART. A summary of all four sets of barriers and how they inhibit the implementation of the stylized model can be found in Table 1.

3.1 Cognitive Barriers

Recent work in behavioral decision theory highlights several problems that can short-circuit the stylized performance-focused model. This literature provides key insights in assessing what kinds of information people are likely to use in their decision making processes, how they weight and integrate that information, and how this information influences the ultimate decision. One cognitive barrier to the effective use of evaluations during the planning part of

the policy process is the reliance on various heuristics for decision making. Rather than collecting, processing, and evaluating all of the available information before making a decision about what action to take, people (and hence agencies composed of people) may rely on "rules of thumb" or shortcuts in the learning process (Tversky & Kahneman, 1974).

One of the most relevant heuristics in the context of our ideal plan-act-evaluate model is the availability heuristic. This shortcut consists of judging the likelihood of an uncertain event by the ease with which instances of similar events can be brought to mind. In the context of government agency decision making, the availability heuristic may have a strong influence on the information gathering and integration steps, determining what evidence key decision-makers deem to be the most relevant for the task at hand and giving more weight to certain kinds of information. This can lead to biases if, for example, the availability heuristic causes policy-makers to focus on only very recent experience or certain kinds of studies (e.g., descriptive case studies or in-house evaluations).

Another important and well-documented tendency is "confirmation bias," or the tendency of people to seek information that confirms their prior beliefs about the correct action (Klayman, 1995). For example, an agency that is already convinced that policy X reduces pollution may look for evidence that policy X indeed decreased pollution in some specific case and discount any evidence that suggests that policy X did not decrease pollution. This heuristic may lead to less than optimal use of information from prior evaluations in program design. It may also bias the results of evaluations themselves, particularly if such assessments are conducted "internally" by individuals with a stake in the evaluations' results.

A third set of cognitive barriers concerns how agencies integrate and assess information in decision-making. Behavioralists have shown that people may be biased toward maintaining the status quo because the strain in weighing the different attributes (both positive and negative) of a decision to change away

from the status quo is demanding. This can lead to decision avoidance (Anderson, 2003). Similarly, research has shown that people tend to prefer errors of omission (doing nothing) to errors of commission (doing the wrong thing) which may also lead to a bias toward the status quo (Spranca, Minsk, & Baron, 1991).

Furthermore, there are cognitive barriers that can interrupt the cycle between planning and action. The disconnect between "knowing what to do" and "doing it" is well illustrated by concept of a "predictable surprise," first introduced by Bazerman and Watkins (2004). These authors define a "predictable surprise" as "an event or set of events that take an individual or group by surprise, despite prior awareness of all of the information necessary to anticipate the events and their consequences." They argue that the terrorist attacks of September 11 and the collapse of Enron represent two examples of contemporary, high profile predictable surprises. The recent catastrophes resulting from Hurricane Katrina, the recent series of financial crises and the Deepwater Horizon oil spill also conform to this definition (Leonhardt, June 6, 2010). In all of these cases, groups of individuals within government recognized that a problem existed, learned about appropriate solutions, and even made recommendations for addressing the problem, but these recommendations were never implemented. Bazerman and Watkins (2004) provide both cognitive and organizational explanations for these failures to connect planning with action. The cognitive barriers include the status quo bias discussed previously (Anderson, 2003; Spranca, et al., 1991) and the tendency for people to undervalue risks, particularly risks involving large losses with low probability. We discuss the organizational roots of predictable surprises in Section 3.3.

3.2 Social Barriers

Cognitive biases afflict individuals. The day to day workings of a performance-based system turn on the ability of an organization (i.e., a government agency) to engage in successful planning, action, and evaluation.

This section explores how interactions among individuals within an organization may produce additional barriers to learning.

An increasingly rich scholarly literature on organizational learning has identified several ways in which membership in an organization shapes individual learning. Perhaps the most important factor involves the way in which an organization socializes those individuals who work for it, influencing their identity. Being a part of an organization creates a sense of shared purpose and a collective set of norms, which influence how an individual sees herself, what goals she pursues, and how she works to achieve those goals. Through the organization, individuals become part of a "community-of-practice" (Brown & Duguid, 1991), interacting and learning from each other and adopting a common viewpoint. These shared norms and common identities serve an important purpose within the organization, facilitating coordination and communication (Kogut & Zander, 1996). However, an individual's group identity also shapes and filters the way he thinks, putting constraints on what options he considers as well as how he searches for information (Kogut & Zander, 1996). At worse, this leads to the type of concurrence-seeking behavior that Janis labeled "groupthink"(Janis, 1973; Janis & Productions, 1983). Characteristics of groupthink that can impede the use of evaluation include a belief in the fundamental morality (righteousness or virtue) of the group and its activities, a shared illusion of unanimity of opinion, and both self- and other- censorship of deviations from the group norm or viewpoint (Janis, 1973; Janis & Productions, 1983). It is perhaps not surprising that people working together on a particular program, often for years, believe deeply in the inherent value of the program activities and may be skeptical of the need for evaluation and the benefits that could be gained from participating in such evaluation.

More generally, identity can be a powerful factor affecting learning, action, and evaluation within an organization. Mendeloff's (2004) comparison of the role of evaluation at the Occupational Safety and Health Administration (OSHA) and the National Highway Transportation Safety Administration

(NHTSA) revealed that OSHA's identity as an enforcement agency contributes to a lack of systematic evaluation, while NHTSA sees itself as a "science-based regulatory agency" and relies more heavily on evaluation. In this case, OSHA's stakeholders play a strong role in reinforcing this identity. The firms and industries OSHA regulates are hostile to evaluation, contributing to an atmosphere in which regulators resist this part of the institutional learning process (Mendeloff, 2004).

These studies are consistent with a study by the Government Accountability Office (2003), which emphasized the importance of creating an "evaluation culture" at agencies that had successfully incorporated systemic assessment into their processes of policy formulation and implementation including the Administration for Children and Families (ACF) in the Department of Health and Human Services and NHTSA. This "institutional commitment to learning from evaluation" (GAO, 2003) had become an important part of these agencies' identities. For example, ACF has a long-history of promoting evaluation. States could request waivers of federal regulations to test innovations in poverty reduction and welfare-to-work programs as long as they committed to rigorous evaluation of these programs. Lessons from these evaluations were critical in the development of the Job Opportunities and Basic Skills Training (JOBS) program as well as the welfare-to-work programs under Temporary Assistance for Needy Families (TANF) (GAO, 2003). NHTSA's evaluation culture consists of a commitment to a three-staged process of learning. The first stage is identifying the nature of the problem and the suite of potential solutions. The second stage is conducting benefit-cost analysis of the potential solutions and settling on a preferred regulatory approach. The final stage is conducting long-term evaluations of the consequence of the new regulation, recognizing that the impacts may not be known for five or more years (GAO, 2003).

3.3 Barriers in Organizational Structure

For decades, scholars of management have understood that any serious attempt to consider the performance of an organization requires attention to its basic institutional structure and whether that "organizational chart" facilitates or impedes the organization's goals (Chandler; Chandler). For evidence of the primary role structure plays in a critique of organizational performance, one need look no further than the U.S. experience following the terrorist attacks on September 11, 2001. Following what by all accounts was a massive performance failure among U.S. intelligence agencies, the immediate proposals to improve performance in this sector focused on restructuring organizational hierarchy, in particular the consolidation of dozens of bureaus and agencies within the Department of Homeland Security and the placement of intelligence agencies within that Department.

This focus on the organizational chart is not surprising. An organization's structure can either obstruct or assist communications, coordination, evaluation and learning. Overly vertical agency structure may result in a series of "fiefdoms," each with its own goals, and projects. Because communication must flow up the hierarchy before managers can pass information across divisional boundaries, this structure limits the ability to learn from others' programs and evaluations. In contrast, overly horizontal structures might create "decision by committee," a situation where everybody is responsible for decisions, and hence, nobody is responsible for decisions. This lack of direct accountability can also impair the policy process.

The 2003 GAO study on agency program assessment offers considerable evidence that organizational structure, networks, and culture have powerful implications for the significance of evaluation in policy-making. In this study, the GAO closely examined five agencies that enjoy a reputation for having a strong performance focus. Its analysts found that the ability to develop collaborative partnerships with other agencies and across federal, state and local levels played a key role in their success. For example, ACF typically issues block

grants to states. Many states experimented with new strategies for helping poor families and ACF required these innovative strategies to be rigorously evaluated. While ACF frequently hired consultants to conduct the evaluations, they still had to build strong collaborative partnerships with state and local personnel upon whom they depended for consistent implementation of the experimental protocols (e.g., compliance with random assignment).

Mendeloff (2004) similarly cites organizational structure as one of the key barriers to performance-based approaches at the Occupational Safety and Health Administration (OSHA). In particular, Mendeloff highlights the fact that the National Institute of Occupational Safety and Health (NIOSH) -- the primary agency responsible for research on occupational health and safety issues -- is not located within OSHA, but rather within the Bureau of Labor Statistics. This feature of OSHA's organizational structure inhibits this agency's ability to conduct evaluations that respond to its information needs and encourage learning and improved performance over time. Bazerman and Watkins (2004) also cite organizational "silos," failures in communication, and duplicative authority across multiple agencies as critical to the development of "predictable surprises."

A further organizational barrier to the effective use of evaluation in policy design involves the level of existing expertise in evaluation methods within different agencies. In 1998, the GAO surveyed 23 different government offices that had conducted some program evaluation during 1995. They found that half of these offices had fewer than 18 full time-equivalent employees (FTEs), and the level of resources spent on evaluation was also low (GAO, 1998). Similarly, Mendeloff (2004) argued that one of the reasons NHTSA performs more evaluations than OSHA is that NHTSA has greater institutional capacity for evaluation activity.

3.4 Incentive Barriers

A final set of barriers to the stylized model concerns the incentives that a particular evaluation process generates. The first incentive barrier stems from

the fact that most policies have multiple goals and multiple outcomes of interest. Not infrequently there are tradeoffs between these goals, such that the promotion of one outcome leads to the demotion of another outcome. Focusing on a subset of outcomes can lead to perverse incentives for program managers. A good example of this common dilemma comes from the federal Job Training Partnership Act (JTPA). The JTPA had many goals, but two of them were to increase earnings, thereby reducing welfare dependence, and to provide training to disadvantaged populations. But there is tension between these two important goals. The most disadvantaged populations require substantially more time, effort, and services to move from welfare to work. Research has repeatedly demonstrated that by measuring success in terms of earnings gains and reductions in welfare payments, program managers reduced services to disadvantaged populations (Heckman, Heinrich, & Smith, 1997; C. J. Heinrich, 1999, 2002). There are also entrenched conflicts between the goals of congressional oversight committees and executive agencies that can reduce incentives for good evaluation. In the past half-century, members of Congress and various Presidential administrations have frequently pledged allegiance to performance-based governance. But these officials also want to please their respective constituencies and these two objectives do not always align. In the wake of the federal government's embrace of PART, several scholars have found some evidence that budget decisions were influenced by PART evaluations, but the effect was small and dominated by political factors (Buell, 2011; Frisco & Stalebrink, 2008; Gilmour & Lewis, 2006a, 2006b; C. Heinrich, 2009). Mid-level managers have little incentive to take evaluation seriously if they see it as disconnected from upper-level decision making.

At the same time, making explicit links between assessment and budgetary allocation can lead to classical principal/agent problems (Grossman & Hart, 1983; Hart & Holmstrom, 1986). In the case of government agency decision making, the principal (Congress or the White House) wants government agencies to work toward one set of objectives and achieve certain performance-based

results outlined in the stylized model. However, there are a number of reasons why the agencies and the people who work in them may not find it in their interest to pursue the principal's goals. One is the desire to protect programs and budgets (Niskanen, 1994). If the agency can manipulate outcome measures (either by choosing metrics or by shaping calculation mechanisms), then the agency has an incentive to overstate performance to protect or enhance its budget. A related concern is that tying performance to budgets can reduce innovation, if agencies shy away from trying new policy approaches for fear of failure. This problem is particularly acute when agencies must use short-term performance measures, leaving little time for learning and program adjustment.

An additional incentive barrier arises from an inability to respond to learning that results from evaluation, which occurs in regulatory agencies when there are statutory limitations on the allowable set of policy changes. For example, the Clean Air Act requires EPA to set ambient air quality standards for six criteria air pollutants at levels "protective of public health," allowing for an adequate margin of error. The agency is not able to consider costs in setting standards, so even if evaluations that reveal that the program is not efficient or cost-effective, the agency cannot respond to such assessments with substantive program revisions. The frustration that arises from the inability to act on evaluations has been demonstrated to significantly weaken their appeal to agency employees in a variety of settings at both the federal and state levels (GAO, 2005; C. Heinrich, 2009; D.P. Moynihan, 2005).

4 Suggestions for the Future

Nearly half a century of intensive efforts in America's federal government to create performance-based policy-making have provided some impetus for careful planning and evaluation, but these efforts have largely failed to make the process of evaluation and learning integral to the regulatory process. We have argued that governmental agencies have structured prior performance-based efforts mostly around the stylized model of evaluation presented in Section 2 and

that those efforts have been significantly hindered by the cognitive, social/cultural, organizational, and incentive barriers discussed in Section 3. In this section we outline our views on how performance-based planning and evaluation could be implemented more successfully, with a particular eye to new regulatory agencies. We make three suggestions: (1) promote an evaluation culture, (2) target priority outcomes, and (3) promote learning through small-scale experimentation.

4.1 Promote an Evaluation Culture

A key reason for the lack of success of earlier performance-based initiatives stems from their adherence to the stylized model of evaluation in Section 2. This model proceeds in a linear fashion from planning to action to evaluation. There is room for evaluation to feedback into planning, but too often evaluation constitutes the end of the process rather than a part of an ongoing cycle. Even more problematically, agencies frequently tack on some process of evaluation after they have fully implemented a program, in response to a mandate from a higher authority. Anecdotally, it appeared that many regulatory programs, scrambled to develop outcome measures to satisfy PART requirements, an approach hardly conducive to the thoughtful construction of meaningful assessment designed to promote learning.

We present a more useful model for thinking about the role of evaluation in the regulatory process in Figure 3. In this cyclical model, evaluation clearly plays a role both as a means of learning about the effects of prior actions and as a means of information for future planning. The other key difference in this stylized model is the central role of culture and norms. If evaluation is to count for more than a check-list that staffers complete and then ignore, agency leaders and mid-level managers must view it as an integral part of the regulatory mission. In other words, a commitment to learning and continual improvement must be embedded in the culture of the agency. The model in Figure 3 highlights the central role of an evaluative culture/norms in the regulatory process.

A key feature of such a culture is that evaluation, data collection, and data analysis become part of the very initial stages of regulatory planning. From the beginning, program officials should be asking questions like:

- If this program is working well (not well), what would we expect to see? What types of data could we collect that would help us measure these impacts?
- If success or failure could vary with several components of the regulatory design, how could we distinguish these later on?
- How will we distinguish the impact of our program from the impacts of other programs and general trends?
- How can we design a program that is flexible enough to respond and adapt to information gained through the evaluation process?

If regulatory officials pose these questions from the beginning of the planning process, re-ask them throughout the process, and consistently feed the answers back into the policy process, the evaluation that occurs at the "end" will produce much more useful results, thereby enabling better policy making in the future.

Although prior research has recognized the critical important of an evaluation culture to the success of performance-based initiatives (GAO, 2003; Mendeloff, 2004; D.P. Moynihan, 2005), the shaping of organizational culture generally eludes oversight agencies. Congressional committees or OIRA can require specific actions of regulatory agencies, but cannot directly change institutional cultures. New regulatory agencies, by contrast, lack the constraints of institutional history. Not yet subject to a particular path dependency, they enjoy an especially good opportunity to build an organizational culture that is performance-based.

4.2 Target Priority Outcomes

Building a culture of evaluation requires that people in the organization feel that the time and energy spent on evaluation is useful and not just a series of bureaucratic hoops through which they must jump. Rigorous evaluation is costly, both in terms of dollars and time. These valuable resources should be devoted to formal performance assessment where it can do the most good.

Unlike PART, which eventually evaluated 97% of government programs, we suggest that future evaluation efforts should make no pretense of comprehensiveness. Instead, officials charged with general regulatory oversight, such as congressional committees or OIRA, should target a set of priority outcomes that a variety of agency programs might influence. With evaluation focusing on a narrower set of programs and outcomes, the government could then more readily expend the resources to ensure that these outcomes are well-measured and that the evaluation is well-conducted. Well-measured outcomes may be short or long term measures as appropriate. Well-conducted evaluations should take care to develop a credible counterfactual. It is worth noting that GPRAMA moves nicely in this direction by having agencies develop goals for agency-wide strategic priorities, but falls short by requiring 12-18 month outcomes for all priorities. For example, a mission-critical goal at EPA might be to improve ecosystem health. But improving ecosystem health is not something that is likely to occur in dramatic fashion in 1.5 years. It is more important in this case to establish a set of short, medium and long term metrics; develop data collection plans to populate those metrics; and develop a process for adapting program design and implementation in response to these metrics.

Finally, the federal government, most plausibly though OIRA, should direct resources toward contexts in which evaluation can lead to substantial learning and evidence-based revision of policy. Programs that are limited in statutory authority and so hard to change without further congressional action should not receive the same kind of intensive evaluation resources. Good targets have some agency discretion, so that information about the impact of particular choices can inform policy revisions or future policies.

4.3 Promote Learning Through Small-Scale Experimentation

Focusing on the success of a set of programs/policies in achieving the agency's strategic mission-related outcomes allows for experimentation, which is a key component of evaluation. Experimentation with new approaches,

however, presents significant challenges if officials lack prior evidence of the impacts of prevailing policies, and if the individual program will be held strictly accountable for positive performance. Some experiments necessarily fail, but a lot of learning can come from failures.

Rather than experimenting program-by-program, the government should seek to develop learning communities around sets of small-scale policy experiments designed to have broad applicability across agency programs. Here we use the term "experiment" broadly and do not necessarily imply randomized controlled trials. Officials in charge of assessment should design all experiments carefully to maximize the potential usefulness of the approach, and then disseminate findings widely. These learning communities should embrace "productive failures," failures of experiments that lead to substantial learning about the viability of different approaches, in addition to experimental successes.

An example of a potentially productive learning community might be one that focuses environmental labeling and information disclosure programs. Labeling and disclosure are used extensively at EPA and DOE to inform the public about environmental risks and product attributes ranging from mercury in fish, lead in housing paint, toxics chemical releases, energy consumption of appliances, fuel efficiency of automobiles, levels of chemicals in drinking water, etc. A learning group might focus on what the research has shown worked well, didn't work well, how to design small experiments to test new hypotheses about information dissemination, and how to adapt existing programs in light of these findings. This is quite distinct from examining the causal impact of individual information programs (which has been done in academia) to examining generalizable lessons from these individual evaluations that may improve programmatic efforts and also identifying key aspects that are still not well understood and promoting experimentation along those particular policy dimensions.

5 Conclusions

Performance-based policymaking has been a part of the public discourse for decades, articulated at the highest levels of government and reflected in the public's desire for a more effective and accountable regulatory system. Despite the widespread appeal of this goal, however, we have outlined several barriers that may help to explain why evaluation continues to play a limited role in many policymaking contexts. Overcoming these barriers will not be easy, particularly for entrenched agencies with established cultures and ways of doing business. Newly established agencies, on the other hand, face a unique opportunity to confront barriers head-on and to design themselves in a way that aligns more closely with a model that weaves evaluation into a more circular policy process. We believe that by building a culture of evaluation, targeting priority outcomes, and using small-scale experiments to promote learning, regulatory agencies that are currently being established could set the example and take an important step toward performance-based policymaking.

There are at least three broad areas of future research in which academics can help further to process of evaluation-based governance that we outline in this paper. First, this study (and most others) focuses on evaluation initiatives at the federal level in the United States. More comparative research is needed to understand how evaluation has or has not been successfully implemented in other countries or at the state and local levels. What are the levers of control? What set of incentives have been used? How has evaluation been institutionalized?

A related set of question arises from the dearth of research on the political economy of evaluation. What we have proposed is a fairly technocratic approach to evaluation-based governance. But agencies exist in a democratic, pluralistic, and frequently politically polarized system. When have successfully learning communities developed? What are the political factors that determine the successful (or unsuccessful) implementation of an evaluation-based regulatory structure?

Finally, there is a need for methodological research and guidance on how best to synthesize a wide range of studies on a particular regulatory topic. Certainly methods of quantitative program evaluation and qualitative evaluation are well developed. And methods for synthesis of quantitative studies in the form of meta-analysis are also well-developed. But our recommendation for the development of cross-cutting learning communities hinges on the ability of these government employees to synthesize and learn from both quantitative and qualitative research. Guidance on how best to do this will be critical for the success of these learning communities.

Bibliography

Anderson, C. J. (2003). The psychology of doing nothing: forms of decision avoidance result from reason and emotion. [Review]. *Psychological bulletin, 129*(1), 139-167.

Bazerman, M., & Watkins, M. (2004). *Predictable Surprises: The Disasters You Should Have Seen Coming and How to Prevent Them*. Boston, MA: Harvard Business School Press.

Bennear, L. S., & Coglianese, C. (2004). Measuring Progress: Program Evaluation of Environmental Policies. *Environment, 47*(2), 22-39.

Brown, J. S., & Duguid, P. (1991). Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation. *Organizational Science, 2*(1), 40-57.

Buell, N. (2011). *An Analysis of the Program Assessment Rating Tool: Measuring the Performance of Federal Environmental and Natural Resource Programs.* Duke University.

Chandler, A. D. *Strategy and structure : chapters in the history of the industrial enterprise*: Cambridge, Mass. : M.I.T. Press, c1990.

Chandler, A. D. *The visible hand : the managerial revolution in American business*: Cambridge, Mass. : Belknap Press, 1977.

DonVito, P. A. (1969). The Essentials of a Planning-Programming-Budgeting System. Santa Monica, CA:: RAND Corporation.

Frisco, V., & Stalebrink, O. J. (2008). Congressional use of the program assessment rating tool. *Public Budgeting & Finance, 28*(2), 1-19.

Gallup. (2010). Majorities in U.S. View Gov't as Too Intrusive and Powerful Retrieved May 10, 2011, from http://www.gallup.com/poll/143624/majorities-view-gov-intrusive-powerful.aspx

Gallup. (2011). Trust in Government Poll Retrieved May 10, 2011, from http://www.gallup.com/poll/5392/trust-government.aspx

GAO. (1997a). Managing for Results: Analytical Challenges in Measuring Performance. . Washington, DC: Government Accountability Office.

GAO. (1997b). Managing for Results: Prospects for Effective Implementation of the Government Performance and Results Act. Washington, DC: Government Accountability Office.

GAO. (1998). Program Evaluation: Agencies Challenged by New Demand for Information on Program Results. Washington, D.C.: Government Accountability Office.

GAO. (2000). Managing for Results: EPA Faces Challenges in Developing Results-Oriented Performance Goals and Measures. Washington, D.C.: Government Accountability Office.

GAO. (2003). Program Evaluation: Culture and Collaborative Partnerships Help Build Agency Capacity. Washington, D.C.: Government Accountability Office.

GAO. (2005). Performance Budgeting: PART focuses attention on program performance, but more can be done to engage Congress. Washington, D.C.: Government Accountability Office.

Gilmour, J. B., & Lewis, D. E. (2006a). Assessing performance budgeting at OMB: The influence of politics, performance, and program size. *Journal of Public Administration Research and Theory, 16*(2), 169.

Gilmour, J. B., & Lewis, D. E. (2006b). Does performance budgeting work? An examination of the Office of Management and Budgetís PART scores. *Public Administration Review, 66*(5), 742-752.
Grossman, S. J., & Hart, O. D. (1983). An analysis of the principal-agent problem. *Econometrica: Journal of the Econometric Society*, 7-45.
Hart, O., & Holmstrom, B. (1986). *The theory of contracts*: Dept. of Economics, Massachusetts Institute of Technology.
Heckman, J., Heinrich, C., & Smith, J. (1997). Assessing the performance of performance standards in public bureaucracies. *The American Economic Review, 87*(2), 389-395.
Heinrich, C. (2009). *How Credible is the Evidence and Does it Matter?: An Analysis of the Program Assessment Rating Tool*. La Follette School of Public Affairs, University of Wisconsin-Madison.
Heinrich, C. J. (1999). Do government bureaucrats make effective use of performance management information? *Journal of Public Administration Research and Theory, 9*(3), 363.
Heinrich, C. J. (2002). Outcomes-based performance management in the public sector: Implications for government accountability and effectiveness. *Public Administration Review, 62*(6), 712-725.
Janis, I. L. (1973). GROUPTHINK AND GROUP DYNAMICS: A SOCIAL PSYCHOLOGICAL ANALYSIS OF DEFECTIVE POLICY DECISIONS*. *Policy Studies Journal, 2*(1), 19-25.
Janis, I. L., & Productions, C. (1983). *Groupthink*: CRM Productions/McGraw-Hill Films.
Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation, 32*, 385-418.
Kogut, B., & Zander, U. (1996). What firms do? Coordination, identity, and learning. *Organization science*, 502-518.
Leonhardt, D. (June 6, 2010). Spillonomics: Underestimating Risk, *The New York Times,* p. MM3.
Mendeloff, J. (2004). Evaluation in Two Safety Regulatory Agencies. Washington, D.C.: AEI-Brookings Joint Center for Regulatory Studies.
Metzenbaum, S. (1998). Making Measurement Matter: The Challenge and Promis of Building a Performance-Focused Environmental Protection System. Washington, D.C.: Brookings Institute.
Moynihan, D. (2009). The Politics Measurement Makes: Performance Management in the Obama Era. *Forum-a Journal of Applied Research in Contemporary Politics, 7*(4), -. doi: Artn 7
Moynihan, D. P. (2005). Goal based learning and the future of performance management. *Public Administration Review, 65*(2), 203-216.
Moynihan, D. P. (2008). *The dynamics of performance management : constructing information and reform*. Washington, D.C.: Georgetown University Press.
Niskanen, W. A., Jr. (1994). *Bureaucracy and Public Economcis*. Aldershot, U.K.: Edward Elgar.
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice* 1. *Journal of Experimental Social Psychology, 27*(1), 76-105.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124.
Williams, D. (2003). Measuring government in the early twentieth century. *Public Administration Review, 63*(6), 643-659.

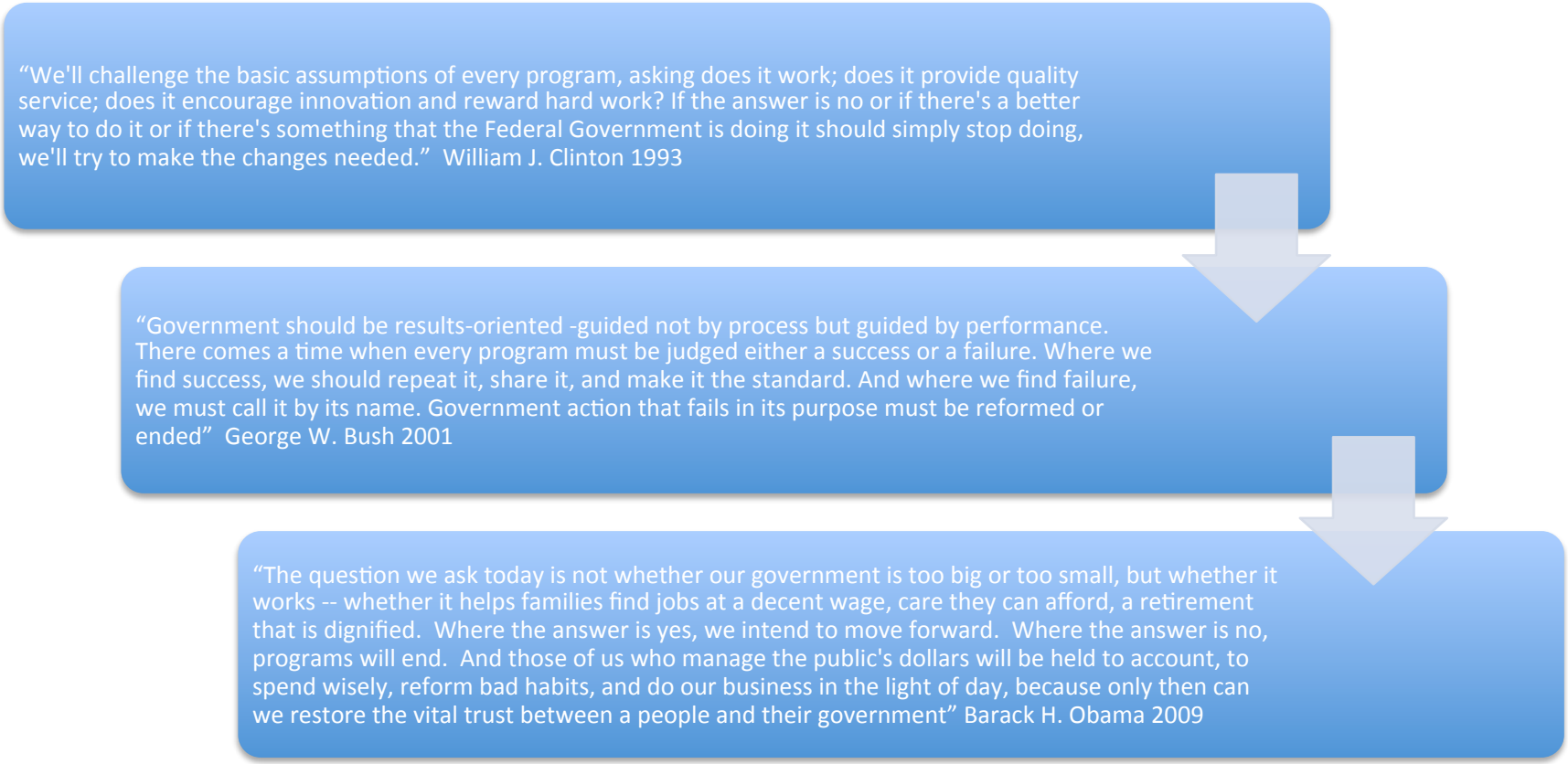

Figure 1: Presidential Views on Evaluation and Performance-Based Governance 1993-2009

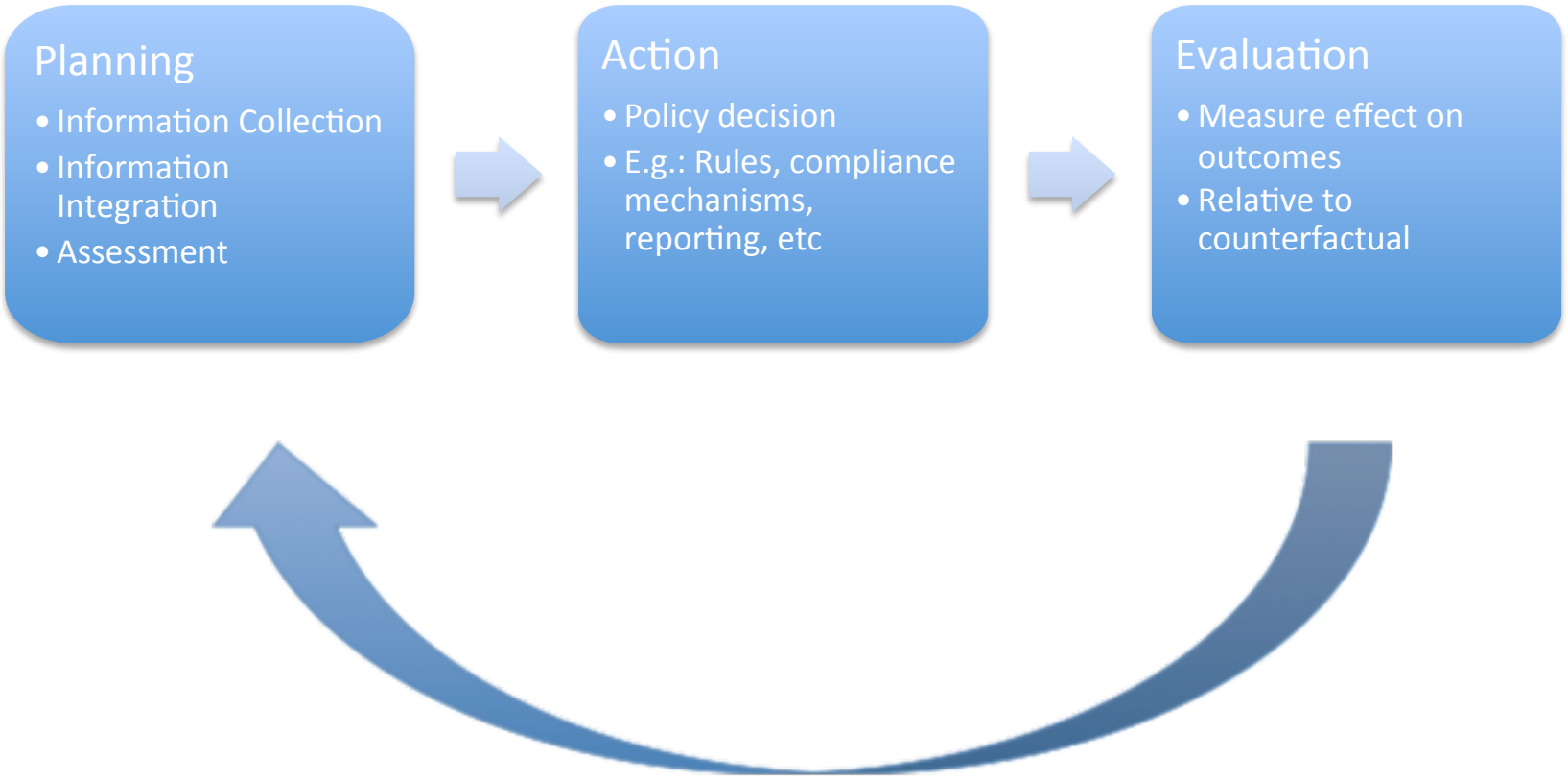

Figure 2: Stylized Model of Evaluation in the Policy Process

Figure 3: Revised Stylized Model of Evaluation in Policy Process

Table 1: Stylized model of Evaluation and potential barriers

	Planning	Action	Evaluation
Stylized Model	Gather all relevant information, including evaluations of past programs Assess validity of information Identify optimal course of action identified	Adopt optimal course of action	Collect data on outcomes Evaluate effect of program on outcomes Make evaluations available to inform future policies
	BARRIERS:		
Cognitive	Reliance on heuristics (e.g., availability heuristic) Confirmation bias Status quo bias Bias toward errors of omission vs. errors of commission	Status quo bias Bias toward errors of omission vs. errors of commission	Confirmation bias
Social	Identity Sense of agency's purpose Communities of practice Cognitive dissonance: hard to accept evidence of past failures	Social rewards/punishments Identity Norms	Evaluation culture Norms that reinforce strong priors and confirmation bias Cognitive dissonance: hard to accept evidence of past failures
Organizational structure	Information sharing "Silos" Hierarchy Communication costs (time, money, effort)	Individuals with knowledge don't have authority to act Budget concerns	Budget for evaluation? Expertise for evaluation? Organizational mandate Collaborative partnerships
Incentives	Incentives to share information Incentives to search for information	Principal-agent problems: Disconnect between desired action and budget-maximizing choice	Incentive to conduct evaluations Fear of revealing poor performance