This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.

Toward a Culture of Consequences

Performance-Based Accountability Systems for Public Services

Brian M. Stecher, Frank Camm, Cheryl L. Damberg, Laura S. Hamilton, Kathleen J. Mullen, Christopher Nelson, Paul Sorensen, Martin Wachs, Allison Yoh, Gail L. Zellman, with Kristin J. Leuschner

RAND EDUCATION

The research described in this report was conducted within RAND Education, a unit of the RAND Corporation, under a grant from a private philanthropic organization.

Library of Congress Cataloging-in-Publication Data

Toward a culture of consequences : performance-based accountability systems for public services / Brian M. Stecher ... [et al.].
p. cm.
Includes bibliographical references.
ISBN 978-0-8330-5015-1 (pbk. : alk. paper)
1. Government accountability—United States. 2. Organizational effectiveness—United States. 3. Performance—Management. I. Stecher, Brian M.

JK421.T79 2010
352.3'5—dc22

2010027552

Cover image courtesy Fotosearch

Published 2010 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

During the past two decades, performance-based accountability—the application of incentives on the basis of measured outcomes as a means of improving services to the public—has gained popularity in a wide range of public service fields. This monograph presents the findings of a cross-sector analysis of the effectiveness of performance-based accountability systems (PBASs) for improving public services in child care, education, health care, public health emergency preparedness (PHEP), and transportation. The purpose of the study was to examine the empirical evidence about the use of PBASs in these sectors and to analyze the factors associated with effective PBAS design, implementation, and outcomes.

The monograph is directed toward decisionmakers charged with designing PBASs for public services—typically, committees consisting of government agency directors, consultants, service providers, and researchers—who want to know how to develop and implement a system effectively. It should also be of interest to policymakers and their staffs who are charged with deciding whether to adopt a PBAS and why and how to evaluate one.

A companion report presents our analytic framework for describing how a PBAS works and uses the framework to identify appropriate questions to ask when studying the operation and impact of PBASs (Camm and Stecher, 2010).

This research was undertaken within RAND Education, a unit of the RAND Corporation. Funding to conduct the study was provided by a private, philanthropic organization.

Questions and comments about this monograph should be directed to the authors:

Brian Stecher
1776 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
Tel: (310) 393-0411 x6579
Fax: (310) 393-4818
Brian_Stecher@rand.org

Frank Camm
1200 South Hayes Street
Arlington, VA 22202-5050
Tel: (703) 413-1100 x5261
Fax: (703) 413-8111
Frank_Camm@rand.org

More information about RAND Education is available at http://www.rand.org/education/.

Contents

Preface iii
Figure and Tables xiii
Summary xv
Acknowledgments xxxi
Abbreviations xxxiii

CHAPTER ONE
Introduction 1
What Is a Performance-Based Accountability System? 4
Problem Recognition and Adoption 5
Design of the Performance-Based Accounting System 6
Implementation and Monitoring 6
Effectiveness 7
Refinement 7
A Note About Terminology 7
Research Approach 9
Cases Examined in This Study 10
Child Care 10
Education 11
Health Care 11
Public Health Emergency Preparedness 14
Transportation 15
Organization of This Monograph 16

CHAPTER TWO
A Historical Perspective on Performance-Based Accountability Systems 19
Origins 19
Twentieth-Century Efforts to Improve Efficiency and Performance 22
Total Quality Management 23
Other Public-Sector Performance Management Initiatives 25
Government Performance and Results Act 27
National Performance Review 30
Recent State and Local Efforts 31
Chapter Summary 33

CHAPTER THREE
Problem Recognition and Adoption 35
Reasons for Adopting a Performance-Based Accountability System 36
Concerns Over Quality Sometimes Build for Many Years 36
Specific Events Can Also Weigh Heavily in System Creation 38
Sometimes One System Leads to Another 38
Influence of Stakeholder and Governance Context 39
Service Providers Are Usually Influential Stakeholders 40
Influence of Service Consumers and Other Stakeholders Tends to Be Episodic 42
Decentralized Governance Structures Provide Opportunity for Stakeholder Influence 43
Systems Are Often Created Without Clear Agreement Among Stakeholders About Key Design Issues 43
Understanding of Service Production Processes 47
There Is Considerable Variation Across Sectors in the Quality of the Knowledge Base for Performance-Based Accountability Systems 47
Tension Between a Performance-Based Accountability System and Other Oversight Structures 49
Chapter Summary 50

CHAPTER FOUR
The Design of an Incentive Structure to Motivate Behavioral Change 53
Whose Behavior Must Change to Meet System Goals? 55
In the Sectors Examined, System Designers Quickly Identified Whose Behavior Needed to Change 55
System Designers Sometimes Seek to Change Organizational Behavior 56
System Designers Must Ensure That Individuals and Organizations Targeted for Change Can See the Connection Between Their Behavior and System Goals 57
A System Should Distinguish the Contribution of Individuals from That of Teams 59
Individuals and Organizations Targeted in the Nine Cases Have Varying Levels of Control Over Desired Changes 60
Incentive Structure Used to Induce Behavioral Change 62
Context Shapes the Incentive Options Available 62
The Size of an Incentive Should Reflect the Value to the Government of Changing the Targeted Behavior 63
Training and Technical Support Can Sometimes Be Used to Enhance Incentives 67
Cases Studied Varied Widely in the Use of Rewards and Sanctions 68
Chapter Summary 70

CHAPTER FIVE
The Design of Measures That Link Performance to Incentives 73
Options Available for Defining Measures 75
Measures Typically Focus on Outputs, Inputs, Processes, or Structures 75
Measures Can Rate Performance on a Continuous Scale or Apply Categories or Thresholds 76
Measures Can Focus on Current Performance or the Change in Performance Over Time 77
A System Can Link Incentives to a Single Measure, to a Composite Measure, or to Multiple Measures 78

A System Might Employ Measures That Are Not Linked to Incentives 79
Factors That Are Important in Choosing Metrics and Measures for a Performance-Based Accountability System 80
Feasibility, Availability, and Cost Considerations Are Paramount 80
Institutional Context Strongly Influences the Choice of Measures 82
Designers Seek to Align Measures with System Goals, Though This Often Proves Challenging 86
Designers Seek Measures That Service Providers Can Either Control or Strongly Influence 89
Designers Seek Measures That Are Resistant to Service Providers' Manipulation 92
Service Providers Want to Understand How Measures Reflect Their Interests and Are Influenced by Their Behaviors 94
Attributes of Measures Chosen and Factors Influencing Choices 96
Chapter Summary 100

CHAPTER SIX
Implementation and Monitoring 103
Common Pitfalls in Implementing a Performance-Based Accountability System 104
Lack of System Experience and Infrastructure Can Pose Operational and Capacity Challenges 105
Unrealistic Timelines Can Create False Expectations Among Stakeholders 106
System Complexity Can Create Confusion in Implementation 107
Failure to Communicate with Stakeholders Can Limit System Effectiveness 107
Stakeholder Resistance Can Undermine System Credibility and Create a Barrier to Change 109
Potential Strategies for Successful Implementation 110
Exploiting Existing Infrastructure and Securing Other Resources Can Help Shorten Implementation 110
Allowing Systems to Evolve Incrementally Can Reduce Mistakes and Increase Buy-In 111

Incorporating a Pilot-Testing Phase Can Head Off Unexpected Problems 113
Communicating Early, Often, and Through Multiple Channels Facilitates Understanding 115
Engaging Stakeholder Support Is Key to System Success 116
Formative Evaluation Can Be Used to Identify and Correct Implementation Problems 117
Chapter Summary 119

CHAPTER SEVEN
Effectiveness of Performance-Based Accountability Systems 121
Evidence of System Effectiveness in the Five Sectors 122
Performance-Based Accountability Systems Have Captured the Attention of Service-Delivery Providers and Users 122
Performance-Based Accountability Systems Have Been Effective in Motivating Behavior Change 123
Performance-Based Accountability Systems Have Helped Providers Focus Attention on Aspects of Service Needing Improvement 124
Some Evidence Links Performance-Based Accountability Systems to Improvements in Long-Term Outcomes 125
Information About Unintended Consequences and Costs 132
Unintended Consequences Vary Across Sectors 132
A Performance-Based Accountability System Might Not Always Be the Most Cost-Effective Option 134
Gaps in Our Knowledge 135
Chapter Summary 136

CHAPTER EIGHT
Motivating and Improving System Evaluation 139
Reasons to Evaluate 140
What Questions Could Evaluations Answer About Performance-Based Accountability Systems? 141
What Does It Cost to Design and Run a Performance-Based Accountability System? 143
How Cost-Effective Are Particular Systems or System Designs? 143

To What Degree Have Performance-Based Accountability Systems Become Learning Organizations? 143
Are Performance-Based Accountability Measures Adequate Proxies for Long-Term System Goals? 144
How Well Do the System's Incentives Work to Alter Behavior? 146
What Unintended Consequences Are Likely to Emerge? 146
What Contextual Factors Might Hinder the System's Implementation? 147
Possible Reasons for the Limited Evidence About System Effectiveness 147
Political Climate Might Discourage Close Scrutiny of Performance-Based Accountability Systems 147
Call for Evaluation Might Be Considered a Sign of Failure 148
Stakeholders Might Be Comfortable with the Status Quo 149
Evaluating Performance-Based Accountability Systems Is Challenging 150
Uncertain Funding Makes Pilot-Testing and Program Staging Less Appealing 154
Cost Is Also a Key Barrier to Evaluation 154
Some Sectors Are More Focused on Service Provision Than on Experimentation 155
Making Accountability and Evaluation More Appealing 156
Reframe Performance-Based Accountability as One of Several Policy Options 157
Embed the Evaluation Design into the System's Implementation 157
Propose Evaluation Designs That Are Developmentally Appropriate 157
Create a Realistic Evaluation Timeframe 158
Assemble Sufficient Evaluation Expertise 158
Interest an Outside Funder in Supporting an Evaluation 160
Designing Evaluations to Maximize Their Value 160
Consider a Separate Measure to Assess the Validity of Reported Gains 160
Include an Appropriate Comparison Group 161
Embed Evaluation into System Implementation 162
Consider Using a Logic Model 163

Chapter Summary 163

CHAPTER NINE
Conclusions 165
Observations on the Structure and Functioning of a Performance-Based Accountability System 166
Performance-Based Accountability Systems Can Be Effective in the Right Circumstances 167
Performance-Based Accountability Systems Motivate Some Changes in Service-Provider Behavior 170
A System's Structural Details Strongly Influence Providers' Responses 170
Initial Success Is Rare, and the Need for Modification Should Be Anticipated 172
A System's Success Should Be Assessed in Relation to Previous Conditions Within the Sector 173
Practical Insights on System Design and Implementation 174
Whether to Pursue a Performance-Based Accountability System 174
High-Level Design Principles 175
Designing Performance Measures 176
Designing Incentives 178
Implementing a Performance-Based Accountability System 179
Evaluating and Improving a System Over Time 182
Areas for Further Research 185
Final Thoughts 186

APPENDIXES
A. The Five Sectors 189
B. Designs for Evaluation 209

Bibliography 217

Figure and Tables

Figure

1.1. Evolution of a Performance-Based Accountability System..... 5

Tables

S.1. Cases Examined in This Study..... xvii
1.1. Performance-Based Accountability System Cases, by Sector, Examined in This Study..... 12
4.1. Whom System Designers Try to Induce to Change Their Behavior..... 61
4.2. Incentive Structures That System Designers Use to Induce Behavioral Change..... 69
5.1. Attributes of Performance Measures Chosen and Factors Influencing Choices 98

Summary

During the past two decades, performance-based accountability systems (PBASs), which link financial or other incentives to measured performance as a means of improving services to the public, have gained popularity in a wide range of service fields. There are many examples. In education, the No Child Left Behind Act of 2001 (NCLB) (Pub. L. 107-110) combined explicit expectations for student performance with well-aligned tests to measure achievement and strong consequences for schools that do not meet program targets. In child care, quality rating and improvement systems (QRISs) establish quality standards, measure and rate providers, and provide incentives and supports for quality improvement. In the transportation sector, cost-plus-time (A+B) contracting is used to streamline highway construction; in health care, there are more than 40 hospitals and more than 100 physician and medical group performance-based accountability (popularly dubbed pay-for-performance, or P4P) programs in place in the United States. There have also been recent efforts to create performance measures and establish consequences related to the nation's efforts to prevent, protect against, respond to, and recover from large-scale public health emergencies.

While PBASs can vary widely across sectors, they share three main components: goals (i.e., one or more long-term outcomes to be achieved), incentives (i.e., rewards or sanctions to motivate changes in individual or organizational behavior to improve performance), and measures (formal mechanisms for monitoring the delivery of services or the attainment of goals).

Today's PBASs grew out of efforts over many years and many countries to manage the private and public organizations that were growing too large to be overseen by a single manager who knew what everyone was doing. These innovative approaches focused on measuring *performance*, which was originally defined fairly narrowly. Over time, notions about what aspects of performance most mattered broadened and changed. By the 1980s, government organizations were linking performance to incentives in an effort to motivate and direct individual performance and improve organizational outcomes.

But while the use of PBASs has spread in the public sector, little is known about whether such programs are having the desired effect. Research suggests that PBASs influence provider behaviors, a first step toward achieving outcomes, but there is currently little evidence concerning the effectiveness of PBASs at achieving their performance goals, or the experiences of governments and agencies at the forefront of this trend. This monograph seeks to address the gap by examining several examples of PBASs, large and small, in a range of public service areas. This study examines nine PBASs that are drawn from five sectors: child care, education, health care, public health emergency preparedness (PHEP), and transportation (Table S.1). The cases we studied provide useful information on the formation, design, operation, and evaluation of PBASs.

The choice of cases was guided by practical as well as theoretical considerations. On the practical side, we wanted to take advantage of the expertise available at RAND, where empirical research is being conducted on a number of performance measurement and accountability systems in different service areas. On the theoretical side, we wanted to include cases in which services are provided primarily by public agencies (education, transportation), as well as sectors in which services are provided primarily by private organizations but in which the public sector has an important role in governance (child care, health care). We also wanted to include at least one instance in which measurement itself was a challenge (PHEP).

Table S.1
Cases Examined in This Study

Sector	PBAS	Key Incentive
Child care	QRISs	Prestige associated with high rating Financial incentives
Education	NCLB	Graduated set of interventions regarding professional development, instruction, staffing, and school governance (e.g., constraints on use of funds)
	P4P	Cash bonuses, salary increases
Health care	Hospital and physician or medical group P4P programs, including quality "report cards"	Financial payments for high performance or improvement, public recognition, transparency (i.e., clarity and openness) of performance results
PHEP	CDC PHEP cooperative agreement	Withholding of federal funds for failure to meet performance benchmarks
Transportation	A+B highway construction contracting	Financial rewards or sanctions based on time to complete
	CAFE standards	Fines for failure to meet minimum average fuel-efficiency standards
	CAA ambient air pollution conformity requirements	Federal transportation funds subject to conformity with ambient air quality standards
	Transit subsidy allocation formulas	Share of state or regional funding for local transit operators

NOTE: CDC = Centers for Disease Control and Prevention. CAFE = Corporate Average Fuel Economy. CAA = Clean Air Act (Pub. L. 88-206 and its amendments).

Research Approach

The research approach included a broad review of literature related to performance measurement and accountability, the development of an analytic framework to structure our internal discussions about research evidence in the five sectors, a 1.5-day integrative workshop that examined various features of PBASs (e.g., context in which the PBAS arose, measures, incentives, and evaluation approaches), and analysis of

sector-specific empirical results and identification of cross-sector principles. Through this process, we attempted to derive principles that might have general applicability beyond the cases we studied.

Findings

Evidence on the effects of nine PBASs in five sectors shows that, under the right circumstances, a PBAS can be an effective strategy for improving the delivery of services to the public. Optimum circumstances include having the following:

- a goal that is widely shared
- measures that are unambiguous and easy to observe
- incentives that apply to individuals or organizations that have control over the relevant inputs and processes
- incentives that are meaningful to those being incentivized
- few competing interests or requirements
- adequate resources to design, implement, and operate the PBAS.

However, these conditions are rarely fully realized, so it is difficult to design and implement PBASs that are uniformly effective. The following sections highlight the major factors that influence PBAS development and effects in the cases we studied.

Decision to Adopt a Performance-Based Accountability System Is Shaped by Political, Historical, and Cultural Contexts

In the cases we examined, the decision to adopt a PBAS was subject to multiple influences. In many sectors, the process was heavily influenced by the preferences of service providers—the very people whose behavior the PBAS sought to shape. In transportation, for instance, PBASs designed to improve local transit funding have often been strongly influenced by the local jurisdictions that are the subject of the PBASs. Given conflicts among stakeholders, it is perhaps not surprising that PBASs often proceed in spite of a lack of clear agreement on what constitutes performance and on who should be held accountable for

what. In many sectors, there is not a sufficiently strong evidence base to provide scientific guidance to would-be PBAS adopters and designers.

The creation of PBASs might be nurtured by the presence of a strong history and culture of performance measurement and accountability. In education, for instance, measurement of student performance has a long history in the United States, and standardized achievement tests are accepted as an indicator of performance for many purposes. However, such a history does not ensure the smooth adoption of a PBAS. Many PBASs, once created, exist in conflict with other PBASs and governance structures. This is especially the case in sectors with a long tradition of measurement and accountability in which service providers receive funds from multiple sources and through many funding mechanisms (e.g., transportation, health care, education).

Selection of Incentive Structures Has Proven Challenging

PBAS designers face three basic design issues:

- determining whose behavior they seek to change (i.e., identifying individuals or organizations to target)
- deciding on the type and size of incentives
- measuring performance and linking these measures to the incentives they have chosen.

In the PBASs we examined, it was fairly easy in most cases to identify the individuals or organizations that are held accountable for improving service activities and reaching the PBAS goals. It has been more challenging, however, to decide which incentive structures to use to affect the desired behaviors.

Context can have a large effect on the incentive structures that PBAS designers choose. For example, when participation in a PBAS is voluntary, designers of PBASs typically use rewards rather than sanctions. We found that, when the designers of a PBAS worked within a regulatory setting (e.g., NCLB, PHEP), sanctions were more common. In contrast, designers of PBASs in which participation was voluntary—child care and A+B contracting, for example—tended to prefer rewards. The size and details of rewards vary widely across the

PBASs we studied. It is unclear how well the magnitude of rewards is correlated with the benefits of the changes that the PBAS designers seek to induce or the effort that service providers, such as doctors and teachers, must make to comply with these changes.

Design of Performance Measures Requires a Balance Among Competing Priorities

The measures used to quantify performance can vary in many dimensions. PBAS designers must consider a number of competing factors when selecting and structuring measures:

- the feasibility, availability, and cost of measures
- the context within which a PBAS operates
- the alignment of measures with PBAS goals
- the degree of control of the monitored party
- resistance to manipulation by the monitored service activity
- understandability.

The selection of performance measures ultimately requires some trade-offs among these factors. PBAS designers seem to prefer measures that can be collected at low cost or that already exist outside the PBAS. To choose among potentially acceptable measures, a PBAS tends to balance two major considerations: the alignment of a measure with the PBAS's goals and the extent to which the individuals or organizations monitored by the PBAS have the ability to control the value of that measure. A natural tension arises from efforts to achieve balance between these objectives. Over time, the parties that a PBAS monitors might find ways to "game" the system, increasing their standing on a measure in ways that are not aligned with the PBAS goals. Perhaps the best-known example of such manipulation in the cases we examined is the act of "teaching to the test" in an educational setting.

Continuing vigilance and flexibility can help a PBAS manage this tension and maintain the balance between policymakers' priorities and the capabilities of the parties the PBAS monitors. Such a balance tends to be easier to achieve when the measures the PBAS uses are understandable and have been communicated to all parties.

Successful Implementation Must Overcome Many Potential Pitfalls

Even a well-designed PBAS might not yield the desired results if it is not executed effectively. Our review of the literature and the nine cases identified several pitfalls that can occur on the implementation process:

- lack of PBAS experience and infrastructure
- unrealistic timelines
- complexity of the PBAS
- failure to communicate
- stakeholder resistance.

There are many strategies available to address these pitfalls. For example, when building a PBAS, exploiting the existing infrastructure, when possible, and implementing in stages can minimize both the time needed for implementation and the disruptive potential of mistakes before they can compound. Incorporating a pilot-testing phase can also head off a number of problems early. Communicating with stakeholders is also integral to the success of the PBAS, while formative monitoring can be important for identifying and correcting problems that occur during implementation.

Evidence of System Effectiveness Is Limited and Leads to Varying Conclusions by Sector

In general, PBASs have not been subject to rigorous evaluation, and the evidence that does exist leads to somewhat different conclusions by sector:

- In education, it is clear that NCLB and other high-stakes testing programs with public reporting and other incentives at the school level have led to changes in teacher behavior; however, teachers seem to respond narrowly in ways that improve measured outputs (i.e., the measures) with less attention to long-term outcomes (i.e., the goals).[1] While student test scores have risen, there is uncer-

[1] We use the following terminology when talking about public service programs and their consequences: a program is a structured activity that transforms *inputs* into *outputs*, which are observable, measurable (e.g., blood pressure, test scores, parts per million of carbon diox-

tainty as to whether these reflect more learning or are to some degree the product of teaching to the test or other approaches to generating apparent improvement.
- In health care, relatively small financial incentives (frequently combined with public reporting) have had some modest effects in improving the quality of care delivered.
- Examples from the transportation sector suggest that large financial incentives can lead to creative solutions, as well as to lobbying to influence the demands of the PBAS regulation. The latter has been the case with the CAFE standards, which require automobile manufacturers to achieve a minimum level of fuel economy for the fleet of vehicles sold each year in the United States.
- It is too soon to judge the effectiveness of PBASs in child care and PHEP.

PBASs also have the potential to cause unintended consequences by incentivizing the wrong kind of behavior or encouraging undesirable effects. For example, in NCLB, attaching public reporting and other incentives to test scores has led to unintended behavioral changes (i.e., teaching to the test) that might be considered undesirable. In the transportation sector, some analysts have argued that CAFE standards prompted auto manufacturers to produce smaller and lighter vehicles, which, in turn, increased the number of crash-related injuries and fatalities, though this conclusion remains subject to some debate. A concern in the health-care sector is that PBASs include a narrow set of performance markers, which might increase physicians' focus on what is measured and reduce their attention to unmeasured effects. However, to date, there is an absence of empirical evidence showing such effects.

---

ide, or CO_2), and easy to associate directly with the program. Ultimately, these outputs affect long-term *outcomes* that are of interest to policymakers (health, achievement, air quality). The outcomes might or might not be measurable, but it is typically difficult to draw a direct connection between the program and these outcomes. Many factors beyond the program's control or even understanding might affect the relationship between the program and the higher-level, broad outcomes relevant to policymakers. As a result, to influence behavior within a program with confidence, an accountability system must focus on measures of *outputs* that can be clearly attributed to the program.

If a PBAS does not initially meet its aims, it does not mean that a PBAS cannot be successful; it might just mean that some of the structural details require further refinement. PBASs are sufficiently complex that initial success is rare, and the need for modification should be anticipated.

Recommendations for System Developers

We offer a number of recommendations for PBAS sponsors, designers, and other stakeholders to consider regarding PBAS design, incentives and performance measurement, implementation, and evaluation.

Design of the Performance-Based Accountability System

Designing a PBAS is a complex undertaking, and many of the decisions that will need to be made are heavily dependent on sector-specific contextual circumstances.

Consider the Factors That Might Hinder or Support the Success of a PBAS to See Whether Conditions Support Its Use. The first issue is to consider whether a PBAS is the best policy approach for the policy concern at hand and whether it might be expected to succeed. From the cases examined, we identified a number of factors that tend to support a successful PBAS implementation:

- broad agreement on the nature of the problem
- broad agreement on PBAS goals
- knowledge that specific changes in inputs, structures, processes, or outputs will lead to improved outcomes
- ability of service providers, through changes in behavior, to exert significant influence on outputs and outcomes
- ability of the implementing organization to modify the incentive structure for service providers
- absence of competing programs that send conflicting signals to service providers
- political context in which it is acceptable for the PBAS to be gradually improved over time

- sufficient resources to create the PBAS and to respond to the incentives.

If a large share of these factors does not hold for the application under consideration, decisionmakers might wish to consider alternative policy options or think about ways to influence the context to create more-positive conditions for a PBAS.

Be Sensitive to the Context for Implementation. It is important to account for constraints and leverage opportunities presented by the context in which the PBAS will be implemented. Such considerations include the extent to which the implementing organization can alter the incentive structure faced by service providers, existing mechanisms that will affect the behavior of service providers (e.g., safety or licensing requirements) or that can be used to support the PBAS (e.g., data collection), and current knowledge of the service activity covered by the PBAS.

Consider Applying Performance Measures and Incentives at Different Functional Levels. If the service-delivery activities are organized hierarchically (e.g., students within classrooms within schools within districts), PBAS designers should consider the application of performance measures and incentives at different functional levels within the activity (e.g., different measures and incentives for school districts, school principals, and teachers or for hospitals, clinics, and doctors). Provided that the performance measures and incentives are structured in a complementary fashion, the results can be additive and mutually reinforcing.

Design the PBAS So That It Can Be Monitored Over Time. To obtain the best results over the long term, it is important to develop a plan for monitoring the PBAS, identifying shortcomings that might be limiting the effectiveness of the PBAS or leading to unintended consequences, and modifying the program as needed.

Incentives and Performance Measurement

The selection of incentives and performance measures is of vital importance to the PBAS. The type and magnitude of the incentives will govern the level of effort providers expend to influence the performance

measures, while the measures will dictate the things on which the service providers should focus and what they might choose to ignore or neglect.

Create an Incentive Structure Compatible with the Culture of the Service Activity. Many options for incentives are available, including cash, promotions, status, recognition, increased autonomy, or access to training or other investment resources. The goal is to select options that will best influence behavior without undermining intrinsic service motivation.

Make the Rewards or Penalties Big Enough to Matter. The size of the incentive should outweigh the effort required by the service provider to improve on the performance measure; otherwise, service providers will simply not make the effort. However, the size of the incentives should not exceed the value obtained from improved provider behavior, since the PBAS would, by definition, not be cost-effective.

Focus on Performance Measures That Matter. Performance measures determine how service providers focus their efforts. To the extent possible, therefore, it makes sense to include those measures believed to have the greatest effect on the broader goals of interest.

Create Measures That Treat Service Providers Fairly. In certain settings, the ability of service providers to influence desired outputs might be limited. When selecting performance measures, PBAS developers should consider the degree to which service providers can influence the criteria of interest. Individuals or organizations should not be held accountable for things they do not control. In such cases, there are other options for creating performance measures that treat service providers fairly:

- Create "risk-adjusted" output measures that account for relevant social, physical, and demographic characteristics of the population served.
- Establish measures based on inputs, structure, or processes rather than on outputs or outcomes.
- Measure relative improvement rather than absolute performance.

Avoid Measures That Focus on a Single Absolute Threshold Score. Although the threshold approach can be intuitively appealing (in the sense that the specified score represents a quality bar that all service providers should strive to achieve), in practice, measures that focus on a single threshold can prove quite problematic. Low achievers with no realistic prospects for achieving the absolute threshold score will have no incentive to seek even modest improvements, while high achievers will have no incentive to strive for further improvement. Alternatives include use of multithreshold scores and measurement of year-over-year improvement.

Implementation

It is possible to create a potentially effective design for a PBAS and then fail to implement the design successfully; thus, special attention needs to be paid to the way the PBAS is implemented.

Implement the Program in Stages. Because most PBASs are quite complex, it is often helpful to develop and introduce different components in sequence, modifying as needed in response to any issues that arise. For example, initial efforts and funding might focus on the development of capacity to measure and report performance, with measures and incentives rolled out over time. Pilot-testing might also be used to assess measures and other design features.

Integrate the PBAS with Existing Performance Databases and Accounting and Personnel Systems. A PBAS is not created in a void; rather, it must be incorporated within existing structures and systems. It is thus important to think through all of the ways in which the PBAS will need to interact with preexisting infrastructure—e.g., performance databases, accounting systems, and personnel systems. These considerations might suggest changes in the design of the PBAS or highlight ways in which the existing infrastructure needs to be modified while the PBAS is being created.

Engage Providers, and, to the Extent Possible, Secure Their Support. To garner the support of providers, it is helpful to develop measures that are credible (i.e., tied to outcomes about which they care), fair (i.e., that account for external circumstances beyond the control of providers), and actionable (i.e., that can be positively influenced

through appropriate actions by the service provider). A good approach is to involve providers in the process of developing the measures and incentives. While, to some degree, it can be expected that service providers will seek to weaken the targets or standards to their benefit, those responsible for implementing and overseeing the PBAS will need to judge whether lowering performance expectations would ultimately undermine the success of the PBAS.

Ensure That Providers and Other Stakeholders Understand Measures and Incentives. Communication is key. Particularly in cases in which there are numerous providers with varying internal support systems to enable engagement—as, for example, with health-care P4P systems and child-care quality ratings—it can be helpful to employ multiple communications channels (e.g., email, website, conference presentations) as appropriate.

Plan for the Likelihood That Certain Measures Will "Top Out." As service providers improve their performance in response to the incentive structure, a growing percentage might achieve the highest possible scores for certain measures. PBAS designers should plan for this eventuality, e.g., by replacing topped-out measures with more-challenging ones or by requiring service providers to maintain a high level of performance for topped-out measures in order to qualify for incentives.

Provide Resources to Support Provider Improvement. It can be valuable to devote program resources to support efforts at improvement. This might involve infrastructure investments or education for providers on becoming more effective.

Evaluation

Ironically, given the spirit of accountability embodied in the PBAS approach, most of the cases reviewed in this study have not been subjected to rigorous evaluation. We believe that it is vitally important to rectify this lack of evaluation. Only through careful monitoring and evaluation can decisionmakers detect problems and take steps to improve the functioning of the PBAS over time.

Consider Using a Third Party to Evaluate the PBAS. Not all organizations that implement a PBAS possess the necessary methodological expertise to conduct a sound programmatic evaluation. Additionally,

many implementing organizations, for understandable reasons, will tend to be biased in favor of positive results. For these reasons, it is beneficial to rely on an independent and qualified third party to conduct an evaluation of the PBAS.

Structure the Evaluation of a PBAS Based on Its Stage of Development. When a system is first developed, it might be most helpful to evaluate implementation activities (e.g., whether appropriate mechanisms for capturing and reporting performance measures have been developed). As the system matures, the focus should shift to evaluating the effects, in terms of observed provider behavior and service outputs, of the performance measures and incentive structure. An evaluation should focus on outputs only after performance measures and incentives have been in place long enough to influence behavior.

Examine the Effects of the PBAS on Both Procedures and Outputs. One approach for doing so is to develop a logic model, a visual representation of the ways in which the PBAS is intended to influence provider behavior. This model can then become the basis for thoughtful monitoring and evaluation and make it easier to plan the evaluation of a PBAS based on its stage of development.

Use the Strongest Possible Research Design Given the Context in Which the PBAS Exists. Options, sorted in order of decreasing rigor, include randomized control trials, regression discontinuity designs, nonequivalent-group designs, lagged implementation designs, and case studies. If certain design aspects are flexible, it might be possible to implement variations in the PBAS coupled with common evaluation frameworks to provide rigorous comparison and help choose the most effective options. Such variations could include different performance measures, different types of incentives, or different incentive levels (e.g., significant versus modest financial rewards).

Implement Additional, Nonincentivized Measures to Verify Improvement and Test for Unintended Consequences. A PBAS might induce service-provider responses that lead to improved performance scores without corresponding improvement in the underlying objectives (e.g., a teacher might invest instructional effort on test-taking strategies that lead to improvement on standardized test scores that overstates actual student gains in mastery of the broader subject matter). To

detect when this might be occurring, it can be helpful to include nonincentivized measures intended to test similar concepts (e.g., additional math and reading exams in alternative test formats to check whether there has been a comparable level of improvement). Nonincentivized measures can also be used to examine whether a focus on the incentivized measures within the PBAS is causing other areas of performance to be neglected.

Link the PBAS Evaluation to a Review and Redesign Process. The true benefits of evaluation come not from simply understanding what is working and what is not, but rather from applying that understanding to improve the functioning of the PBAS. Evaluation should thus be embedded within a broader framework for monitoring and continuing to refine the PBAS over time.

Areas for Further Research

Because so few of the PBASs that we examined have been subjected to rigorous testing and evaluation, there are a number of fundamental questions that our study cannot answer about the design, implementation, and performance of PBASs. Policymakers would benefit from research—both within individual sectors and across sectors—on the short- and long-term impact of PBASs, the elements of a PBAS that are most important in determining its effectiveness, and the cost and cost-effectiveness of PBASs, particularly in comparison to other policy approaches.

Concluding Thoughts

This study suggests that PBASs represent a promising policy option for improving the quality of service-delivery activities in many contexts. The evidence supports continued experimentation with and adoption of this approach in appropriate circumstances. Even so, the appropriate design for a PBAS and, ultimately, its prospects for success are highly dependent on the context in which it will operate. Because PBASs are

typically complex, getting all of the details right with the initial implementation is rare.

Ongoing system evaluation and monitoring should be viewed, to a far greater extent than in prior efforts, as an integral component of the PBAS. Evaluation and monitoring provide the necessary information to refine and improve the functioning of the system over time. Additionally, more-thorough evaluation and monitoring of PBASs will lead, gradually, to a richer evidence base that should help future decisionmakers understand (1) the circumstances under which a PBAS would be an effective and cost-effective policy instrument and (2) the most appropriate design features to employ when developing a PBAS for a given set of circumstances.

Acknowledgments

A number of individuals made important contributions to the work described in this monograph. Sylvia Montoya conducted a literature review that we used to provide background information and to frame the discussions of accountability in each sector. In addition, participants at a 2008 workshop reacted to our early thinking about PBASs, providing helpful, critical feedback and offering new perspectives on our work. These participants included Dominic Brewer, University of Southern California; Jon Christianson, University of Minnesota; Carl Cohn, Claremont Graduate University; William Gormley Jr., Georgetown University; LeRoy Graymer, University of California, Los Angeles (UCLA), Extension; Jayetta Hecker, U.S. Government Accountability Office; Joan Herman, UCLA; Sylvia Hysong, Baylor College of Medicine; Richard Little, University of Southern California; Betsey Lyman, California Department of Public Health; Kathy Malaske-Samu, Los Angeles County Office of Child Care; Meera Mani, Preschool California; Steven Pickrell, Cambridge Systematics; Nico Potterat, Los Angeles Health Care Plan; Gery Ryan, RAND Corporation; Linda Smith, National Association for Child Care Resource and Referral Agencies; Joan Sollenberger, California Department of Transportation; Donna Spiker, SRI International; Sam Stebbins, University of Pittsburgh; Mark Steinmeyer, Smith Richardson Foundation; Michael Stoto, Georgetown University; Fred Tempes, WestEd; and Craig Thomas, Centers for Disease Control and Prevention. Donna White expertly coordinated the workshop and provided invaluable administrative assistance on other aspects of this project.

The quality of this document was substantially improved through several stages of review. Our RAND colleague Richard Neu reviewed an early draft of this monograph and provided constructive suggestions that led to important changes. We are also grateful to William Gormley and Harry Hatry, who served as our formal peer reviewers, and to Cathy Stasz, RAND Education Quality Assurance Manager, for their thoughtful reviews.

Abbreviations

A+B	cost plus time
AHA	American Hospital Association
AMA	American Medical Association
AYP	adequate yearly progress
CAA	Clean Air Act of 1963
CAFE	Corporate Average Fuel Economy
CDA	child-development associate
CDC	Centers for Disease Control and Prevention
CMS	Centers for Medicare and Medicaid Services
DoD	U.S. Department of Defense
DOT	U.S. Department of Transportation
ECERS-R	Early Childhood Environment Rating Scale, Revised
EPA	U.S. Environmental Protection Agency
EPCA	Energy Policy and Conservation Act
FAH	Federation of American Hospitals
FHWA	Federal Highway Administration

GE	General Electric
GPRA	Government Performance and Results Act
HEDIS	Healthcare Effectiveness Data and Information Set
HHS	U.S. Department of Health and Human Services
HMO	health maintenance organization
IOM	Institute of Medicine
MCT	minimum competency testing
MIT	Massachusetts Institute of Technology
mpg	miles per gallon
NACCHO	National Association of County and City Health Officials
NAEYC	National Association for the Education of Young Children
NCLB	No Child Left Behind Act of 2001
NCQA	National Committee for Quality Assurance
NEPPS	National Environmental Performance Partnership System
NPR	National Performance Review
NTD	National Transit Database
OMB	Office of Management and Budget
P4P	pay for performance
P4R	pay for reporting
PAHPA	Pandemic and All-Hazards Preparedness Act
PAR	Performance and Accountability Report
PART	Program Assessment Rating Tool

PBAS	performance-based accountability system
PHEP	public health emergency preparedness
PPBS	Planning, Programming, and Budgeting System
QRIS	quality rating and improvement system
R&R	resource and referral agency
RCT	randomized control trial
RTPA	regional transportation planning agency
SBR	standards-based reform
SEP-14	Special Experimental Project 14
SIP	state implementation plan
SUV	sport-utility vehicle
TIMSS	Third International Mathematics and Science Study
TQM	Total Quality Management

CHAPTER ONE

Introduction

With much fanfare, the No Child Left Behind Act of 2001 (NCLB) (Pub. L. 107-110) initiated an era of *performance-based accountability* in federal education policy. In an unusual display of bipartisan support, Democrats (led by Sen. Edward Kennedy and Rep. George Miller) and Republicans (including President George W. Bush) carved out a strategy to hold schools accountable for the performance of their students. The legislation set the lofty goal of all students being proficient in reading and mathematics by 2014. States were required to adopt grade-level standards in reading and mathematics, test all students (grades 3–8 and one high-school grade) in these subjects, and impose a series of progressively stricter interventions if schools did not make sustained annual progress toward 100-percent proficiency.

An interest in improving the quality of American schools was not new. The 1983 publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983) documented the poor performance of U.S. students in English and mathematics. Concerns accelerated in the 1990s in the wake of the disappointing performance of U.S. students on the Third International Mathematics and Science Study (TIMSS); for example, in 1994–1995, American eighth-graders ranked 18th out of 25 participating countries in mathematics,[1] and the scores were no better in 1999. The poor performance of urban, minority, and limited-English-proficient students on state tests and the National Assessment of Educational Progress also caused alarm.

[1] U.S. students scored lower than Singapore and Japan and other Asian countries, but also lower than the Russian Federation, Ireland, and New Zealand (Beaton et al., 1996).

What was different with NCLB was its approach toward improving education: a combination of explicit expectations for student performance, well-aligned tests to measure achievement, and strong consequences for schools that do not meet program targets.[2] There had been some earlier state-level efforts that drew on some of the ideas that would be brought together in NCLB. For example, Kentucky and Maryland had implemented systems that required scores for every school to be published so the public could see how well students were doing and press for reforms where needed. Such public "shaming" was a modest incentive but one that proved to be a potent lever for change in at least some cases (Koretz et al., 1996). However, research suggests that reporting alone is often inadequate for producing widespread change (Hanushek and Raymond, 2005), and, perhaps in response to this perceived inadequacy, other states supplemented the reporting of scores by providing rewards to high-performing schools and sending support teams to assist low-performing schools. NCLB applied the concept of performance-based accountability on the national scale. A key component of the legislation was an emphasis on outputs relevant to ultimate outcomes rather than on the means used to achieve those outputs. Schools were required to meet the program's broad goal of achieving proficiency for all students in reading and mathematics by 2014. But it was left up to states to determine the specific goals for their students (curriculum standards), find ways to measure achievement of those goals (large-scale testing), set progressively higher targets for achievement, and intervene when schools failed to achieve their targets, to the point of reconstituting schools if necessary.

NCLB is an example of a phenomenon that is increasingly common in situations in which governments regulate the delivery of services, whether those services are delivered by the government or by nonprofit or for-profit organizations: the adoption of what we will call performance-based accountability systems (PBASs). As seen with NCLB, a PBAS typically has three main components: goals, incen-

[2] Standards and assessments had been introduced in the Improving America's Schools Act (Pub. L. 103-382, 1994), the previous reauthorization of the Elementary and Secondary Education Act (part of NCLB).

tives, and measures. For example, the Centers for Medicare and Medicaid Services (CMS) is seeking to improve the quality of health care delivered under Medicare by experimenting with systems that reward health-care providers on the basis of measurable outputs relevant to patients. Similarly, many state transportation agencies seek to reduce the time needed to complete road projects by offering bonuses to road construction and repair contractors who complete the project within a set time. Regulatory bodies in other sectors, such as child care and public health emergency preparedness (PHEP), have also adopted PBASs to improve services in these areas.

The development of these systems in the public sphere grew out of the application of quality-based assessment systems in the private sector, which seek to emphasize customer priorities and outcomes while reducing costs. Such systems are now applied in almost all types of commercial industries and corporate settings through such techniques as lean production[3] and Six Sigma.[4] The history of PBASs is explored in more detail in Chapter Two.

But, while the use of PBASs has spread in the public sector, little is known about whether such programs are having the desired effect. For example, the success of NCLB is still being debated. Some researchers find increasing test scores in many states, as well as evidence that the gaps between the performance levels of majority and minority students

---

[3] *Lean production* is a popular name for the Toyota Production System, which was publicly documented in Ohno and Bodek (1988). Womack, Jones, and Roos (1991) coined the term *lean production* when they documented a multiyear research effort at the Massachusetts Institute of Technology (MIT) to benchmark the Toyota Production System against other approaches to automobile design, production, and marketing in Japan, North America, and Europe. Womack and Jones (2003) uses examples from many industries to show how lean production can dramatically improve the performance of processes outside the automobile industry.

[4] Six Sigma emerged from quality-control efforts at Motorola during the 1980s. This work built directly on Crosby (1979). Motorola launched a formal, company-wide Six Sigma program in 1987 to reduce the incidence of defects on its assembly lines. This was documented in Harry (1988). Six Sigma quickly spread to General Electric (GE), Allied Signal, and Citicorp through peer-to-peer benchmarking. Welch (2001) documents the early experience of Six Sigma at GE, one of the principal early advocates. Pande and Neuman (2000) and Huesing (2008) take a broader perspective.

are declining (Center on Education Policy, 2009; Dee and Jacob, 2009; Neal and Schanzenbach, 2007). Others report problems, such as teachers teaching to the test (which might lead to score inflation), schools giving too much emphasis to reading and mathematics (which might reduce learning in other subjects), and the unfairness of using measures defined in terms of status rather than change in performance (Stecher, 2002; McCall, Kingsbury, and Olson, 2004). Given the size and scope of the educational enterprise in the United States, it might be too soon to fully evaluate NCLB's impact.

A broader problem is that there is currently little evidence concerning the effectiveness of PBASs, or the experiences of governments and agencies at the forefront of this trend. This monograph seeks to address the gap by examining several examples of PBASs, large and small, in a range of public service areas. This study examines nine PBASs, which are drawn from five public service sectors: child care, education, health care, PHEP, and transportation. The cases we studied provide useful information on PBAS formation, design, operation, and evaluation and should be of interest to government officials who are considering this approach for regulating public services and to policy and agency staff charged with implementing such programs.

In order to understand PBASs in detail, we developed a framework for examining the components of a PBAS. We describe this framework next, though with a brief description of our methods and of the sectors examined.

What Is a Performance-Based Accountability System?

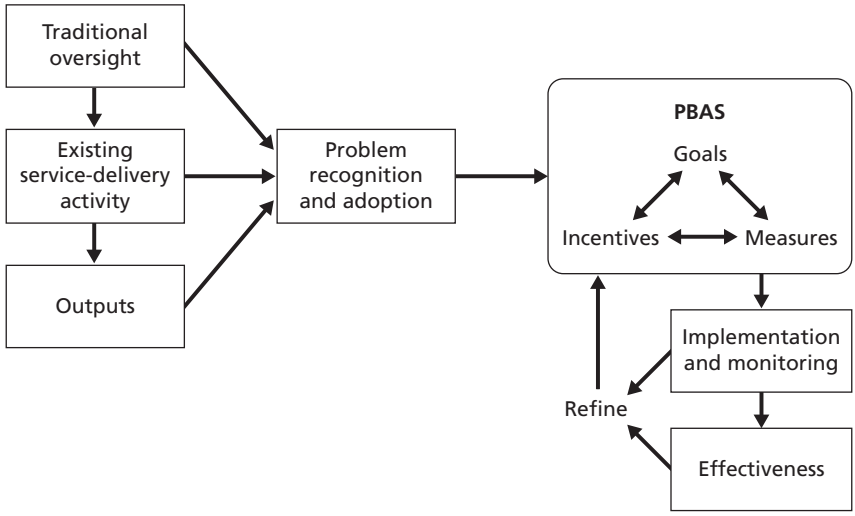
A PBAS provides a way to refocus resources (both human and financial) to achieve long-term performance goals. A PBAS is a mechanism for motivating and managing change. Typically, a PBAS has three main components: goals (i.e., one or more long-term outcomes to be achieved), incentives (i.e., rewards or sanctions to motivate changes in individual or organizational behavior to improve performance), and measures (formal mechanisms for monitoring the delivery of services or the attainment of goals).

The introduction and operation of a PBAS is illustrated in Figure 1.1. This framework presupposes an existing service-delivery activity provided by public or private organizations and regulated directly or indirectly by a government entity. For example, a state department of education is responsible for ensuring the quality of education provided in the public schools; a child-welfare agency is responsible for ensuring the safety of three- and four-year-old children receiving child care in public or private settings. Either of these agencies might adopt a PBAS as a strategy to achieve its long-term goals.

Problem Recognition and Adoption

The PBAS typically arises out of the recognition of some sort of problem or perceived deficiency in services or outcomes. If stakeholders perceive that there is a gap between the outputs that are produced and the goals they value, they might demand reform, which might take the form of a PBAS. The organizational and political context in which the service-delivery activity is embedded plays a role in shaping the eventual form and operation of the PBAS.

Figure 1.1
Evolution of a Performance-Based Accountability System



RAND *MG1019-1.1*

Design of the Performance-Based Accounting System

The recognition of a problem in an existing service-delivery activity, which is depicted in the center of Figure 1.1, leads to the design of the PBAS. The creation of the PBAS involves decisions about three interrelated components: goals, incentives, and measures.

- First, policymakers must agree on a set of *goals* or desired long-term outcomes for the service-delivery activity; these are usually expressed in general, nonquantified terms (e.g., world-class achievement, efficient public transportation, high-quality child care). These goals define what the service-delivery activity is supposed to achieve under the new regime of the PBAS.
- The second piece of the PBAS is an *incentive* structure that assigns rewards or sanctions (or some combination thereof) to individuals or organizations to try to motivate changes in their behavior. The incentives need not be financial; we include in our definition nonfinancial consequences that might motivate changes in provider behavior, such as greater autonomy, loss of control, or public reporting.
- The third element of the PBAS is a set of *measures* that can be used as the basis for applying the incentives to the people and the units that deliver the services. The designers of a PBAS must choose a way to define performance in order to implement the incentive structure and encourage better performance on the part of service providers.

Implementation and Monitoring

After a PBAS is designed and before it is fully operational, there is a period of implementation, which is depicted at the lower right of the figure. During this period, data collection and reporting systems are established and structures are put in place to judge whether targets are met, identify those who warrant incentives, and enact the incentives. In addition, mechanisms to audit the operation of the PBAS might be put in place. These monitoring functions might include secondary data collection, such as participants' knowledge of system rules and opinions about system fairness, and independent auditing of operations.

Effectiveness

The bottom right of Figure 1.1 illustrates efforts to judge the impact of the PBAS on the service-delivery activity. If designed and implemented appropriately, the PBAS encourages those delivering the service to take actions that improve measured outputs in the short term and promote desired outcomes in the long term.[5] Evaluation provides evidence about the effectiveness of the PBAS.

Refinement

Finally, evidence about PBAS operation and effectiveness, as well as implementation, can be used by the designers of the system to improve its performance. Such refinements might entail changes in the goals, design of incentives, or quantification of metrics.

A Note About Terminology

Despite broad interest in making government activities more accountable, there does not appear to be any general agreement on what accountability is. Some approach the idea in terms of the creation of accounts—in effect, the creation of performance measurements and the placement of them in easily accessible, transparent accounts. From this perspective, accountability ensures transparency (i.e., clarity and openness) about what government agencies are actually doing for citizens. Once such accounts exist, citizens are free to use them as they please.

A narrower interpretation views accountability as implying consequences for performance. When an activity provider is held accountable for its performance, the activity presumably suffers if the provider does not perform well and thrives if it does. We prefer this narrower interpretation. No matter how transparent and accessible a set of accounts

---

[5] Standard approaches to program evaluation envision program processes that transform inputs into outputs, which, in turn, ultimately affect the program outcomes that really interest policymakers. Our application of this approach accepts that the *outcomes* of a program can almost never be directly observed and measured. As a result, any accountability system must rely on measures of program *outputs* that can be measured. Throughout, we use this distinction to draw a line between *outputs* and *outcomes* when outcomes relevant to policymakers cannot be directly observed and measured.

is, an activity provider cannot be said to be held accountable for its performance unless it faces consequences for that performance—hence our focus on performance-measurement systems with incentives that explicitly link the performance of a service activity to consequences for the service activity, its managers, and its employees. In this report, a PBAS, by definition, includes an incentive structure to translate performance measures into consequences in a predictable way.[6] Incentives might be direct, such as cash bonuses or interventions that mandate changes in practice, or they might be indirect, such as public reporting designed to influence user choices.

In our discussion of PBASs, we focus on four broad categories of individuals and organizations that are involved in the creation and operation of the systems.

- *Sponsor* refers to the organization, administrative agency, or other policymaker that governs some activity and decides to enact the PBAS. For NCLB and the Clean Air Act (CAA), the key sponsor was the U.S. Congress, with the assistance of the U.S. Department of Education and the U.S. Environmental Protection Agency (EPA).
- *Designers* are those individuals or organizations that develop and implement the PBAS in accordance with the broad goals established by the stakeholders.
- A *service provider* is the organization that delivers the service the stakeholders are hoping to improve. In health care, the service providers are medical groups, physicians, hospitals, and home health providers; in transit, they are bus companies and other transit operators.
- *Users* or *consumers* are the individuals who receive the services.

When talking about issues related to PBASs, many researchers use the terms *principal* and *agent*. An *agent* is a party that acts on behalf of a *principal*; the *principal* wants something done and turns to an *agent*

[6] For a useful discussion of this perspective, see Artley (2001, pp. 5–8).

to do it rather than doing it itself.[7] Sponsors might be referred to as *principals* and service providers as *agents*.

Research Approach

To examine PBASs, we selected for study a heterogeneous set of service areas: child care, education, health care, PHEP, and transportation.[8] The research approach consisted of the following components:

- a broad review of literature related to performance measurement and accountability
- development of an analytic framework (Figure 1.1), which was used to structure our internal discussions about research evidence in the five sectors
- a 1.5-day integrative workshop at the RAND offices in Santa Monica, California; the workshop examined various features associated with PBASs, including the context in which the PBAS arose, measures, incentives, and evaluation approaches. In addition to project staff, approximately 30 researchers, policymakers, and agency administrators drawn from the five sectors attended the workshop.
- analysis of sector-specific empirical results and identifying cross-sector principles. Through this process, we attempted to derive principles that held in the five sectors and might have general applicability beyond them.

---

[7] According to the *Restatement of the Law of Agency*,

> Agency is the fiduciary relationship that arises when one person (a "principal") manifests assent to another person (an "agent") that the agent shall act on the principal's behalf and subject to the principal's control, and the agent manifests or otherwise consents so to act. (American Law Institute, 2006, ¶1.01)

[8] We examined a single type of PBAS in child care, health care, and emergency preparedness. In education, we examined two different approaches to PBAS, and, in transportation, four distinct PBASs.

Cases Examined in This Study

We chose to focus our inquiry on work in which RAND was already engaged. This focus enabled us to exploit our existing knowledge and build an in-house team that could work together to discuss key issues and further the work.

We selected for study nine cases drawn from five policy areas or sectors: child care, education, health care, PHEP, and transportation. The choice of cases was guided by practical as well as theoretical considerations. On the practical side, we wanted to take advantage of the expertise available at RAND, where empirical research is being conducted on a number of performance measurement and accountability systems in different service areas. On the theoretical side, we wanted to include cases in which services are provided primarily by public agencies (education, transportation), as well as sectors in which services are provided primarily by private organizations but where the public sector has an important role in governance (child care, health care). We also wanted to include at least one instance in which measurement itself was a challenge (PHEP). Our research team also included one member with extensive experience studying PBASs in the context of defense contracting.

Table 1.1 summarizes by case the key features of the PBASs examined in this study. For a fuller description of each case, please see Appendix A.

Child Care

Child care is funded and delivered by public agencies at all levels of government and by a variety of private organizations and occurs in a range of settings, including free-standing centers, public school campuses, churches, community centers, and family homes. Until quite recently, quality standards in child care were largely limited to state licensing requirements, such as the adequacy and safety of a program's physical environment. However, the growing policy attention focused on K–12 accountability has raised questions about how well child-care programs are preparing children for school, leading to efforts to improve quality. Quality rating and improvement systems (QRISs) represent the

most popular current approach to doing so and are the key PBASs in the sector. QRISs focus primarily on assessing and improving program inputs and processes, including child-staff ratios, group size, staff education and training, and indicators of the classroom environment. QRISs produce a single, easy-to-understand rating for each provider, much like restaurant ratings in some cities; however, participation in QRISs is voluntary.

Education

Public education in the United States is primarily the responsibility of state and local governments, with the federal government playing a relatively small role focused on equity (e.g., supplemental funding to address the challenges of poverty, disabilities). Despite a strong tradition of local control, standardized measures of outputs, such as tests, have been used in education for decades to monitor school performance. In the 1990s, some states began to adopt "standards-based" accountability systems, which included standards, assessments, publication of scores, and, in some cases, rewards and sanctions. In 2001, the idea of standards-based reform was incorporated into federal legislation (NCLB), which is essentially a PBAS with schools and districts as the unit of accountability. NCLB requires all states to create accountability systems that include state standards in reading, mathematics, and science and annual testing of students in selected grades in these subjects and is thus the predominant PBAS in this sector.

Pay for performance (P4P) is also being adopted in some districts and states. P4P systems in education usually pay bonuses directly to teachers or principals for meeting specific performance criteria, usually in terms of student achievement but, in some cases, including other outputs relevant to students (such as graduation) or educators' practices.

Health Care

Health care in the United States is primarily provided by the private sector, and the providers are reimbursed for their services in large measure by private payers (60 percent compared with 25 percent by

Table 1.1
Performance-Based Accountability System Cases, by Sector, Examined in This Study

Sector	PBAS	Goal(s)	Key Incentives	Measures
Child care	QRISs	Improve quality of child care	Prestige associated with high rating Financial incentives	Measures of child-staff ratios, group size, staff education and training, etc.
Education	NCLB	Achieve academic proficiency, graduation	Graduated set of interventions regarding professional development, instruction, staffing, school governance (e.g., constraints on use of funds)	Reading and mathematics test results
	P4P	Improve student achievement, other student outcomes, and educator practices	Cash bonuses, salary increases	Value-added scores based on standardized achievement tests
Health care	Hospital and physician/medical group P4P programs	Improve quality of health care	Financial payments for high performance or improvement	Clinical quality measures (e.g., proportion of patients in a certain risk group who receive a specific type of evidence-based care)
	Quality report cards	Reduce resource use and overuse of services	Public recognition, transparency (i.e., clarity and openness) of performance results	Patient experience with receiving care Resource use or cost-efficiency measures

Table 1.1—Continued

Sector	PBAS	Goal(s)	Key Incentives	Measures
PHEP	CDC PHEP cooperative agreement	Ensure state and local jurisdictions' preparedness against large-scale public health emergencies	Withholding of federal funds for failure to meet performance benchmarks	Measures for medical countermeasure delivery, other aspects of PHEP
Transportation	A+B highway construction contracting	Reduce time to complete road projects	Financial rewards or sanctions based on time to complete	Time required to complete construction
	CAFE standards	Foster energy independence and (more recently) reduce greenhouse-gas emissions through improved fuel economy	Fines for failure to meet minimum average fuel-efficiency standards	Sales-weighted average fuel economy for passenger cars and light trucks within each model year
	CAA ambient air pollution conformity requirements	Ensure that regional transportation plans support improved air quality	Federal transportation funds subject to conformity with ambient air quality standards	Forecasted emissions of EPA-regulated "criteria" pollutants (e.g., carbon monoxide, ozone) based on planned transportation investments
	Transit subsidy allocation formulas	Improve performance and efficiency of transit service	Share of state or regional funding for local transit operators	Allocation based on formula combining both nonperformance (e.g., population size) and performance (e.g., total passengers served) measures

NOTE: CDC = Centers for Disease Control and Prevention. A+B = cost plus time. CAFE = Corporate Average Fuel Economy.

public health insurance, i.e., Medicare and Medicaid).[9] However, CMS remains the dominant purchaser in the market for health-care services because use of health services is disproportionately greater among the senior population who are insured through Medicaid and Medicare. The measures typically used in health-care PBAS programs focus on quality, although, recently, an increasing number of PBASs are evolving to include measures of resource use or cost-efficiency. As of 2009, there were more than 100 P4P programs focused on physicians and more than 40 such programs focused on hospitals. All these programs were initiated by the private sector. The National Committee for Quality Assurance (NCQA), through initial funding from employers and CMS in the late 1980s and early 1990s, developed the Healthcare Effectiveness Data and Information Set (HEDIS) to provide quality information about plans to employers. HEDIS measurement and public accountability provided the initial foundation for many of today's PBASs in health care, and, today, many health plans, employers, consumer advocacy groups, and government agencies publish quality report cards to assist individuals in choosing their health-care providers. CMS's current "pay-for-reporting" programs for hospitals and physicians and various pilot demonstrations of P4P are setting the stage for performance-based accountability within Medicare. At present, CMS is operating P4P demonstrations that target hospitals, physician group practices, end-stage renal disease facilities, nursing homes, and home health workers.

Public Health Emergency Preparedness

PHEP involves efforts to prevent, protect against, quickly respond to, and recover from large-scale health emergencies, including bioterrorism, naturally occurring disease outbreaks, and natural disasters. Primary legal responsibility for PHEP (as with other aspects of public health) lies with state health departments, which delegate varying degrees of responsibility to local health departments. While PHEP efforts are typically coordinated and led by governmental public health agencies,

---

[9] Roughly 15 percent have no health insurance coverage and typically rely on emergency services.

PHEP also requires active involvement by health-care systems, emergency management, law enforcement, communities, and individuals. During the late 1990s, increasing concern about the threat of weapons of mass destruction led to a small federal effort to build state and local ability to prepare for large-scale public health emergencies. These efforts grew substantially after the terrorist attacks of September 11, 2001, and the anthrax attacks of October 2001. The two most important federal PHEP programs focus on hospital preparedness (the U.S. Department of Health and Human Services, or HHS, Hospital Preparedness Program) and all-hazards public health preparedness (CDC's PHEP cooperative agreement). The PHEP cooperative agreement, which is the case on which we focus in this study, requires grantees (including the 50 states, four separately funded large cities, and eight territories) to report data on performance metrics for two program areas: (1) mass medical countermeasure delivery and (2) all other aspects of PHEP (e.g., epidemiological investigation, surveillance, incident management, crisis and emergency risk communication). The Pandemic and All Hazards Preparedness Act of 2006 (PAHPA) (Pub. L. 109-417, 2006) required that the Centers for Disease Control and Prevention (CDC), as of 2009, withhold federal funding for failure to meet performance benchmarks.

Transportation

PBASs are of growing interest in the field of transportation planning and policy. Federal legislation is expected to be enacted in the near future that will renew the national transportation funding program for the following six years. Many politicians, interest groups, and scholars are advocating that familiar formulas by which federal funds are distributed to states for particular programs should be replaced by funding arrangements that are performance-based (National Transportation Policy Project, 2009). Thus far, many state and federal programs purport to measure and report on the performance of the transportation system, but relatively few include any sort of accountability requirements. Current debates suggest that funding should, in the future, be tied more directly to measures of the attainment of major programmatic objectives, such as improved mobility, increased acces-

sibility, and congestion reduction, yet it is not clear that consensus can be reached on approaches by which to measure the attainment of these objectives.

We were unable to find for this study transportation programs that incorporated PBASs, but several specific PBASs were identified within or related to transportation. These were in general narrower in scope than some of the pending proposals. For example, we included in this study examination of road construction contracts that provide bonuses for early completion of a road project and penalties for late project delivery (A+B contracting). Also included was the CAFE program of the federal government, which, in pursuit of environmental goals, financially penalizes automobile manufacturers that fail to achieve improvements in fuel economy. A third example is provided by an attempt to financially reward public transit systems that increase their daily patronage in relation to peer organization. Finally, we looked at penalties imposed on regions under the CAA amendments when their regional transportation plans failed to result in specified targets for the reduction of air pollutant emissions. While these four examples cannot, even when taken together, suggest how a more integrated PBAS might work in the field of transportation, they provide many lessons that should influence the design of such a system over the coming few years.

Organization of This Monograph

The remainder of this monograph is divided into nine chapters:

- Chapter Two provides background information on the history of PBASs for readers who would like to know more about the evolution of these systems and their migration from the private to the public sector.
- Chapter Three describes the events and recognition of problems motivating the development of PBASs examined in this project.
- Chapter Four describes the design of PBAS incentive structures to motivate behavioral change.

- Chapter Five discusses the design of measures that link performance measurement to an incentive structure.
- Chapter Six describes the implementation and monitoring of PBASs, including common pitfalls to implementing PBASs across sectors, as well as strategies for addressing these challenges.
- Chapter Seven examines what is known about the effectiveness of PBASs in improving services and whether those enhanced services have led to meeting desired system goals.
- Chapter Eight focuses on the evaluation of PBASs; it explores why evaluation was uncommon in the sectors we examined, offers justifications for conducting evaluation of these systems, and suggests key features to be included in PBAS evaluations.
- Chapter Nine provides concluding observations on the design and implementation of PBASs, as well as recommendations to assist policymakers in considering, designing, and implementing a PBAS.
- Appendix A provides a brief description of each of the five sectors and the nine related PBASs covered in this monograph. Appendix B describes research designs that can be used to evaluate the impact of PBASs.

CHAPTER TWO

A Historical Perspective on Performance-Based Accountability Systems

Today's PBASs grew out of efforts over many years and many countries to manage the private and public organizations that were growing too large to be overseen by a single manager who knew what everyone was doing. The innovative approaches that were tried focused on measuring performance, which was originally defined fairly narrowly. As these efforts developed, notions about what aspects of performance most mattered broadened and changed over time. By the 1980s, government organizations were linking performance to incentives in an effort to motivate and direct individual performance and improve organizational outcomes. The PBASs on which this book focuses represent some of these new ways of incentivizing performance in the delivery of public services. Understanding this history helps to clarify where the PBASs that we emphasize fit in the broader efforts to improve the performance of government activities and might suggest future directions as well.[1]

Origins

Scholars point to two events in the second half of the 19th century that might be thought of as the earliest examples of scientific management. At the time, privately owned railroads were the largest enterprises in

[1] William Gormley and Harry Hatry were especially helpful in advising us on how to sort through the complex history of performance management in American governments, especially from the 1960s on.

the United States, far larger than any government activity. By 1855, Daniel C. McCallum, head of the Erie Railroad, realized that he had no way of knowing which of the increasing numbers of managers he was overseeing were doing a good job, and which were not. He wanted a way to determine which of "the officers . . . conduct their business with the greatest economy" and provide him a way to "indicate, in a manner not to be mistaken, the relative ability and fitness of each for the position he occupies" (Chandler, 1977, p. 115). By 1875, the Louisville and Nashville Railroad (L&N) had developed a summary metric of this kind that it built up from 68 distinct sources of data (Chandler, 1977, p. 116).

At about that time in England, an effort was under way to improve conditions in English hospitals. This effort was led by Florence Nightingale, a nurse who was intent on reducing hospital mortality rates due to unsanitary conditions, and William Farr, a physician interested in social reforms. In 1863, Nightingale published *Notes on Hospitals*, a public report card that compared hospital mortality rates (Iezzoni, 1996; Gormley and Weimer, 1999). Despite its methodological flaws, Nightingale's report publicly revealed that some hospitals had very poor performance, as reflected by high mortality rates. This early public accountability effort was an important impetus for reforms that were subsequently undertaken to improve conditions in hospitals—and for measurement and accountability efforts that routinely occur in health-care settings today.

A generation later, the Progressive reform movement in the United States began to bring performance measurement into American government activities. Three men with expertise in social statistics, municipal accounting, and Frederick Winslow Taylor's[2] scientific management—William H. Allen, Henry Breure, and Frederick A. Cleveland—founded the New York City Bureau of Municipal Research in 1907

[2] Frederick Winslow Taylor (1856–1915) was an American mechanical engineer who sought to improve industrial efficiency. He is regarded as the father of scientific management.

to help improve the accountability of the newly empowered executive branch of the city government:[3]

> The role of performance measurement was to reconcile administrative discretion with accountability. Unelected administrators were to be given sufficient discretion to accomplish their assigned goals, but in return, they would be monitored by elected representatives of the public . . . to verify that the will of the legislature or the elected executive was being followed. (Williams, 2003, p. 651).

The bureau shifted the monitoring of government agencies from a more-traditional emphasis on money management and ad hoc collection of relatively vague data to systematic and precise measurement of the "costs and results" of municipal functions:

> In addition to financial records, there was a focus on work records and records of outputs, outcomes, social indicators, and needs. . . . Such reports should assign costs narrowly to functions so that they could be comparable across time, jurisdictions, organizations, and work units. . . . These data were used to communicate accountability information to office holders and the public, and to inform budget making and productivity improvement. Through measurement, the bureau sought to substitute facts for vague impressions. Facts served as a check against error when holding administrators accountable, allocating resources, or improving work processes. (Williams, 2003, pp. 646–647, 649)

The perceived success of this approach led to the creation of a similar bureau in Philadelphia in 1908 and comparable bureaus in 13 other cities by 1915. The bureau itself ultimately became a training center, which prepared personnel with similar skills for work in many other jurisdictions.

---

[3] This discussion of performance measurement in New York City is based on Williams (2003).

Twentieth-Century Efforts to Improve Efficiency and Performance

These early efforts were followed by a series of government reforms in the 20th century, all designed to improve efficiency and performance and all contributing to today's emphasis on performance measurement and management. Like the earlier efforts instituted by the railroads, 20th-century reforms began in private manufacturing firms, which, in the 1920s, adapted industrial engineering methods to management problems, increasing the role of precise performance metrics in management. Such scientific management came to the government in the 1930s. The establishment of clear lines of hierarchical authority and the pursuit of efficiency through consolidation, coordination, and the elimination of redundancy and overlap, were heralded as the answer to government inefficiencies. This interest in scientific methods grew during World War II, when improving performance became a patriotic task, and persisted well into the 1950s.[4]

In the 1950s, the focus of government agencies changed. While they had earlier worked to identify and remove systemic problems in pursuit of new, more-efficient production processes, the approach now took on a more indignant, muckraking tone, with management hunting for "waste, fraud, and abuse," wherever it might lie. The goal was not simply to identify and remove impediments but to identify errant and unproductive individuals and shame or punish them. Where the earlier approach had invited steady monitoring of a standard set of performance measures, this new approach looked for "wrongdoing" wherever it might arise, inviting a view of performance measurement that felt more like a forensic or investigative activity than an engineering activity.

Two other developments in the 1950s set the stage for PBASs. First, system analysis was developed at the RAND Corporation. This new analytic approach created performance metrics precisely crafted to address specific public policy questions, first in the defense policy domain (Hitch and McKean, 1960) and then, as the concept of the

[4] This discussion draws on Light (1997).

Great Society came to the fore in the 1960s, in domestic policy areas.[5] RAND applied this approach to ongoing management of entire government agencies and created a set of accounts called a Planning, Programming, and Budgeting System (PPBS) (Novick, 1965). The U.S. Department of Defense (DoD) currently uses a descendant of this early system to manage its resources. In its effort to link resources to the ultimate defense capabilities that DoD produced, PPBS can be seen as a precursor of more-recent efforts to link budgets to output-oriented performance goals based on agency capabilities.

Second, formal program evaluation became an important new activity in nondefense federal agencies. Big government programs began to include set-aside funds for program evaluation. The Urban Institute became an early leader of efforts to measure performance in program evaluations at all levels of government. For example, in cooperation with the International City Management Association, the Urban Institute produced the first edition of a report now issued annually comparing the performance of about 200 local governments.[6] As program evaluation came of age as a formal discipline, logic modeling—the creation of a systematic representation of the relationships among a program's resources, planned activities, outputs, outcomes, and impacts—became a major policy-analysis tool. State and local governments have pioneered significant efforts to use systematic performance measurement to improvement the performance of government activities, setting the stage for similar federal government initiatives during the 1990s. For example, efforts in Texas had a strong effect on the federal National Performance Review (NPR) initiated by the Clinton administration.

Total Quality Management

By the final quarter of the 20th century, a new approach to managing large organizations was coming of age in the private sector, the product

[5] This discussion draws on Hatry (1999).

[6] The most recent edition of this report is Hatry et al. (2006).

of Japan's efforts to recover from its defeat in World War II. In the 1940s, Americans considered Japanese products to be cheap and shoddy. Japanese industrial leaders recognized this problem and decided to change their image. They invited several American quality leaders to Japan to ask for help in achieving these goals. The invited experts, including W. Edwards Deming, Joseph M. Juran, and Armand V. Feigenbaum, came up with an approach, referred to as Total Quality Management (TQM), that grew out of quality control but went much further: It encompassed all levels of a manufacturing organization. Kaoru Ishikawa also discussed "total quality control" in Japan, which, according to his explanation, means "company-wide quality control" that involves all employees, from top management to the workers.

At its core, TQM is a management approach to long-term success through customer satisfaction. In a TQM effort, all members of an organization participate in improving processes, products, services, and the culture in which they work. While much attention is focused on improving products, TQM goes beyond the exclusive product focus of quality-control efforts, which was introduced in the 1920s, to detect and fix problems along the production line and thereby avoid the manufacture of faulty products. TQM focused on traditional quality control but also on ways to motivate employees to perform better and help them develop the capacity to do so through education and self-improvement. It includes, along with a customer focus, a commitment to continuous improvement and the integration of quality improvement into the total organization. A signal example of this new approach was the establishment of quality circles, groups of workers who volunteer to meet and discuss ways to improve all aspects of their workplace; their ideas are shared with management.

TQM became widespread in Japan and contributed to a very different perception of Japanese products; American corporations took notice that Japanese market share was improving and that Japanese products were of higher quality. By the 1970s, American corporations began to consider adopting the concept into their management approaches. These considerations fostered a great deal of debate concerning whether cooperative and communal Japanese business practices, however effective they appeared to be in turning around the quality

and success of Japanese products, find traction in more-individualistic American enterprises. A small number of firms, including Xerox and Motorola, implemented effective versions of TQM during the 1980s. But it was not until the 1990s, when academic studies had fully dissected TQM, that American firms understood that they could successfully emulate such approaches, such as the Toyota Production System, in their own settings.

A few governments in the United States began to experiment with TQM at the same time. By 1990, the state governments of California, Florida, Texas, and Wisconsin already had a variety of TQM programs in place.[7] By 1990, the federal government had set up the Federal Quality Institute to train federal employees on TQM methods.[8] Early experience quickly made it clear that, to succeed, government users would have to adapt TQM to the environment in which they operated.[9] Today, the Baldrige National Quality Program issues the Malcolm Baldrige National Quality Award, the federal government's primary recognition for organizations that implement TQM effectively, and maintains criteria specially framed for government organizations.[10] Federal, state, and local government agencies today make extensive use of current variants of TQM, lean production, and Six Sigma.

Other Public-Sector Performance Management Initiatives

Meanwhile, American policymakers increasingly watched the United Kingdom, where Margaret Thatcher was working to implement her smaller-government principles by transferring many government activities to the private sector and attempting to improve the performance

---

7 For more information, see Carr and Littman (1990).

8 For information on early federal experience, see Keehley (1990).

9 For an early discussion of the challenges involved, see Swiss (1992).

10 For a useful overview of ten early government experiences with TQM (primarily at the federal level), see Cohen and Brand (1993).

of what remained in the public sector.[11] As the approach developed, it increasingly distinguished *what* the government expected from its agencies (what outcomes the government's customers, its citizens, want) from *how* they would provide it (what the government and its contractors would do to meet those demands). The approach gave agencies more discretion to choose *how* to pursue their responsibilities in exchange for accepting more accountability for *what* they delivered. Its efforts brought the British government to an understanding of performance-based management quite compatible with the central principles of TQM. But this understanding was now rooted firmly in the public sector.

The British government experience produced two important new insights. First, the British successes provided a proof of principle that successful private-sector methods could have broad, similarly positive effects in the public sector. In particular, treating taxpayers as customers as well as voters is a reasonable and doable government task. By viewing taxpayers as customers as well, the agency necessarily focuses outside itself, on its customers, and is more responsive to its customers' priorities. Second, if a primary goal of a government agency is to serve customers, then perhaps the government budgeting system should reward those agencies that do a better job of serving them. This second idea provided the basis for performance budgeting, in which each agency's budget in the future depends directly on its performance in the past.

The United States was watching Thatcher's efforts with interest, but it was Australia and New Zealand that implemented them, often more aggressively than Thatcher did herself. By the 1980s and 1990s, applications of performance measurement and performance budgeting had emerged in Australia and especially New Zealand, where, to date, they have been implemented more fully than anywhere else.[12]

---

[11] For a useful discussion of the factors that led to reform in the United Kingdom, the institutional setting within which it began, and the way it unfolded from 1979 into the 1990s, see Ferlie (1998). This discussion has direct implications for the expansion of performance-based management in other settings.

[12] For a comparison of approaches in these countries, at the national and provincial level, see Martin and Jobin (2004). New Zealand was among the first to adopt this result-oriented

This performance-centric approach has come to be known as the new public management.[13]

Government Performance and Results Act

In the United States, the federal government responded as well, but more slowly. Interest in and commitment to these new approaches in Washington reflected strong interest by cities and states, which implemented a range of innovative management approaches. By the early 1990s, there was strong bipartisan support for these ideas, leading to the passage of the Government Performance and Results Act (GPRA) (Pub. L. 103-62) in 1993.

GPRA codified, in statute, what it meant to pursue performance-based management in a federal agency. Among other things, GPRA required every federal agency to develop a strategic plan that included a mission statement, goals, and objectives, "including outcome-related goals and objectives, . . . [and] a description of the program evaluations used in establishing or revising general goals and objectives." It also required every agency to prepare an annual performance plan that establishes performance goals and expresses those goals in an "objective, quantifiable, and measurable form," establishes performance indicators, and provides a basis for comparing actual program results with the performance goals established. Agencies were also mandated to report yearly on program performance for the previous fiscal year and evaluate the performance plan for the current fiscal year in light of the previous year's performance. Interestingly, given the importance of

budgeting and management approach in the late 1980s. In addition to the countries mentioned in the text, Denmark, Finland, the Netherlands, Sweden, and the United States followed in the 1990s. Austria, Germany, and Switzerland launched similar moves a bit later (see Curristine, 2005).

[13] Because *new public management* means different things to different people, some doubt the usefulness of the term. We use it here as a broad catchall for the range of efforts that emerged in the 1980s and 1990s to make government agencies more responsive to what their constituents wanted. These initiatives have tended to share important attributes of TQM, particularly the focus on aligning all of an organization's efforts to a single purpose—serving the customer.

state and local governments to the effective implementation of many federal policies, GPRA was mute concerning the roles that other levels of government might play in a new approach to performance measurement and management.[14]

A George Mason University team is conducting an ongoing assessment of the annual Performance and Accountability Reports (PARs) submitted by the 24 agencies covered by the Chief Financial Officers Act (Pub. L. 101-576, 1990), which together account for 99 percent of federal spending (Feller, 2002, p. 444).[15] The assessment reveals mixed GPRA implementation success to date, in part because the metrics used by the George Mason University evaluation team are themselves highly aggregated and abstracted. For example, the U.S. Department of Transportation (DOT) was given the highest score of all 24 agencies on "forward looking leadership" for providing an informative explanation for each missed target and setting challenging targets for each individual measure; presumably, a high score could be achieved the following year as well for more missed targets and challenging future ones. Many in Washington view GPRA as a burden with little payoff. Critics note, for example, that many agencies have created an office to generate performance reports for GPRA that has little connection to the offices that make policy decisions.

Impatience with the pace of GPRA implementation, and particularly of efforts to link performance data and budget decisions, led the Office of Management and Budget (OMB) in 2002 to develop a mechanism called the Program Assessment Rating Tool (PART) to help budget examiners and federal managers measure the effectiveness of government programs.[16] PART uses simple questionnaires to assess government programs each year in four areas: program purpose and design, strategic planning, program management, and program results. PART uses weights to aggregate raw scores in these four areas into a single numerical rating and then uses this rating to determine whether

[14] This discussion of federalism is based on Radin (2003).

[15] This effort reviews the PAR documents but does not examine how these documents affect decisionmaking or actual performance in the agencies that produce them.

[16] The discussion of PART draws on Hatry (2008) and Hughes and Shull (2005).

performance-based management in the program should be assessed as "effective," "moderately effective," "adequate," "ineffective," or "results not demonstrated."

While PART received a good deal more attention in the George W. Bush administration than did the reports associated with GPRA, econometric analysis of the effect of PART scores on the program budgets that OMB proposed has yielded mixed findings that associate both "promise and problems" with the performance-measurement program (Gilmour and Lewis, 2006a, 2006b).[17] On the positive side, this analysis found that, in some circumstances, when all other factors (e.g., politics) are considered, PART scores for programs were positively correlated with the budgets that OMB proposed for those programs. On the negative side, this occurred only for programs originally created by Democratic presidents or in departments with agendas more closely aligned with Democratic than with Republican priorities. Put another way, although PART scores did not displace traditional political forces in budgeting, they did appear to help OMB direct its attention to the poorest-performing programs in the departments in which OMB sought cuts for political reasons. PART scores did not ultimately affect the budgets that Congress approved for these programs, perhaps because Congress detected more politics than analytic performance measurement underlying the broad patterns of budgeting that OMB sent to Congress.

More broadly, critics have questioned whether the progress documented from year to year in PART metrics reflects real increases in the application of performance-based management, apparent application with little actual use behind the façade, or increasingly lax standards for judging progress. Congress has voiced great skepticism about the objectivity of PART assessments; PART has had little effect on Con-

[17] Richard Nathan (2005) also offers a mixed review of PART. He recognizes many shortcomings in PART but mainly prefers to start with what we have rather than starting over on performance-based management. He seeks broader participation in the choice and definition of performance measures, participation that would make the measures more "intergovernmentally sophisticated" and better informed by formal social science and evaluation. He also seeks greater transparency in the methods used to define them.

gress's budget-building activities. The actual effects of the more-formal GPRA-based plans and reports have not been formally assessed.

Looking forward, many suggest giving less attention to the elements of GPRA and PART that highlight the process of performance measurement and more to the actual level of performance of government agencies. This view is part of a broader concern that formal, mandated performance measurement too easily becomes a compliance program, with little or no effect on actual agency or legislative decisionmaking. Performance-measurement programs also often lose sight of the importance of measuring outcome- and output-oriented performance. By staying focused on key performance factors, agencies that prepare statements about their performance can potentially simplify those statements, effectively increasing their accessibility and potentially their actual use in decisionmaking.

National Performance Review

Under the Clinton administration, the NPR initiated a wide range of pilots that sought to test ideas from TQM and the new public management in a federal setting.[18] State and local governments expanded their development of performance metrics and their application of these metrics to decisions relevant to budgeting, outsourcing, and process improvement.[19] Nongovernmental organizations developed a variety of scorecards that collected information on the performance of a wide range of activities and publicized this information in the hope that the information would change the conduct of these activities for the better.[20]

---

[18] For an early report on the nature of the NPR that conveys how it felt to be involved in it, see Kettl and DiIulio (1995). For a follow-up report on progress, but still from close quarters, see Kettl (1998).

[19] For a global overview that places the new public management in historical perspective, see Lynn (2006).

[20] For a useful overview, see Gormley and Weimer (1999).

Recent State and Local Efforts

Despite the absence of a federal mandate from GPRA, state and local governments have continued to pursue performance-measurement and performance-management initiatives. A number of state and local surveys find widespread interest in and use of new public management tools and approaches (see Brudney, Hebert, and Wright, 1999).[21] For example, a 1996 survey by the Governmental Accounting Standards Board and National Academy of Public Administration (1997) found that 44 percent of the respondent municipalities had developed performance measures for a substantial number of programs; 37 percent reported that they used these measures in decisionmaking processes, such as budgeting, performance evaluation, and strategic planning for substantial numbers of programs. By the summer of 2000, respondents to another survey of 300 local governments reported a pervasive use of performance measurement.[22] Although respondents were not enthusiastic about the effectiveness of performance measurement, the introduction of performance measurement had led to improvements in communication within and among branches of governments and to informed discussion of governmental activities. Use of performance measures to reward or sanction behavior remained unusual. Only 20 percent of respondents reported use of rewards for outstanding performance; only 10 percent reported use of sanctions for poor performance.

A good example of a performance-measurement program that has affected policy outcomes at the state (and federal) level is the National Environmental Performance Partnership System (NEPPS), which has made enforcement of more than a dozen separate federal environmental laws more performance-oriented.[23] Since 1995, EPA and individual states have negotiated specific performance goals. EPA has since granted these states increased flexibility to design and implement cost-

---

[21] Questions about the survey design and low response rate raise some doubts about the validity of its findings. For a good assessment of the survey and interpretation of its findings, see Thompson (2002). See also Poister and Streib (1999).

[22] For details, see Melkers and Willoughby (2005).

[23] The material in this paragraph draws on Metzenbaum (2003) and Herb et al. (2000).

effective initiatives that allow achievement of these goals. Performance measures created and sustained through these partnerships have "added value" by simplifying the identification of successful strategies and benchmarking of strategies across states and the targeting of resources on the most cost-effective strategies and on the resolution of problems revealed by the measures. NEPPS has also "greatly increased public access to information about state environmental performance and their plans to improve it" (Metzenbaum, 2003, p. 5). Today, 34 states participate in these partnerships (EPA, undated).

The evidence on broader implementation and effectiveness of performance management at the state level is limited and mixed at best. Data collected in 2000 by the Government Performance Project on state governments found that heightened expectations about managing for results (one approach to performance measurement) had not been met.[24] State governments had generally introduced the first half of the recipe—a focus on results—while neglecting the second component—an increase in managerial authority. Managerial flexibility, especially in the areas of human-resource management and financial management, remained limited. An April 2002 survey of midwestern mayors' perception of the value of performance measurement found that a significant portion of the responding mayors still did not know whether their city services had performance measures.[25] Those who did understand their performance metrics said that the metrics helped them to learn about city department accomplishments, gauge citizen-level satisfaction with public services, monitor long-term trends in service performance, establish measurable performance goals or targets, have solid information to present to the public, and help evaluate the overall performance of the city manager. That is, metrics were used primarily to support decisionmaking, not to motivate performance of individuals or departments through formal incentive structures.

---

[24] For details, see Moynihan (2006).

[25] For details, see Ho (2006).

Chapter Summary

This brief history of the use of performance measurement and incentives in the public and private sectors provides background for the analyses that follow. The discussion highlights the fact that PBASs represent an evolution rather than a revolution in policy. This perspective offers an appropriate starting point for our study, because it suggests that these systems will continue to evolve. Thus, the findings we present should not be seen as a "vote" in favor of or against PBASs but rather as part of the accumulating body of information that will shape change and modification of this policy tool in the future.

CHAPTER THREE

Problem Recognition and Adoption

We now begin our detailed examination of nine examples of PBASs drawn from five sectors. This chapter focuses on the process by which concerns about service production are recognized and PBASs adopted to address those deficiencies. It also examines some of the most important features of the political, governance, and organizational context into which PBASs are born. The conditions that lead to the adoption and design of a PBAS might be expected to affect its eventual implementation and impact. Furthermore, it is important to evaluate PBASs, at least in part, in terms of the goals they seek to achieve. Thus, it is important to examine those conditions before we examine more-specific design decisions (Chapters Four and Five), implementation (Chapter Six), effectiveness (Chapter Seven), and evaluation (Chapter Eight) later in the monograph.

Specifically, this chapter examines the following questions about the nine PBASs examined in this monograph:

- What concerns motivated adoption of the PBASs?
- What is the nature of the stakeholder and governance contexts in which the PBASs were adopted, and to what extent was there consensus on key design issues related to the PBASs?
- To what extent could PBAS designers rely on evidence about how to improve the service production process?
- How did the PBAS fit with other oversight structures and policy instruments?

Before proceeding, it is important to emphasize that the evidence used in this chapter comes only from transportation, education, child care, PHEP, and health care. While evidence from these sectors helps contextualize findings presented later in the monograph, it is not possible for us to make broad statements about what factors might lead a sector to adopt a PBAS over a different accountability system.

Reasons for Adopting a Performance-Based Accountability System

Adoption of the PBASs we examined was motivated in large part by concerns about the quality of service delivery. In some cases, these concerns arose gradually as the result of changes over time in observed performance, service production, or use; in other cases, they appeared to be dramatic shocks to public consciousness spurred by a single event or series of events in rapid succession.

Concerns Over Quality Sometimes Build for Many Years

In some of the sectors we studied, a worrisome trend in performance became apparent over the course of many years, until, at some point, there was wider acknowledgement that existing governance mechanisms were not performing adequately. For example, in education, there is a long tradition of using standardized tests to monitor school performance, dating back to an initiative in Boston in the 1840s (Resnick, 1982). As the technology of achievement testing developed over subsequent decades, the use of standardized tests expanded. Attention to test results became particularly prominent in the 1970s, when a movement called minimum competency testing (MCT) took hold (Jaeger, 1982). MCT was intended to assure policymakers and the public that all students had mastered basic skills, and often involved the use of test scores to hold individual students accountable by denying graduation or grade promotion. In the 1980s, arguments about the need for testing were bolstered by newly available data on national and international performance (Hamilton, 2003). Advocates for school reform used rankings from international assessments to make the case that

American students were falling behind their international counterparts and that this posed a threat to the nation's security, an argument that was expressed in vivid language in the 1983 publication *A Nation at Risk* (National Committee on Excellence in Education, 1983). These concerns led a number of states to adopt PBASs in the form of high-stakes testing systems in the 1990s. These PBASs provided the impetus for adding performance accountability elements to subsequent reauthorizations of the Elementary and Secondary Education Act, culminating with NCLB in 2001.

Similarly, in health care, there was a long tradition of measuring performance internally for doctors to use for quality improvement. In this case, the health-care community and policymakers were driven to take action due to mounting evidence of substantial quality and safety deficits (McGlynn et al., 2003) and the publication of two key documents, *Crossing the Quality Chasm* (IOM, 2001) and *To Err Is Human* (Kohn, Corrigan, and Donaldson, 1999). Changes in the underlying context of the health sector also created momentum: With the shift to managed care, concerns arose that providers were incentivized to withhold services as a cost-saving measure. This led to explicit efforts to measure the performance of managed-care plans and to hold them publicly accountable.

Much the same dynamic was evident in the transportation PBASs we examined. For example, in the past several decades, traffic congestion in most urban areas has grown persistently worse. Against that general backdrop, traffic congestion can be especially severe when heavily traveled routes are partially or fully closed due to construction activities, leading to potentially significant economic losses due to wasted time and fuel. These factors stimulated experimentation and subsequent adoption of several strategies for expediting construction work, most notably A+B contracting.

PBASs might also be created on the heels of previously unsatisfactory efforts to address a problem. This was true in health care, where physicians' internal measurement efforts were insufficient to ensure that recommended care was provided. Similarly, direct efforts to link performance to accountability are sometimes preceded by indi-

rect efforts, such as the publication of performance reports.[1] In market-based sectors, such as health care or child care, report cards might be intended to encourage consumers (patients or parents, respectively) to choose higher-quality providers or to encourage or even to shame poor performers into making efforts to improve.

Specific Events Can Also Weigh Heavily in System Creation

Sometimes it takes a special event to call attention to a previously ignored problem or to signal a shift in its importance. PHEP presents a clear example of the impact of a single event. Until recently, the federal role in PHEP was limited to providing largely ad hoc assistance and coordination during large-scale incidents that stretched state and local capabilities. During the late 1990s, increasing concern about the threat of weapons of mass destruction led to a small federal effort to build state and local ability to prepare for large-scale public health emergencies. Even though there was already some concern after the World Trade Center bombings in 1993 and the sarin gas attack in Tokyo in 1995, emergency-preparedness efforts grew considerably after September 11, 2001, and the anthrax attacks of October 2001.

The transportation sector also provides an example of the effect a single event can have in leading to the adoption of PBASs. The 1973–1974 oil shocks resulting from the Arab oil embargo triggered severe negative effects for the U.S. economy, galvanizing public support for actions to reduce the nation's dependence on foreign oil. One result of the embargo was the implementation of CAFE standards, which were introduced with the Energy Policy and Conservation Act of 1975 (EPCA) (Pub. L. 94-163).

Sometimes One System Leads to Another

Already-established PBASs might drive the creation of additional PBASs within a given sector. For example, in education, P4P systems that reward teachers for student performance appear to have been cre-

[1] Publishing performance information might be said to create an incentive, particularly where the information might shape consumer or voter choices. Chapter Four provides an in-depth examination of how incentives are used in PBASs.

ated at the district level partly in response to support for accountability that coalesced around the earlier NCLB. Although experimentation with P4P dates back many decades, there has been rapid growth in recent years as a result of increased emphasis on accountability, as well as the recognition that the testing and data systems put in place in response to accountability requirements can also support P4P (Podgursky and Springer, 2008).

The introduction of a PBAS in one sector can also raise concerns about performance in another sector. For example, although studies have consistently found that, on average, child-care quality is mediocre (e.g., Peisner-Feinberg and Berchinal, 1997; Mohan, Reef, and Sarkar, 2006; Karoly, Ghosh-Dastidar, Zellman, Perlman, and Fernyhough, 2008), the sector has not focused much attention on improving program quality. Until quite recently, quality standards were largely defined by state licensing requirements, which generally set a fairly low bar. Licensing is focused primarily on the adequacy and safety of a program's physical environment, including fencing, square footage, and protection of children's well-being (i.e., whether electrical outlets are covered, whether cleaning supplies are locked up). However, the growing policy attention focused on K–12 accountability through NCLB has raised questions about the outcomes of child-care programs, particularly school readiness. These questions have led the sector to focus on quality and devise ways to improve it. Of these, QRISs represent the most popular current approach.

Influence of Stakeholder and Governance Context

We now consider the nature of the stakeholder and governance contexts in which these PBASs were adopted and the extent to which there was consensus on key design issues related to the PBASs. PBASs are introduced into policy contexts that include other potentially competing policy instruments (e.g., regulatory regimes), patterns of influence, balances of power, and cultures. These contextual factors shape the types of PBASs that are put into place and the trajectories they take after their birth.

Service Providers Are Usually Influential Stakeholders

PBASs are, to a large extent, efforts by funders and authorities to shape the behavior of service providers. In the cases we examined, however, the lines of influence usually work in both directions, with service providers also playing an active role in shaping PBASs and other policy decisions.

In transportation, for instance, PBASs designed to improve local transit funding are often strongly influenced by the local jurisdictions that are the subject of the PBASs. Regional transportation planning agencies (RTPAs) are typically governed by boards of commissioners that include representatives of the local jurisdictions within the RTPA. If the transit funding formula applied by an RTPA includes performance measures, and if those measures cause certain transit operators to lose funds, the officials who represent those operators have both the motive and the means either to eliminate or to water down the performance-based elements of the funding formula. Evidence suggests that this has often occurred in practice.[2]

Similarly, in PHEP, the guidance associated with federal grants to state and local health departments is typically shaped not only by congressional and executive-branch priorities but also by a wide range of associations representing state and local health departments and professional associations (e.g., representing epidemiologists or laboratorians).

In health care, physicians and hospital providers have considerable influence over policies designed to affect their behavior, particularly in the area of payment policy. National associations of providers—such as the American Medical Association (AMA), the Federation of American Hospitals (FAH), and the American Hospital Association (AHA)—have helped shape national policy around payment policies and performance measurement. These organizations and their state-level counterparts are heavily engaged in the political and regulatory processes, particularly with respect to what occurs within the context of the federal Medicare program. Additionally, physicians and hospitals have been very influential at the local level in determining the

[2] See, for example, Cook and Lawrie (2004); Hartman, Kurtz, and Winn (1994); and Fielding, Mundle, and Misner (1982).

shape of private-sector programs that measure their performance, publicly report the information, and use the results to reward better performers through differential pay and steer patients to higher-quality, lower-cost providers through differential copayments (i.e., vis-à-vis provider tiering).

In practice, health-care providers have pushed back strongly on many performance-measurement efforts, even threatening to stop providing services (potentially creating access problems) or to bring legal action if their concerns are not considered in the design of the accountability system. The heavy influence of health-care providers in the design of a health-care PBAS was illustrated recently in a legal settlement reached between the attorney general of the state of New York and CIGNA health plan. The settlement occurred in response to a complaint filed by AMA and the Medical Society of the State of New York. Physicians reacted negatively to CIGNA's creation of a narrow network of providers that it would offer to plan members (who would be financially incentivized to use providers in the narrow network by means of lower out-of-pocket payments). To be included in the network of preferred providers, a physician had to demonstrate good performance on quality and cost metrics (Attorney General of the State of New York, 2007). The physicians who filed the complaint objected to the lack of clarity and openness concerning the methods used by the health plan to classify physicians and rejected the validity of these methods. In the settlement, the attorney general noted,

> because measuring physician performance is relatively new, complex and rapidly evolving, the need for transparency, accuracy and oversight in the process is great. In addition, when the sponsor is an insurer, the profit motive may affect its program of physician measurement and/or reporting. This is a potential conflict of interest and therefore requires scrutiny, disclosure and oversight by appropriate authorities. (Attorney General of the State of New York, 2007, pp. 1–2)

In the end, the providers prevailed when the attorney general required that health plans that operate these PBASs ensure "accuracy and transparency of information, and oversight of the process,"

which would be determined by an independent monitor of each health plan's practices. This settlement set a national precedent for how private health plan–sponsored PBASs would be required to operate if they wanted to avoid legal action that could stop their use. In response to this, health plans have been working with NCQA to conduct independent audits of their PBAS program methods and to make their methodologies transparent.

In child care, the influence of service providers is less clear. Here, decisions about the design of PBASs are typically made not by the child-care provider organizations themselves, which are typically small and politically inactive, but by state employees working in agencies that will manage the system; sometimes, politicians weigh in. In many instances, states seeking to design PBASs for child care turn to other states for models. Large national organizations, such as the National Association for the Education of Young Children (NAEYC), which accredits child-care programs, and the National Association of Child Care Resource and Referral Agencies (NACCRRA), which provides resource and referral services in many systems, also play a role.

Influence of Service Consumers and Other Stakeholders Tends to Be Episodic

For the most part, consumers appear to wield less influence over policy than other stakeholders in the sectors we examined. This might be because performance measures are somewhat "arcane" and generally have low salience for consumers (Gormley and Boccuti, 2001). Such influence, when it does occur, often comes in the form of public attention and scrutiny, which might be generated by high-profile activists, elected officials, or the media. For instance, in education, the op-ed and commentary pages of newspapers often offer points of view about educational accountability that probably have some influence on the positions of some policymakers. Frequently, this sort of broader attention contributes to problem recognition and the initial decision to create a PBAS to solve an identified problem.

Decentralized Governance Structures Provide Opportunity for Stakeholder Influence

Most of the sectors we examined were characterized by wide distribution of legal authority and funding sources. This, in turn, provides multiple entry points for stakeholder influence. In public health and education, for instance, primary policymaking authority resides in states, which often delegate authority to local health departments and school districts. Similarly, in several sectors, many of the resources and much of the discretion lie in the hands of nongovernmental entities. In health care, only about 25 percent of the population is covered by public health insurance (i.e., the elderly through Medicare and low-income individuals through Medicaid), while 60 percent of the population is covered by private health insurance plans. While there is some governmental presence in health care, the cultural context in the United States favors private health insurance and the delivery of health care through the private market.

And, in two of the transportation cases, the provision of services falls either predominantly or exclusively within the private sector. With A+B contracting, a large share of construction activities are conducted by private firms; CAFE standards were enacted by the national government but apply to major vehicle manufacturers, all of which are private companies.

Systems Are Often Created Without Clear Agreement Among Stakeholders About Key Design Issues

Given the decentralization of political influence in the sectors we examined, it is perhaps not surprising to find that there are usually differences of opinion about the desirability and general contours of PBASs. What might be more surprising is that PBASs are created in spite of this lack of consensus on how to define quality performance and whom to hold accountable.

How to Define Performance. For most of the transportation cases, there has been relatively little conflict about how to define performance for PBASs, in part because they have thus far been designed around very specific goals—e.g., speeding construction project delivery, reducing harmful air pollutants, or reducing fuel consumption.

For the CAA, applicable ambient air-quality standards were developed by EPA through a scientific process that assessed the harm that different levels of pollutants could cause to humans and the environment. With CAFE standards, federal legislation specifies that manufacturers be judged based on average fuel economy, as measured by EPA, of the passenger vehicle and light-duty truck fleets that they sell each year. For A+B contracting, the goal of reducing construction time can be easily measured in terms of the number of days from start to finish. An effort is now being undertaken, however, to base transportation funding in the future on much broader, more inclusive, and complex measures of transportation system performance, such as mobility, safety, energy efficiency, and economic growth. This is proving much more difficult to do.[3]

In education, the understanding of *performance* is more complicated. There is general agreement about the broad goals of the system and notions of performance: producing high-school graduates with high levels of achievement, advanced skills, preparation for careers or further education, commitment to community and democracy, respect for others, and so on. However, there is less agreement about which specific elements of performance should be measured and assessed. While the public and policymakers seem content with standardized test scores as a fair measure of student achievement, there remains considerable disagreement among educators about whether existing standardized tests measure meaningful aspects of learning and cover desired outcomes adequately.

The story is similar in child care, with agreement on broad goals but little agreement on specifics. Stakeholders generally agree that the goal is to improve the quality of care, which will improve child outcomes. However, there is considerable disagreement about which out-

---

[3] The National Transportation Policy Project's report, *Performance Driven: A New Vision for U.S. Transportation Policy* (2009), proposes restructuring federal programs, updating the criteria for formulas, and creating a performance-based system that directly ties transportation spending to broader national goals. These goals include economic growth, connectivity, accessibility, safety, energy security, and environmental protection. States would be measured on how greatly they improve access, lower congestion and petroleum consumption, reduce carbon dioxide (CO_2) emissions, and decrease fatalities and injuries.

comes matter most. Many focus on kindergarten readiness, in part because they think that being prepared academically for school will lead to longer-term payoffs in terms of educational achievement or earning capability. But some kindergarten teachers and other educators have pushed against the implied focus on academics to achieve this goal, arguing that learning to regulate emotions, delay gratification, and follow instructions, among other skills, are more important and age-appropriate objectives.

In health care, there has been conflict about which aspects of quality should be measured—whether to focus on process measures (i.e., whether a physician performs a certain procedure) or outcomes of care (i.e., what happens to a patient as a result of the care received). Tension has also arisen with respect to the goals of the program; PBAS sponsors and the affected providers differ on whether the focus should be solely on quality performance or also on cost performance. Providers have pushed hard against including cost metrics in accountability systems, expressing concern that PBASs disguise themselves as quality-improvement tools but are really cost-cutting tools. However, the enormous growth in health spending has forced public- and private-sector sponsors of PBASs to focus on costs as well as quality in their accountability efforts, despite provider pushback.

Who Should Be Held Accountable for Performance. Another common source of debate and conflict involves *who* should be held accountable. Generally, service providers prefer to be held accountable only for those aspects of service production over which they have clear and direct control. In PHEP, for instance, health departments contend that providing adequate security for mass vaccine or antibiotic-dispensing operations relies heavily on law-enforcement agencies over which they have little control. Thus, they often caution against PBASs that would hold them responsible for security or argue that they should held accountable only for such activities as building partnerships with and coordinating with security agencies.

In health care, the accountability issue has played out in several ways. Providers have asserted that they should not be held accountable for things that are outside of their control. For example, reducing morbidity and mortality associated with diabetes requires good

blood-sugar control and weight loss—both of which are influenced by the patient's behavior. In this situation, providers are more comfortable with being held responsible for ensuring that the patient's blood sugar is monitored and counseling the patient on weight control and diet because patients vary greatly in their compliance with physician recommendations and adherence to prescribed treatments. In the case of mortality outcomes, risk adjustment to account for differences in the patients (e.g., age, other comorbidities) has been used to help level the playing field among providers when making comparisons to account for risk factors with which the patient presents upon admission to the hospital.

Conflicts have also arisen regarding how broadly or narrowly the organizational unit of accountability should be defined. For example, accountability issues occur in health care regarding the most appropriate target for behavioral change—is it the physician, the practice site, the larger medical group, or an integrated delivery system of physician groups and hospitals? Some PBAS sponsors assert that the widest variation in practice is at the physician level, so attention should be focused there; however, providers emphasize that the delivery of care requires systems and that it is more appropriate to measure at higher levels of the organization that encompass the system elements.

In education, while there is evidence that individual teachers play a significant role in student achievement, there has been a reluctance to make teachers the object of accountability. Most PBASs in education use schools as the unit of accountability, a focus that acknowledges the difficulty of isolating the contribution of individual teachers to student learning, given the role of other staff members and the whole school environment. Recently, with advances in value-added statistical analyses, some PBASs (namely, P4P) are focusing on teachers as the unit of accountability. Some aspects of the system also locate accountability with students, e.g., high-school exit examinations.

Understanding of Service Production Processes

The existence of a strong knowledge base about the drivers of performance in a sector might help inform debates about who should be held accountable for what, thus easing the process of developing stakeholder consensus around a PBAS.

There Is Considerable Variation Across Sectors in the Quality of the Knowledge Base for Performance-Based Accountability Systems

We can think about this knowledge in terms of what is known about the *service production function*—the structures and activities involved in producing the service and how they are related to outputs and outcomes. For instance, the health-care production function represents what is known about the relationship between the health-care delivery structure (e.g., the number of providers within a community and the presence of electronic data systems to manage patient care), health-care activities (including various forms of treatment and other actions to enhance patient health), and health outcomes (e.g., quality of life, morbidity, mortality).

In health care, the growing body of evidence from clinical trials and other sources has led to a reasonably strong understanding of the linkages between clinical interventions and health outcomes for some conditions. An example is the use of beta-blockers after a heart attack to prevent a recurrence. Similarly, achieving blood-sugar control levels of less than seven in diabetics has been shown to reduce complications (e.g., foot amputations, blindness). However, there are also many areas of health care in which the production function is not well understood, and, consequently, there is greater uncertainty about what to measure. Advances in the science base often mean that knowledge of the production function changes over time, which has required health PBASs to continually monitor the evidence and make changes to performance measures.

For A+B contracting and CAFE standards in the transportation sector, this relationship is relatively well understood. For example, with A+B contracting, there are several known strategies for speeding up construction, including, most obviously, working more shifts per day

and more days per week. Similarly, with CAFE standards, there are many known design and engineering options to make vehicles more fuel-efficient—lighter materials, more-aerodynamic profiles, advanced engine systems, and the like. With the CAA, the science of how tailpipe emissions of various pollutants affect air quality is well understood, although it is worth offering one caveat. Specifically, under the CAA, metropolitan PBAS planning organizations are required to show, via modeling, that future transportation investments will not stimulate future travel decisions that adversely affect air quality in the region. Given the inherent uncertainty in forecasting the future, such predictions might be fraught with error.[4]

In education, by contrast, knowledge of the production function is much weaker. Research demonstrates that teacher quality has a large impact on student achievement growth, but the research has not yet identified the characteristics or practices that are associated with high-quality teaching. Easy-to-measure attributes, such as years of experience or whether the teacher has a standard certification, have very small effects at best, and attempts to measure the effects of specific instructional practices have produced few clear findings. Similarly, research on the structural features of schools (e.g., class size) has produced mixed results; under experimental conditions, reducing class size led to significant gains in achievement (Finn, 2002), but, when implemented on a large scale, the effects were quite modest (Stecher, 2002). Overall, the education research literature does not identify specific strategies for raising student achievement broadly.

Clearly, weaknesses in knowledge about the service production function have not stopped sectors from adopting PBASs, something suggested by the fact that, in spite of variation in knowledge described above, each of the sectors we examined has adopted a PBAS. Yet, uncertainty about issues, such as whose behavior is most responsible for service outcomes, sets the stage for continued conflict over such techni-

[4] As noted earlier, there is a movement within the transportation sector nationally to apply PBASs to much broader elements of the transportation program, in which the relationships between specific activities and outcomes are not terribly well understood.

cal issues as whom to incentivize and what aspects of performance to measure—issues we take up in the next two chapters.

Tension Between a Performance-Based Accountability System and Other Oversight Structures

The final question explored in this chapter concerns how a PBAS fits with other oversight structures and policy instruments. PBASs often operate alongside other PBASs and other oversight structures (e.g., regulatory frameworks). This is especially the case in sectors with a long tradition of measurement and accountability (e.g., transportation, health care, education), in which service providers receive funds from multiple sources and through many funding mechanisms. The presence of other oversight structures will certainly influence the creation of a PBAS, which might be constrained by existing rules and might require changes to existing regulations in order to operate effectively. For example, the myriad public- and private-sector governance structures that exist within local health-care markets create competing demands for the attention of service providers, and, unless their efforts align with the existing governance structures, PBASs are challenged to garner physicians' attention.

In education, primary responsibility for schooling is vested in the states; thus, states have long had regulatory systems in place to govern schools. Some are very centralized; others are quite decentralized, pushing such decisions as curriculum down to the local district level. NCLB was layered on top of these existing state systems. In most states, the NCLB rules were different from the state rules, usually setting tougher requirements and imposing stronger sanctions. The threat of withholding federal resources (about 8 percent of education funding in most states) led states to bring their systems in line with the NCLB rules. Additionally, in some instances, pursuit of NCLB requirements might conflict with teacher union contracts. Similarly, in child care, PBASs have typically been created in isolation, with little regard for how they interact with other programs. The end result is that the "signal" from any one PBAS might be reduced by "noise" from other systems.

In transportation, by contrast, PBASs were simply incorporated within existing governance structures. With CAFE, for instance, the task of verifying compliance with the standards and issuing penalties, where appropriate, was assigned to the National Highway Traffic Safety Administration, which already regulated other vehicle features related to safety (e.g., passenger-restraint systems). For the CAA, metropolitan planning organizations were already required to conduct periodic comprehensive transportation planning efforts to qualify for federal funds; examining the air-quality impacts of future transportation investments was simply added to the required modeling capabilities of these planning activities.

Chapter Summary

While the exact sequence of events leading up to the decision to adopt a PBAS is unique to each circumstance, some similarities are apparent. All PBASs begin with the recognition of some problem that the current governance structure has been shown to be incapable of addressing. In some cases, it might gradually become apparent that existing mechanisms are not adequate to reverse a worrisome trend—e.g., in education, licensure and accreditation have not stemmed the tide of failing schools and students. Or, sometimes the awareness of the need for change might occur in response to a shift in the underlying context of the sector itself, as was the case in health care and A+B contracting. It is even possible for a PBAS to *change* the underlying context of its own or a related sector, prompting the creation of additional PBASs. In other cases, an important event or new finding might shock stakeholders into action—the events of September 11, 2001, for example, brought heightened attention to the need for better emergency preparedness. Finally, in many cases, PBASs are created on the heels of previously unsatisfactory efforts to address the problem and, in some cases, are inspired by other PBASs. The bottom line in all of these cases is that current efforts and mechanisms are not enough to solve the problem, and something new is needed.

Political rhetoric and how-to guides on designing PBASs usually portray them as the product of relatively dispassionate discussions about performance. In reality, the process that gives birth to PBASs is usually less straightforward and less rational. In the cases we examined, the decision to adopt a PBAS was heavily influenced by the preferences of service providers—the very people whose behavior the PBAS sought to shape. Given conflicts among these stakeholders, it is perhaps not surprising that PBASs often proceed in spite of a lack of clear agreement on what constitutes *performance* and on who should be held accountable for what. In many sectors, there is not a sufficiently strong evidence base to provide scientific guidance to would-be PBAS adopters and designers. Finally, all PBASs are created in a context with other governance structures; tension among these structures is to be expected because the PBAS is created to ameliorate a problem the existing structures did not address. In short, PBASs often inherit many of the very same political conflicts they seek to transcend. These challenges need to be addressed when designing the PBAS.

CHAPTER FOUR

The Design of an Incentive Structure to Motivate Behavioral Change

As noted in previous chapters, a PBAS is designed to change service providers' behavior by creating incentives for change that are linked to performance measures. This chapter focuses on the incentive structure, or the decisions required in assigning rewards and sanctions to individuals or organizations. Incentives can include (1) the potential effect on a service provider's reputation of reporting its performance, (2) specific resources made available to the organization based on the service provider's level of performance, (3) specific promotion and training opportunities or bonuses made available to individual employees who work within the service-provider organization, and (4) sanctions for individuals or organizations that fail to achieve certain behavioral changes that the PBAS sponsors and designers seek.

The designers of a PBAS must translate their performance goals into measurable activities and specific incentives about which service providers or users care. Such incentives might include

- an organization's budget allocation, market share, access to high-quality personnel, or level of profit
- an individual's income, opportunities, or commendation
- users' opportunities to consume the services and products they value.

Once PBAS designers determine the goals for the PBAS and the incentive structure needed to induce change, they need to determine how to measure performance and which specific measures should be used to link performance to incentives.

This chapter explores two basic design issues that arise in creating a PBAS:

- Whose behavior must change for the PBAS to succeed?
- What form of incentive structure should be used to induce change, and should support or assistance be provided to promote the desired kinds of change?

Chapter Five continues the discussion by focusing on a third design issue: How should performance be measured, and, in particular, how should performance measures be used to link performance to incentives?[1]

This ordering of chapters might suggest that the identification of those individuals or groups held accountable in the PBAS and the types of incentives used to motivate change necessarily precede the act of determining which performance measures to use. Instead, the activities are highly interconnected (see Figure 1.1 in Chapter One). To illustrate, one could reasonably argue that the process should begin with identifying performance measures of interest (such as rates of mam-

[1] Since the 1990s, many articles and books have offered advice on how to manage change effectively in large, complex organizations. One point that arises in a variety of forms in this literature is that organizations can change only if the people within these organizations change their behavior in a coordinated fashion. Incentive structures play an important role in supporting such coordination. Moore et al. (2002, p. xvii) synthesizes views on this point by asking the following questions:

> If the people who must change their behavior on the job to make a [new practice] succeed in fact change their behavior, will they benefit? If they do not change their behavior, will they be punished? An organization can use any combination of monetary inducements, awards, career actions, and other incentives that is compatible with its corporate culture to reward or punish its employees. But if no personal benefit flows from changing their behavior on the job, or the risk of adverse consequences increases, employees see little personal connection to the success of the change. Because it is typically more comfortable to avoid change, they will often choose the status quo.

Moore et al. (2002) draws on many sources. See, for example, Kotter (1996), in which step 6 of eight key steps advises change managers to ensure that employees who support change are recognized and rewarded. Kotter views each of his eight steps as critical to success; failure to heed any one of them is likely to impede progress. Many others make the same point in different ways.

mograms offered to women at 40 years of age), and then determining the individuals or organizations that should be targeted to ensure that women are offered mammograms (e.g., physicians, health maintenance organizations, or HMOs). For the purposes of this monograph, however, we describe the identification of incentives before discussing measures; this follows the logic that desired goals are carried out by various individuals or organizations, which are, in turn, driven by incentives and penalties.

Whose Behavior Must Change to Meet System Goals?

After identifying high-level goals, PBAS designers must determine whose behavior they want to target for change. In the context of PBASs, the question is how the designers of the PBAS can create a PBAS that induces a service provider or user to do what the designers want done—especially when the service provider or user might prefer to do something else.[2]

In the Sectors Examined, System Designers Quickly Identified Whose Behavior Needed to Change

In each sector we examined, the designers of a PBAS were able to determine fairly quickly who had to change their behavior to achieve the goals of the PBAS. In child care, for example, two groups were identified as being important. First, *child-care providers* must decide what level of quality to offer; second, *parents* must decide whether to buy child care and, if so, what level of quality to buy. Therefore, it was appropriate for the design of a child-care PBAS to employ incentives

[2] During his review of this monograph, William Gormley raised an important and closely related question: *How many* people need to change their behavior for a PBAS to achieve its goals? Ultimately, many people must change their behavior to achieve real improvements. But the incentive structure of the PBAS might need to affect only a small number directly, relying on their example to diffuse change. For example, even if a PBAS directly induced only a small number of child-care centers to improve their quality, additional government funding that flowed toward this small group following the change could dramatize the value of change and persuade others to follow suit. Our analysis did not reveal any empirical insight into this issue. More-targeted attention to this question could be quite helpful.

that target the behavior of these two types of actors. In this case, the PBAS was designed to make child-care providers' quality ratings available to parents, who, in turn, could choose among child-care providers and thereby signal the quality of services they demanded.

In PHEP, it was somewhat more difficult to identify the organizations whose behavior needed to change. There is variability in who is responsible for securing medical warehouses: sometimes the National Guard; other times, state police or private contractors. This means that it is not always immediately clear whose behavior needs to change. Most often, the decision is made to hold health departments accountable for ensuring that security plans and capabilities are in place, but there is often pushback from health departments, which point out that they have limited control over many of the personnel who actually provide security (for instance, the health director cannot compel the police chief to deploy officers).

System Designers Sometimes Seek to Change Organizational Behavior

The designers of a PBAS sometimes seek to change the behavior of the organizations—such as a health-care provider organization or a school—that provide the services or products. An organization that provides a service obviously cannot change its own behavior unless its units and individuals within the units change theirs. Thus, the designers of a PBAS must choose whether to focus attention on the organization as a whole, on some intermediate level of the organization (e.g., the units or subdivisions within the organization), or on the individuals within the organization. PBAS designers can try to motivate an organization as a whole and rely on the organization's internal governance structures to convey direction to units and individuals. For example, when seeking to change the behavior of child-care providers, PBAS designers have focused primarily on provider organizations, not their employees. These designers targeted the organizations, which tend to be small, and depend on them to induce desired changes among their individual employees. This decision makes sense given the high level of staff turnover in many child-care settings and the relatively low level of

professional training among staff in many.[3] In contrast, P4P in health care has targeted individual physicians within hospital settings and HMO systems to provide a determined level of service.

System Designers Must Ensure That Individuals and Organizations Targeted for Change Can See the Connection Between Their Behavior and System Goals

In general, many organizations and individuals can contribute toward the goals of a PBAS. Can the designers of the PBAS identify and isolate the specific contributions of those individuals or organizations held accountable in the PBAS? For example, can PBAS designers determine the individuals or organizations responsible for the level of student performance in a school district? Education planners, curriculum designers, resource managers, principals, teachers, coaches, counselors, students, and the students' parents and peers all play some role.

In order for individuals or organizations to respond effectively to PBAS incentives, they need to understand the connection—whether positive or negative—between their behavior and the goals of the PBAS. Telling a student that a PBAS is designed to raise the performance of the students in the bottom 10 percent of the student population, as determined by measured performance, for example, is unlikely to attract the student's immediate attention, even if she knows that she is in the bottom 10 percent. If she heard only that, what could the individual student do to change the situation? The goal becomes more meaningful as we move up through the school system to levels with broader responsibility for adjusting all policies and practices relevant to the lowest-performing 10 percent of the student body.

The more closely a PBAS designer can match an incentive to the behaviors or activities that an individual or organization controls, the easier it will be for that individual or organization to respond effec-

[3] However, incentives focused on individual staff, e.g., free tuition for early-childhood education (ECE) classes for staff working in settings that achieve a specified star rating, are common and have been very popular. In one state that offered such incentives, training facilities were overwhelmed (Zellman and Perlman, 2008).

tively to the incentive.[4] For example, even if a student in the bottom 10 percent of his class is offered a reward for improving, the student might not know how to do better. Fryer (2010) finds that student performance increases if cash incentives are given for changing their behaviors (inputs) but not if incentives are given for test scores (outputs). His interpretation is that students did not know how to turn their excitement about earning the incentives into measured achievement. In contrast, a teacher who is told that he will be rewarded if these low-performing students improve their performance might be able to adjust some activities that he controls, e.g., the presentation of material, the lessons used from the curriculum, the supplementary materials, and the amount of time spent with individual students. In deciding whom to hold accountable for change, PBAS designers must consider what outputs are under the control of relevant individuals or organizations. Different organizational levels—school system administrators, resource managers, principals, and so on—can have differing levels of control over decisions about class size, scheduling of classes, resources available for tutoring and other help beyond the classroom, teacher placement, selection of textbooks, day-to-day curriculum, and so on. PBAS designers should take into account the decisionmaking authority and control possessed by the individuals or organizations held accountable for changing the quality of services.[5]

---

[4] Note that individuals and organizations tend to have more information than others about the things they control. The argument here is not about the presence of information, which in itself is important. It is about changing the incentive structure in which individuals and organizations use the information they have—inducing them to use that information to advance higher-level goals.

[5] It is rarely possible to identify one individual or part of an organization that has the sole control over some output of interest. So we do not want to suggest that complete control is required for an incentive to work. Rather, the more effective control an individual or organization has over a specific effect that a PBAS incentivizes, the more effective the incentive applied to that individual or organization is likely to be.

A System Should Distinguish the Contribution of Individuals from That of Teams

Another design issue concerns whether individuals or teams can most effectively ensure progress toward the goals of a PBAS. Rewarding individuals in a setting in which teamwork is important can be counterproductive if doing so induces individuals to distinguish their performance from that of others on the team.[6] Similarly, if organizations (or their units) are most productive in cooperation with each other, rewarding individual organizations or units will motivate them to act in ways that produce individual benefit while degrading the joint efforts of all units taken together. However, if individual effort is most critical to a PBAS's success, then incentives should be designed to target relevant individuals. This is especially true if PBAS designers have the capability to measure an individual's performance and determine how the individual's performance, relative to others' performance, contributes to the PBAS's goals.

If teams are determined to be the best targets for change, then PBAS designers must also determine how large the targeted teams should be. The more important team effort is, and the larger the teams relevant to such effort are, the more productive it is likely to be for the operators of a PBAS to target larger, more-inclusive teams, offices, or organizations with specific incentives. In the absence of significant team effects, targeting larger groups can also advance the goals that define a PBAS's success if the operators of the PBAS have difficulty looking within a larger team or organization and identifying who exactly is responsible for the success that, from the PBAS operators' perspective, is more clearly manifested at a higher organizational level.

The issue of whether to target individuals or groups can be challenging to resolve. In health care, for example, it might be very difficult to determine who in a large HMO has contributed most to an individ-

---

[6] W. Edwards Deming developed this argument as part of his approach to improving organizational management, which came to be known as TQM. See Chapter Two for further discussion. For a useful, application-oriented overview of TQM, see George and Weimerskirch (1994). Many more-recent, high-profile proponents of TQM have placed a similar, increased focus on enabling and sustaining teams. See, for example, Davenport (1993) and Senge (1990).

ual patient's weight loss. If a PBAS is designed to reduce patient weight, it will likely be easier to link changes in average patient weight to the HMOs that treat patients than to link one patient's weight to the decisions of an individual doctor. If, on the other hand, a PBAS is designed to improve the application of specific standards of care and its operators can directly observe the decisions a doctor makes with specific patients in specific circumstances, a narrowly and precisely focused incentive directed at individual doctors could offer the best incentive. In some cases, hybrid models might be appropriate as ways to signal the importance of both group and individual contributions. In education, for example, P4P initiatives, such as the Teacher Advancement Program and the IMPACT program in Washington, D.C., include measures of both individual and team performance. Chapter Five returns to this theme in a discussion about measuring performance.[7]

Individuals and Organizations Targeted in the Nine Cases Have Varying Levels of Control Over Desired Changes

Table 4.1 summarizes goals of the PBASs examined in our study and identifies the individuals or organizations held accountable for changing their behaviors. Given the goals, the targeted individuals or organizations appear to be appropriate in the child-care and transportation PBASs. In contrast, those targeted in education and in health-care P4P PBASs have limited control over the changes expected of them; in these sectors, reaching consensus on the definition of *performance* might be difficult. For example, in the case of NCLB, teacher union contracts limit a superintendent's ability to reallocate human resources

[7] Where a PBAS lacks the information required to sort through the issues discussed in the text, it is probably safer for it to target an organization than the individuals in the organization. The organization, which we would expect to understand its own processes better than any outside observer, then has the option of using its internal governance structures to align individuals' incentives with its new goals. This approach would not work if the outside observer wanted individuals to make very specific changes that the organization did not value or understand. For example, a PBAS might want to induce very specific diagnostic procedures among doctors. But such knowledge on the part of the PBAS in itself presumes considerable knowledge of how the organization works—for example, who has responsibility for executing what diagnostic procedure under what circumstances in a particular medical practice.

Table 4.1
Whom System Designers Try to Induce to Change Their Behavior

Sector/PBAS	Change Sought	Parties Targeted for Change
Child care, QRISs	Improved quality of child care	Organizations that provide child care, parents
Education, NCLB	Improved academic proficiency, graduation	Districts, schools
Education, P4P	Improved student achievement, other outcomes, educator practices	Teachers, principals
Health care, P4P	Improved quality of health care, reduced resource use and overuse of services	Doctors
PHEP	Improve capacity to address large-scale emergencies	State and local public health departments
Transportation, A+B	Reduced time to complete road projects	Contractors
Transportation, CAA	Improved air quality	Metropolitan planning organizations
Transportation, CAFE	Energy independence and reduced greenhouse-gas emissions through improved fuel economy	Vehicle manufacturers
Transportation, transit	Improved performance and efficiency of transit service	Transit districts

among schools. Similar limits on control exist in health-care P4Ps; health departments generally create public risk communication messages but must rely on governors' or mayors' offices or media outlets to distribute them. This limits health departments' ability to get messages out in a timely manner, which is critical during high-tempo disasters. Chapter Five examines this question in some detail. The PHEP PBAS case might face similar challenges if public health departments can control only a portion of the resources and planning relevant to PHEP; however, this PBAS is too young to yield evidence at this time.

Incentive Structure Used to Induce Behavioral Change

Once PBAS designers have identified the individuals or organizations whose behavior should change, they must decide what form of incentive to use. The context in which PBAS designers work is likely to shape the range of incentive options available. Within that set of options, the designers of the PBAS must decide (1) how large an incentive to offer and whether to make the incentive positive (to reward success) or negative (to penalize failure) and (2) whether and how to integrate rewards and penalties with training and technical assistance designed to help the service providers improve their performance to reach the PBAS goals.

Context Shapes the Incentive Options Available

Three attributes of context are likely to shape the design of incentives.

How Much Authority Do PBAS Designers Have to Shape Behavior? In the absence of regulatory authority, the operators of a PBAS might be able only to collect and disseminate information about the performance of the service providers and users they oversee, hoping this information will induce them to change the service that they provide or consume. With regulatory authority or when coordinated with regulatory authority, PBAS designers can use sanctions to induce behavioral change.

How Much Control Do PBAS Designers Have Over Resources? If PBAS designers control the allocation of public resources, they can reward high-performing organizations over low-performing ones. This is the intent of performance-based budgeting, which seeks to expand organizations that perform well and shrink those that do not. Such a degree of control over resources is unusual, however, and, where it exists, such PBASs might still be riddled with challenges. For example, in transit subsidy allocation, political logic favors the distribution of public resources equally across districts, and this has undermined the ability of PBAS designers to reward high performers with more public resources. Efforts to alter the distribution of resources in school districts have been hampered by collective-bargaining agreements that prohibit the allocation of funds based on performance.

More typically, the PBAS designers can control the expenditure of resources on cash and noncash awards or influence personnel management decisions (e.g., who is promoted, who receives valuable training and experience). In the absence of control over resources, the operators of the PBAS must fall back on any regulatory authority or the dissemination of information.

When the PBAS Designers Control Resources, What Options Will They Have on How to Apply Them? Using resources to provide cash bonuses and high–cash value awards, such as cars or travel, is much easier in the private sector than it is in the public sector. Even in the private sector, organizations tend to have their own culture defining the types of rewards that are appropriate. In all likelihood, PBAS designers will apply the resources they can influence in a way that is compatible with the culture of the service providers. For example, PBAS designers in one situation might reward good performers with resources for additional training, while another might reward good performers with public recognition or praise.[8]

In the case of child-care PBASs, specific incentive structures differ from state to state. The simplest PBASs periodically audit each child-care provider that opts to participate, then give each provider a simple rating. PBAS designers in other states have chosen to incentivize providers and parents more aggressively. They created PBASs that include tiered reimbursement schemes, in which the level of payment for children eligible for a subsidy is linked to a provider's quality rating. Such subsidies encourage providers to incur the costs necessary to offer higher levels of quality and, in turn, encourage parents to seek higher levels of quality.

The Size of an Incentive Should Reflect the Value to the Government of Changing the Targeted Behavior

A PBAS induces individuals or organizations to invest their efforts in a way that promotes the PBAS's goals. Any incentive, therefore, must be large enough to offset the cost of service providers' efforts to change

[8] For a discussion of this point and its implications for the development of incentives in organizations with different cultures, see Camm et al. (2001).

their behavior when they reap the rewards or avoid the sanctions. And, if PBAS designers are offering a reward, the reward should not be greater than the value of the desired change.

For example, A+B contracting works by inducing contractors to incur higher construction costs in order to speed construction. These costs might be associated with paying workers overtime, negotiating deals with suppliers that give the contractors speedier access to their supplies, or using machinery and methods that allow construction to move faster but cost more than conventional construction machinery and methods. To speed construction, the designers of a PBAS must reward a successful contractor enough to offset the costs of such increased effort. Presumably, the more the designers of a PBAS are willing to pay for rapid construction, the more a contractor will be willing to pay to complete construction more rapidly.

So how large an award should the PBAS designers offer for more-rapid construction? The simple answer from the principles of cost-benefit analysis is that they should raise the reward until the last dollar increases the speed of construction just enough to be worth an extra dollar to the designers of the PBAS.[9] How high is that? The answer is inherently political and is likely to vary from one jurisdiction and PBAS to another.[10] In the types of PBASs we examined, we found almost no evidence about the dollar value of benefits that a PBAS designer might use to directly compare incremental costs and benefits.

But the basic principle is general: Designers of a PBAS should be willing to increase the reward by a dollar, as long as they can expect this dollar to induce a change in service providers' behavior that yields more

---

[9] For general principles, see Mishan (1976). For implications of those principles in the design of incentives meant to shape behavior between organizations, see Laffont and Tirole (1993). These standard sources speak of costs and benefits in monetized terms. Thinking more metaphorically, analysts and policymakers can use the principles that these works present more generally, defining costs and benefits in whatever terms they prefer.

[10] Even when a transportation department seeks to measure the economic value of congestion to decide how much to pay to reduce congestion, as many departments do, the cost-benefit analysis used to inform public investments is fraught with political considerations. In most circumstances, the tools used to measure expected benefit and costs are ultimately the servants of government agencies that operate in a highly politicized environment.

than a dollar of benefit.[11] Similarly, they should be willing to increase incrementally the cost of a sanction as long as it does not exceed the incremental benefit gained when service providers change their behavior. For example, the designers of a health-care PBAS should be willing to pay a doctor a reward for pursuing a standard of care if the benefit of that care exceeds the cost of the reward paid. The PBAS designers who enforce the CAA should be willing to impose restrictions on a city's development decisions to enforce compliance as long as the social costs of these restrictions do not exceed the social costs of being out of compliance with the CAA.

If the PBAS's incentive structures are properly designed, the incentives will reveal the relevant social values of the PBAS's goals. However, the designers of PBASs can pursue the same goals from different directions. For example, a doctor can use many different approaches to manage the level of cholesterol in a patient's blood. If the designers of a PBAS incentivized each of these approaches with separate rewards or sanctions, we could, in principle, observe whether the rewards and sanctions collectively implied the same level of social value associated with lowering cholesterol. If they did not, the designers of the PBAS should adjust the rewards and sanctions until they were equally cost-effective—that is, each tended to yield a change in cholesterol level that was compatible with the level of effort a doctor had to make to comply with the incentive structure.[12]

We should not infer from the simplicity of this principle that it is easy to implement in practice. Identifying consensus levels of the social costs and values involved in each case will be challenging, but

---

[11] For example, commercial airlines use A+B contracts with aircraft-maintenance companies to incentivize the maintainers to operate in a cost-effective maintenance time. These contracts typically state a target number of calendar days for a maintenance action—say, 25 days. If the maintainer takes a day more, the maintainer must provide the airline an equivalent aircraft to use for a day or pay the airline to lease such an aircraft. If the maintainer finishes a day early, the airline pays the maintainer a bonus equal to the cost of leasing an equivalent aircraft for a day. Methods are available for applying this approach in a government setting. See for example, Keating and Loredo (2006). Again, the formal methods applied here provide heuristics for valuing benefits that are harder to monetize formally.

[12] For an example of this approach, see Dixon and Garber (2001).

this principle does offer a simple framework in which to pursue such complex social judgments.

Our observations of PBASs revealed that, in practice, decisions about incentive structures (e.g., whom to target, the amount and form of reward or sanction) are difficult and can produce unintended consequences. In the education sector, for example, there is strong evidence of teaching to the test—disproportionately focusing teaching on the anticipated content of the tests that will measure performance—and of teachers working hardest to improve the performance of the students closest to the thresholds in an incentive structure and, hence, of the students whose behavior is most important to the performance measures that the designers of the PBAS have chosen.[13] We observed comparable responses in other sectors. PBAS designers therefore should be aware of these reactions and be prepared to adjust the design of the PBAS incentive structure and performance metrics when they observe unanticipated and undesirable behavioral responses.

Another problem occurs when PBAS designers do not have adequate control over resources or regulatory authority to create an incentive structure that complies with these principles. In such cases, designers should strive for "second-best" solutions by shifting resources and regulatory sanctions from a place where they have a small effect on progress toward the PBAS goals to another place where these same resources or sanctions have a larger effect (Lipsey and Lancaster, 1956–1957).[14]

In practice, of course, this is hard to do. The ability to adjust the design of incentives requires careful, ongoing monitoring and evaluation of the PBAS in operation—and we found very little evidence that PBAS designers monitor PBASs after implementation.[15] Perhaps this absence of evaluation explains the paucity of evidence in all of our cases

---

[13] It is easier to discuss this topic in depth with a better appreciation of the role of performance measures. As a result, we defer more-detailed discussion to Chapter Six.

[14] Note that, in our context, we can expect to address the "second best" only metaphorically. For a useful discussion of the effects of second-best reasoning on the design of incentives, see Laffont and Tirole (1993).

[15] See Chapters Eight and Nine for details on what we observed.

about how individuals and organizations actually respond to changes in the level of rewards or sanctions they face.

Training and Technical Support Can Sometimes Be Used to Enhance Incentives

Training and technical support play important roles in the child-care and education PBASs we examined but were generally not as important in the other cases. Under some child-care PBASs, provider staff are eligible for training only when the provider scores above a certain threshold; this gives a provider an additional benefit for reaching a target level of quality. Some child-care PBASs focus training and technical support on areas in which a provider needs the most help. In the case of NCLB, school districts are required to devote additional technical support resources to schools that fail to meet performance standards. Such support presumably helps the affected providers improve their performance in the future and so can be seen as providing benefits in the context of a PBAS. These benefits are clearer for child-care providers because support comes from the outside and adds to the total level of resources available. For school districts, on the other hand, the support must be funded out of district funds. Consequently, the requirement to provide such support actually constrains the districts' freedom to allocate their limited resources as they would prefer. The support might lead to future benefits for the schools affected, but it often makes the districts in which these schools operate less able to do what they would have preferred in the absence of NCLB.

The provision of technical support is especially useful when PBAS designers want to teach service providers how to provide services more cost-effectively. The offer of technical support presumes that someone outside a provider's organization knows how the provider produces its services and can therefore offer advice. In education, when the designers of NCLB prescribed the provision of technical support, they directed districts with low-performing schools to find and fund appropriate experts. That is, a district that performs poorly must redirect its scarce resources toward technical support, the content of which remains under the district's control. While a district must use its own funds and might choose the type of technical support to provide, in

effect, the designers of NCLB gave the district an incentive to do the training *right*—however the district might choose to do it. If test scores improve, the district will not have to continue to redirect its resources toward technical support.

When the designers of a PBAS offer technical support as a pure benefit to underperforming providers, these designers risk rewarding low performance. Low achievers can potentially continue to have access to such a free benefit by limiting their own performance. Of course, if the support is useful only in improving performance, there is no clear benefit for a low-performing service provider to overuse the support. However, if low performance leads to support in a fungible form—for example, the provision of resources without clear restrictions, or staff training that can be used to acquire skills for a higher-paying job at a higher-performing provider—the support can be counterproductive to the PBAS goals. To mitigate these potential problems, the designers of child-care PBASs constrain the types of technical support provided; generally, support is closely linked to quality ratings and focuses on those areas in which a provider performed worst. Culture in the education sector also encourages the recipients of technical support to get through it and beyond it quickly, helping to limit any potential for overuse.

Communication can be important in helping the individuals or organizations being held accountable to understand the changes sought, motivation behind these changes, the rewards and sanctions they stand to face, and how to respond effectively (see Moore et al., 2002). Chapter Five discusses efforts to ensure that those monitored understand the consequences of specific changes in their behavior in the hope that this will strengthen the incentives that PBAS designers apply.

Cases Studied Varied Widely in the Use of Rewards and Sanctions

Table 4.2 summarizes the incentive structures that we found in our PBAS cases and displays a wide variation in the use of rewards and sanctions. Context appears to play a dominant role in the designers' choices, but it is not by itself determinative. Designers of PBASs that arose in regulatory governance settings—NCLB, PHEP, CAA enforce-

Table 4.2
Incentive Structures That System Designers Use to Induce Behavioral Change

Sector/PBAS	Parties Targeted for Change	Incentive Structure Employed
Child care, QRISs	Companies that provide child care; parents	Summary performance score presented as public information Performance-based subsidies External technical support
Education, NCLB	Districts, schools	Sanctions based on several performance scores
Education, P4P	Teachers, principals	Bonuses, professional opportunities based on several performance scores
Health care, P4P	Doctors	Payments based on a summary performance score
PHEP	Public health departments	Potential sanctions in the form of denied federal funding
Transportation, A+B	Contractors	Contract price based on simple performance score
Transportation, CAA	Local governments	Various sanctions based on several performance scores
Transportation, CAFE	Vehicle manufacturers	Specific fines based on two performance scores
Transportation, transit	Transit districts	Performance-based budget allocation

ment, and CAFE standards enforcement—tended to choose sanctions commensurate with the priorities that prevail in those settings. The effective sanctions associated with NCLB remain somewhat diffuse and uncertain, perhaps a consequence of the complexities of coordinating federal, state, and local roles in the provision of education. The federal funds at risk for metropolitan areas that do not comply with the CAA are much better defined. The CAFE standards define sanctions in even more precise and simple terms and induce easily predictable outcomes for behavioral change.

Designers of PBASs in which participation was voluntary—child care and A+B contracting, for example—tended to prefer rewards. The rewards used in A+B contracting are quite explicit, but there is limited evidence on how well they reflect underlying political priorities on how quickly to complete construction. Performance-based subsidies for child care emerged more as a product of the PBASs than of the environments in which they operated: Once PBASs existed and could distinguish different levels of quality in child care, tiered reimbursement became politically attractive because public funds could be directed toward higher-quality providers and encourage improvement. Such tiered reimbursement allowed for fairly substantial differentials in payment in a way that other incentives could not.

The size and details of rewards vary widely across the PBASs we studied. There is uncertainty about how well the magnitude of rewards is correlated with the benefits of the changes that the PBAS designers seek to induce or the effort that doctors and teachers must make to comply with these changes.

The public-transit PBAS is the only example of performance-based budgeting in our sample. It ultimately suffered the same fate that many performance-based budgeting initiatives have suffered elsewhere—difficulty sustaining consensus on how to measure performance. Chapter Six addresses this problem in more detail.

Chapter Summary

PBAS designers face three basic design issues: (1) determining whose behavior they seek to change (i.e., identifying individuals or organizations to target), (2) deciding on an incentive structure (that broadly includes both the type and size of incentives), and (3) measuring performance and linking these measures to the incentives they have chosen. This chapter addresses the first two issues; Chapter Five addresses the third. In the PBASs we examined, it is fairly easy to identify those who are held accountable for improving service activities and reaching the PBAS goals. It is more challenging, however, to decide which incentive structures to use to affect the desired behaviors.

In principle, the PBAS designers should examine how individuals, units, or entire organizations can affect the goals of the PBAS, then create incentives that are well defined and well linked to activities that each targeted individual or organization can affect or control in some way. To the extent that teamwork is important in such efforts, the designers of the PBAS should avoid targeting individuals, and, similarly, if PBAS designers cannot accurately observe or measure the contributions of individuals, incentives should be defined for the appropriate groups—organizations as a whole, units, offices, teams, and so on—even if changes in individual effort are of greatest importance. Some PBAS designers in our cases followed these guidelines closely. Those that did not provide incentives for groups instead tended to target individuals who held control over some activities relevant to the provider's performance. As we see in the next chapter, this decision has consequences for the way PBAS designers have chosen to measure performance.

We expected context to have a large effect on the incentive structures that PBAS designers choose, and, indeed, context has been extremely influential in most of the PBASs we examined. For example, when participation in a PBAS is voluntary, designers of PBASs used rewards rather than sanctions. When the operators of a PBAS work within a regulatory setting, sanctions were more common. Within the constraints that any context imposes, simple normative guidelines are available to help the designers and operators of PBASs choose the level and distribution of rewards and sanctions in their incentive structures. But we found little evidence that the design of the incentive structures we examined is consistent with such guidelines. It appears that little is known about the factors that would be relevant to implementing such guidelines. Use of training and technical assistance is not as common as normative guidelines would have led us to expect. But PBASs that emphasize such assistance appear well designed to mitigate potential dysfunctional effects of such use.

CHAPTER FIVE

The Design of Measures That Link Performance to Incentives

As discussed in Chapter Four, a key task for the designers of a PBAS is to identify the individuals or organizations that, through strategic changes in behavior, should be capable of improving the quality of delivered services. Next, the PBAS designers must develop (a) a set of incentives to motivate the intended behavioral changes and (b) a set of measures to gauge the performance of service providers and serve as a basis for the application of the incentives. Chapter Four considered the question of incentive structures; in this chapter, we turn our attention to the development of performance measures.[1]

As a general rule, the designers of a PBAS will begin by considering *what* aspects of service should be measured. Next, the designers must consider *how*, specifically, to operationalize the measures. This typically involves identifying data sources, as well as processes for collecting, transforming, aggregating, and classifying the data to generate the measures. Within the academic literature, the terms *metrics* and *measures* are often used to distinguish between the what and the how of performance measurement. Unfortunately, though, these terms are

[1] For a global perspective on performance measurement, see Lynn (2006). Other overviews of experience outside the United States include Curristine (2005), V. Martin and Jobin (2004), Atkinson (2005), and Sterck (2007). For evidence on growth in the use of performance measurement in the U.S. federal government, see Ellig (2007). Evidence from recent surveys on performance measurement in U.S. state and local governments and cooperatives can be found in Brudney, Hebert, and Wright (1999); Governmental Accounting Standards Board and National Academy of Public Administration (1997); Poister and Streib (1999); Moon and deLeon (2001); Moynihan and Ingraham (2004); Wang (2002); Melkers and Willoughby (2005); and Carman (2007).

not applied uniformly across disciplines. To avoid potential confusion, we have therefore adopted the term *measure* to cover both concepts; the specific meaning in each instance should be evident from the context in which the term appears.

PBAS designers face a broad array of choices when designing performance measures. For example, they can choose to do any of the following:

- Assess performance in terms of outcomes, outputs, processes, structural attributes, or inputs.
- Consider external factors that might influence the level of effort that a service provider must exert to achieve any particular level of performance.
- Express measures in the form of a continuous scale, or define categories or thresholds that take on special significance (for example, scores above or below the 75th percentile).
- Define performance in terms of the current value of a measure, or focus on its rate of change over time.
- Use a single measure, a composite measure based on the aggregation of a set of related component measures, or a set of independent measures that are not aggregated.
- Include certain performance measures within the incentive structure, and simply monitor other measures without consequences.

This chapter explores these variations, focusing on two questions relevant to the development of measures for a PBAS:

- What options are available for designing measures to link performance to an incentive structure?
- What factors have been most important in influencing the choices of measures for the PBAS cases that we examined?

Options Available for Defining Measures

Measures Typically Focus on Outputs, Inputs, Processes, or Structures

Normative guidelines for performance measurement typically advise developing measures that are closely related to desired outcomes.[2] In practice, though, it often proves difficult to incorporate outcome measures within a PBAS. In some cases, the intended outcomes—for instance, the kindergarten readiness of a child following a preschool program or the preparedness of a health department to respond effectively to a large-scale public health emergency—are either conceptually difficult or prohibitively expensive to measure. In other cases, the outcomes of interest—such as the degree to which a secondary education program prepares a student to become a productive and engaged member of society—do not become evident until years after the service was provided. For these reasons, many PBASs employ measures based on direct outputs of a service-delivery activity, such as standardized test scores in the education case, that are believed to correlate with the ultimate outcomes of interest.

To the extent that they serve as a proxy for intended outcomes, the use of output measures within a PBAS offers inherent appeal. If, however, the outputs are strongly influenced by external factors beyond the service provider's control, then the application of rewards or sanctions based on such measures might be viewed as unfair. In education, for example, one would more likely expect to see higher standardized test scores at a school that serves wealthier, native English-speaking stu-

---

[2] Again, when studying PBASs, we find it useful to draw a simple distinction between *outcomes* and *outputs*. In our discussion, *outcomes* are long-term goals about which policymakers care; *outputs* are observable, measurable things that a PBAS can link directly to the behavior of the activity it monitors. The literature on program evaluation often refers to outputs, defined in this way, as *intermediate outcomes*. We prefer the simpler distinction used throughout this monograph to highlight the important difference between what PBASs can effectively monitor and what policymakers actually care about. With that caveat in mind, representative examples of normative handbooks that promote alignment of performance measures to outcomes, as we defined them, are Gormley and Weimer (1999) and Hatry (1999). More broadly, such alignment could be offered as the central tenet of the quality movement. See, for example, Levine and Luck (1994) and Kaplan and Norton (1996).

dents than at a school that serves a higher percentage of lower-income or immigrant children, regardless of the quality of the instruction. In a similar vein, it is certainly easier to develop an effective and well-patronized transit system in a densely populated urban core than in a lower-density suburb or rural area.

To overcome the fairness concern, PBASs might instead rely on measures related to service inputs (e.g., education and training of child-care providers), structural elements (e.g., the use of electronic medical records in a hospital setting), or processes (e.g., adherence to accepted standards of care for a particular illness). As we shall see, this is not the only way to overcome concerns about unfairness, but it is a common approach.

Measures Can Rate Performance on a Continuous Scale or Apply Categories or Thresholds

After determining the aspects of performance to be measured, the designers of a PBAS must next decide how to rate a service provider's performance for each measure as a basis for the application of incentives. One option is to rely on a continuous scale—for example, a score that falls in the range of 0 to 100. Alternatively, the rating of performance can be based on categories or threshold values—for example, scores above the 75th percentile for a particular measure might be rated as "good," and scores below rated as "requiring further improvement." Though such decisions might seem to be technical details of minor importance, the detailed structure of the performance measures together with the incentive structure send an explicit message to service providers about where they should target their improvement efforts and how much energy they should expend.

Consider, for example, the development of a PBAS for education based on standardized test scores. One option would be to measure and rate performance based on the average of all student test scores within the school or for a particular teacher. In this case, the improvement of any individual student will also improve the overall score. Teachers will thus be motivated to improve the test scores of all students, though they might choose to focus additional attention on those students whom they believe have the potential for the most signifi-

cant gains. The option employed in NCLB, on the other hand, is to measure the percentage of students whose test scores meet or exceed a particular score designated as representing proficiency. In this case, a teacher would have little motivation to devote much attention to high-achieving students who will easily exceed the proficiency threshold without additional assistance. Nor would the teacher be likely to focus much on very low-performing students who have little hope of meeting the proficiency score regardless of the teacher's efforts. Instead, the teacher would logically expend the most effort on students whose scores are likely to fall at or near the proficiency cutoff mark, in order to increase the overall percentage of students whose scores qualify as proficient. There is evidence of this type of response from large-scale quantitative studies as well as smaller-scale descriptive studies (Booher-Jennings, 2005; Neal and Schanzenbach, 2007; Stecher et al., 2008).

Neither of these options, of course, is inherently right or wrong; such a judgment ultimately depends on the improvement goals of the PBAS. The key point is that the detailed structure of a measure will strongly influence the behavioral response of service providers and should thus receive careful attention in the design of a PBAS.

Measures Can Focus on Current Performance or the Change in Performance Over Time

A related choice that PBAS designers must make is whether a measure should focus on current performance or how the level of performance has changed over time. Imagine, for example, that an incentive structure has been set up to reward the top 5 percent of service providers based on current performance. In this case, high-performing providers would be likely to strive for continued improvement, vying with one another to be ranked in the top 5 percent and qualify for the incentives. Moderate- and low-performing providers, on the other hand, might conclude that they have little chance of making the top 5 percent and thus not try to improve at all. As a counter-example, imagine that an incentive structure provides rewards to the 5 percent of service providers whose performance has improved the most in the past year. Now, low- and moderate-performing providers would certainly have the incentive to strive for continued performance; such providers,

after all, have the greatest opportunities for improvement. Conversely, though, higher-performing providers might determine that there is not much room for further improvement. Both approaches thus have strengths and limitations. Fortunately, the choice need not be exclusive; one could certainly structure a PBAS that measures and rewards both top performers and most-improved performers.

A System Can Link Incentives to a Single Measure, to a Composite Measure, or to Multiple Measures

Whether a PBAS includes a single measure or multiple measures depends on the goals of the PBAS. With A+B contracting, for example, the principal goal is to speed the delivery time for highway construction projects. Thus, only a single measure—days to complete—is needed. In other cases, however, PBAS designers might be interested in multiple dimensions of performance, thus requiring multiple measures. With NCLB, reading and math are both viewed as important and are included separately in the measurement of performance. In PHEP, the assessment of readiness includes a broad range of factors—the ability to mobilize staff on short notice; the ability to marshal and deploy the resources to transport significant quantities of countermeasures through a complex distribution network; the ability to provide adequate security for personnel, sites, and assets; and the like.

If multiple measures are employed but there is only a single set of incentives, it will be necessary to combine the individual measures into a composite measure that provides the basis for applying the incentives. The CAA case provides a relatively simple example. If, on the basis of modeling results, a region's transportation improvement plan is expected to prevent compliance with *any* of EPA's criteria-pollutant measures (e.g., carbon monoxide, ground-level ozone, fine particulate matter), then federal funding for the improvements will be withheld (Bae, 2004). In other cases, the translation of component measures into a single composite measure is more complex, involving various scaling and weighting functions.[3] In such circumstances, PBAS design-

[3] Despite the quantitative look of a composite measure that brings many subordinate measures together, the rules and weights used to define the composite measure often reflect

ers should take care to ensure that service providers understand how the composite measure is derived and are thus in a position to take actions that will improve their scores and, in turn, support the goals of the PBAS.

Finally, it is possible to have multiple sets of incentives that apply to multiple measures of performance, in which case the development of composite measures is not needed. Within an education PBAS, a high-school principal might face incentives related to the graduation rate, while individual teachers might face incentives related to the standardized test scores of students in their classes.

A System Might Employ Measures That Are Not Linked to Incentives

There are at least two possible motivations for creating measures within a PBAS that are not tied to incentives. First, such measures might help determine whether the additional attention that service providers devote to improving performance for incentivized measures is resulting in unacceptable declines in other aspects of performance. Within an education PBAS that provides incentives based on math and reading scores, for example, one could develop nonincentivized measures for such subjects as civics, history, or the physical sciences to see how the quality of instruction in those areas is being affected.

Second, it is possible that service providers might employ strategies (e.g., teaching to the test or even outright cheating) intended to improve the *measure* of performance without corresponding gains in *actual* performance. To detect when this might be occurring, a PBAS can employ nonincentivized measures that evaluate similar aspects of performance through alternate means (e.g., additional math and reading tests in a different format). As we discuss again in Chapter Eight, significant disparities between the scores for incentivized and nonincentivized measures would constitute evidence of a problem that merits further attention and possible adjustments to the structure of the PBAS.

---

strong value judgments with little analytic foundation. When incentivization requires the creation of such composite measures, the importance of understanding the analytic basis—or lack thereof—for the rules and weights that define them obviously increases.

Factors That Are Important in Choosing Metrics and Measures for a Performance-Based Accountability System

Which of these options for developing measures were chosen for the PBASs that we examined, and why? And could the PBAS designers have chosen better approaches? Our analysis reveals six factors that appear to have been important in guiding the selection and structuring of measures within the programs examined in this study:

- feasibility, availability, and cost of measures
- context within which the PBAS operates
- ability to align measures with goals
- degree of control of the monitored party
- resistance to manipulation
- understandability.

Feasibility, Availability, and Cost Considerations Are Paramount

Repeatedly, within the cases that we examined, questions related to feasibility, availability, and cost were important, often critical, factors in the selection and design of measures.

Feasibility. For a measure to be feasible, it must rely on data that either exist or could, in principle, be collected. While this poses some constraints, the question of *when* the measure could be captured creates even greater limitations. The outcomes that a PBAS seeks to affect often unfold far into the future. This effectively deters the inclusion of outcome-based measures within many PBASs, as it does not become possible to measure realized performance until many years after the service has been provided. Even then, it might be difficult to distinguish the service provider's effect on these outcomes from the effects of many other factors in play.

Consider, for example, child care, which presumably influences the future character development, educational attainment, and skill sets of a child. Even if a child-care PBAS seeks to improve these outcomes, it cannot measure, during the present, how the actions of a child-care provider will ultimately affect the children it currently serves. The same can be argued, by close analogy, for education and health-care PBASs.

This challenge is perhaps greatest for PHEP, in which PBASs seek to induce planning and investment decisions today that should improve the response to future public health emergencies that (1) tend to occur only rarely and (2) are highly unpredictable in terms of location, time, type, likelihood, and magnitude.

Availability. In some cases, a PBAS can incorporate measures that have already been defined or captured for other purposes. Many transit-funding allocation PBASs, for example, have made use of statistics that are already reported to the Federal Transit Administration's National Transit Database. In health care, many PBASs have adopted standard quality measures developed by NCQA. While individual PBASs must still collect the data to compute these measures, this approach is viewed as appealing, given that the measures provide broadly accepted benchmarks against which to gauge performance.

If there are no existing measures, a PBAS might be able to draw on existing data sources. For example, PHEP has benefited greatly from tapping into preexisting federal requirements that states and local jurisdictions undertake response exercises; these often provide valuable occasions to collect data on performance (e.g., timeliness of staff notification, mobilization of staff, throughput in dispensing medications during an emergency).

Cost. PBAS designers will typically, and understandably, pay close attention to the costs and benefits of using alternative measures. At least two observations emerge from the cases that we examined. First, designers typically avoid the inclusion of measures that would be very expensive to collect (unless, of course, the measures are already captured for some other purpose). In health care, for example, the most detailed and complete output or outcome measures would require manual review and data extraction from numerous medical charts, and this would be quite costly. Accordingly, many health-care PBASs have instead used less expensive surrogates (e.g., sampling a smaller number of charts, or measuring processes rather than outputs or outcomes).

Second, the designers of a PBAS often choose to include just a small number of measures closely aligned with the central goals of the PBAS. Having a smaller number of measures certainly reduces administrative costs, but there is also a subtler consideration at play. When

the designers of a PBAS consider adding measures to address more-tangential dimensions of performance, the debate among stakeholders over what the priorities of the PBAS should be is likely to grow more intense. This increases the political cost of building and sustaining consensus and support for the PBAS. In education, for example, it has been fairly easy to gain support for measuring student performance in mathematics and reading, as these are widely viewed as key foundational skills. Proposals to measure performance in science, social studies, or other subjects, however, have provoked greater opposition due to less agreement on the relative importance of these other areas.

Institutional Context Strongly Influences the Choice of Measures

As discussed in prior chapters, context plays an important role in the emergence and initial design of PBASs. Here, we focus on how several elements of context have affected choices of measures in the cases that we examined.

Existing Governance and Regulatory Systems. The governance and regulatory systems in place when a PBAS is developed determine the set of issues that can be taken for granted as PBAS designers consider how to measure performance. In education, for instance, a PBAS need not focus on accreditation or certification, as mechanisms are already in place to establish minimum standards for educational facilities and teacher education. Similarly, A+B contracts can focus on performance in terms of the time required to complete construction and simply reference detailed regulatory requirements related to design, engineering, work safety, and traffic safety.

Sectoral Experience with Performance Measurement. From the cases we examined, a strong tradition of performance measurement within a sector does not appear to be a prerequisite for developing a PBAS, but it can certainly influence the design of measures. In particular, a preexisting culture of performance measurement creates an opportunity for the identification and adoption of more-suitable and sophisticated performance measures, though this opportunity is not always seized.

In health care, there is a strong tradition of applying scientific findings to shape clinical practice. In general, PBASs within health

care have adopted evidence-based performance measures linked to desired outcomes. As new empirical information becomes available, those responsible for implementing health-care PBASs have often revised or replaced measures. Early efforts to promote accountability in health care focused on assessing structural features associated with access to care, such as the number of beds or providers in relation to the size of the population or access to after-hours care. As the evidence base regarding the factors associated with reduced morbidity and mortality in health care expanded in the past three decades, the focus of accountability efforts shifted toward measuring processes and outputs associated with improved patient outcomes.[4] Examples include the use of beta-blockers after a heart attack to prevent reoccurrence and the maintenance of appropriately low blood-sugar levels in diabetic patients to reduce such complications as blindness and foot ulcers and infections that can lead to amputation.

Education also has a strong tradition of measuring performance, but mainly with respect to gauging student achievement through the use of standardized tests. In contrast, there is much less experience with, and technical capability for, measuring the performance of specific teachers or administrators in a more nuanced fashion. Indeed, many education providers oppose efforts to base incentives on measures of their own practice, arguing that are no adequate measures of instructional practices or management activities or that it would be difficult to control for such external factors as the demographic characteristics of the student body. The limited attention to teacher-level measures is likely to change, however, as a result of recently funded programs in the U.S. Department of Education. Several states have

[4] As noted earlier, the program-evaluation literature often prefers the phrase *intermediate outcome* or *intermediate output*. During his review of this monograph, Harry Hatry reminded us that, when an organization succeeds in using an incentive to induce a specific effect, it should be rewarded and encouraged by framing that effect as an outcome, not just an output. We do not reject that perspective. But our analysis has identified unintended consequences almost everywhere PBASs are applied. As a result, we prefer to encourage PBASs to monitor themselves, learn from their experience, and adjust the effects they monitor in pursuit of outcomes that they cannot directly control. To do that, we prefer to focus on outputs, thus reminding designers that they can choose which things to observe and measure and adjust that choice as needed in their ongoing pursuit of their real goals.

received funding to implement growth-based measures of achievement, and the Race to the Top initiative creates incentives for states to calculate growth or value-added measures of student achievement at the individual teacher level.

Transportation is another sector with a strong tradition of measurement. Indeed, the research literature discusses many hundreds of possible measures for evaluating various attributes of transportation facilities and services, and many of these are routinely tracked by local, state, or federal transportation agencies (see, for example, Cambridge Systematics, 2000). Generally speaking, however, many of the measures used in transportation are rather narrow in physical or conceptual scope, tracking such issues as the pavement quality on a specific stretch of road or the number of riders served by a transit system. Much less common are measures that attempt to relate the performance of the transportation system to broader social goals, such as support for economic activity or the facilitation of accessibility within an urban environment. Further, no firm consensus has emerged on how to use available measures to guide the allocation of funds among transportation systems on the basis of performance. In the absence of such a consensus, PBASs that allocate funds have had little success sustaining the use of performance measures over time. In fact, the idea of using performance measures to allocate funds among jurisdictions runs contrary to the traditional application of funding formulas that place a premium on notions of geographic equity—that is, formulas that are designed to ensure that each jurisdiction receives its "fair share" (B. Taylor, 2004).

In contrast, there was little experience with performance measurement in child care prior to the development of PBASs within the sector. Indeed, the implementation of these PBASs appears to have created a new demand for performance measurement. To this day, the selection of specific measures in child-care PBASs continues to evolve in response to professional beliefs about what constitutes high-quality child care, though there is little evidence to support a professional consensus on which particular measures to include in PBASs to promote desired outcomes.

Understanding of the Service Production Process. As discussed in Chapter Three, there is a clear distinction in the degree to which

the designers of a PBAS understand in detail the inputs, structures, and processes through which services are produced or provided; there is also variation in how this understanding has been used to select performance measures. Some PBASs are designed to incentivize fairly specific changes in the service production process, suggesting that the PBAS designers are quite knowledgeable about the inner workings of the service activity and hold strong views about the types of modifications that should lead to improvement. Other PBASs focus solely on service outputs, devoting little if any attention to the specific manner in which improvements are achieved.

Despite differences in the level of understanding about relevant production processes, PBASs in both child care and health care tend to emphasize measures of inputs or processes over measures of outputs. In the case of child care, there is little in the way of formal evaluation research on which to draw. According to the empirical evidence that does exist, however, professionals in the field have developed consensus beliefs about the types of processes—for example, the use of a formal curriculum or low child-staff ratios—that should lead to better outcomes for the child, and most child-care PBASs employ measures that assess quality in terms of these processes. For health care, there is a greater wealth of empirical evidence on the effects of different courses of patient treatment and the benefits of certain infrastructure investments, such as converting medical records to electronic format. Healthcare PBASs thus tend to highlight specific approaches to patient care and employ measures that assess performance in terms of compliance with the highlighted approaches.

In contrast, the PBAS examples drawn from the education and transportation sectors tend to focus on measures related to outputs rather than inputs, structures, or processes. Most education PBASs, for example, measure student test scores, not the conditions or methods used in the classroom. The incentives in A+B contracting are based solely on the time required to complete a project: Provided that the contractor complies with relevant safety regulations and engineering standards, the contract is not concerned with the management techniques used in a project (Strong, 2006). CAFE standards, likewise, focus on the measured fuel efficiency of vehicles sold by a manufac-

turer, not on the design and engineering methods used to achieve the fuel-efficiency goals (T. Martin, 2005).

Although knowledge about the internal workings of the service-delivery processes can shape the measures chosen by PBAS designers, other factors, such as feasibility, often appear to be more important. For example, in child care, there is little consensus about how to measure such outcomes as kindergarten readiness in a cost-effective manner, so it is much easier to assess providers on the basis of adherence to particular processes guidelines. In the case of health care, although many of the outputs of interest—for instance, the success or failure of certain procedures—can be readily observed and tracked, providers are understandably concerned that such outputs might be strongly influenced by factors beyond their control. Patients who are in poorer health to begin with, for example, will reduce the likelihood of successful outcomes, regardless of the quality of care.

Designers Seek to Align Measures with System Goals, Though This Often Proves Challenging

For all of the cases we examined, it appears that the designers of the PBASs have striven for measures intended to align service activities with the goals of the PBAS. In particular, PBAS designers have adopted measures that cover a relatively small number of "essential" goals. There are at least two reasons for this choice. First, reducing the number of performance measures reduces the cost of implementing the PBAS, for both administrators and service providers. Second, focusing on just a small number of goals enables service providers to concentrate their efforts on the goals viewed as most important rather than diffusing their effort against a much broader set of objectives.

Within the health-care domain, for example, the Institute of Medicine (IOM) has outlined a performance-measurement framework that includes six key constructs often used as a basis for selecting PBAS performance measures: (1) safety, (2) timeliness of and access to care for patients, (3) effectiveness, (4) equity, (5) efficiency, and (6) patient-centered care, rated in terms of the quality of doctor-patient interaction and coordination of care. Some health-care PBASs are now evolving to include measures of cost-efficiency, a goal that is receiving increased

attention in light of the spiraling costs of medical care. Often, this appears in the guise of "value-based purchasing," which encompasses both quality and cost components in the accounting of performance.

In education, similarly, almost all test-based accountability systems include measures of reading and mathematics. NCLB, for example, includes just three output measures: graduation rate (for high schools) and reading and mathematics test scores (grades 3–8 and high school).[5] Most P4P systems in education also focus on reading and math achievement, though some include other subjects as well.

Turning now to the challenges, it can be difficult to align measures with essential goals unless those goals are well defined, consensus-based, and, preferably, quantitative. Of the PBAS cases that we examined, A+B contracting has the most clearly stated goal: reducing the time required to complete road construction projects. A construction job is complete when government engineers certify that it is ready for unrestricted public use; on that date, the government and contractor can agree on calendar time to completion, and they can then use a simple formula to calculate the final payment for the job. But things are not so simple elsewhere. Some child-care advocates, for example, argue that child-care programs should focus on academic preparation for kindergarten, while others are more concerned with children's physical, social, and emotional development.

In sectors that confront multiple, and often competing, goals, the cases we reviewed suggest that PBAS designers will typically adopt measures that address just a subset of these—specifically, as noted above, those viewed as most essential. However, including measures that address just a subset of relevant goals can also lead to unintended consequences. To the extent that what is measured reflects a narrow set of performance markers, the net effect can be an increased emphasis on what is measured, with considerably less attention devoted to other relevant concerns. For example, several decades of research in education reveal widespread unintended consequences stemming from formal testing in just a few subjects, including reallocation of time and effort

[5] Science testing is now required as well, after a two-year delay, but is not part of NCLB's accountability index or the reward structure.

from untested subjects to tested ones. Analogous effects occur in health care, in which some PBASs increase emphasis on measured areas and ignore unmeasured ones.[6] In PHEP, relatively strong emphasis on performance measurement for mass medical countermeasure delivery has led state and local health departments to invest in those capabilities at the expense of others. In some cases, the unintended consequences can prove pernicious. For example, measures in a health-care PBAS that might induce physicians to overprescribe antibiotics for certain illnesses could cause drug-intolerant strains of the bacteria to form.[7]

The occurrence of unintended effects does not mean that we should judge the PBASs in question as failures, but it does point to the importance of monitoring measures and refining them if they are not providing accurate information or if they are having negative effects. Provided that the incentive structure used within a PBAS is sufficient to induce changes in the behavior of service providers, any unanticipated disparities between the goals of the PBAS and the measures it employs will induce misalignments in provider effort. Given the complexity of the environment in which many PBASs operate, such disparities are common—perhaps inevitable. To make a PBAS as socially useful as possible, it is thus desirable to build in the capacity to observe the effects of its operation over time and update or refine the measures as experience dictates. Unfortunately, the political environments in which many PBASs operate can make such continuing adaptation and realignment difficult. Chapter Six addresses this point in greater detail.

---

[6] This might be appropriate if the measures focus on critical deficit areas. In such circumstances, however, the measures should shift over time as certain deficit areas are addressed and others emerge.

[7] An incomplete application of an antibiotic can, in effect, screen strains of bacteria, killing the vulnerable ones in a patient and allowing the less vulnerable ones to persist and spread as drug-resistant strains. Such a strain might not hurt the patient immediately treated; in fact, treatment might be incomplete because the patient required no further antibiotics to recover. But, over time, the strain can threaten other patients more sensitive to the strain, patients now unable to use the antibiotic to which the strain has developed resistance. If a treatment standard does not consider this secondary concern, which looks beyond any individual patient in immediate treatment, preventing development of drug-resistant strains can become a "forgotten goal."

Designers Seek Measures That Service Providers Can Either Control or Strongly Influence

PBAS designers strive to develop performance measures that service providers, through their efforts, can either control or strongly influence. There are two arguments for this. First, if a PBAS offers rewards to a provider based on positive results that would have occurred even without the provider's effort, it can be viewed as an inefficient allocation of resources. Second, if a PBAS either withholds rewards or imposes sanctions based on negative results falling beyond the control of the service provider, the service provider will view it—quite reasonably—as unfair. Over time, the repeated occurrence of either of these outcomes is likely to generate resistance to the PBAS, leading to either its modification or its elimination.

For PBASs in education, health care, and transit-funding allocation, the question of control can sometimes be problematic. In education, teachers, schools, and even school districts are ultimately limited in their ability to affect student test scores. Many of the factors that influence student capabilities—socioeconomic status, previous educational experience, and so on—fall beyond the control of teachers and schools. Further, although schools and teachers can help students learn, they cannot control the learning process directly. Other factors, including parental and peer attitudes and the personal commitment of the students, are vitally important. In such circumstances, measuring the performance of teachers, schools, or school districts on the basis of a single year's test scores, without regard to the background of the students and families being served, can be viewed as unfair. Indeed, attaching incentives to such measures can induce undesired and unintended consequences, as discussed in Chapter Four.

Closely analogous circumstances arise in health care. The health status of patients depends on many factors beyond the control of medical providers, including genetic background, socioeconomic status, and previous medical treatment. Doctors and hospitals are able to affect the health status of these patients only from the first day they see them forward. Even then, as in education, the personal commitments of patients and those close to them have a large effect on their health status. Thus, holding doctors and hospitals entirely accountable for

the success or failure of medical treatments can likewise be viewed as unfair. One possible response among service providers, logical though clearly undesired, would be to refuse treatment to those patients at higher risk of poor outcomes.

It has also proven challenging to develop performance measures for transit systems that treat jurisdictions equitably. Consider, for example, the differences between rural and urban transit systems. In rural areas, low-density development patterns make it extremely difficult to attract riders in sufficient concentration to develop an extensive and well-patronized service. In contrast, urban areas, with much higher population densities and a greater concentration of workplaces and recreational destinations, are much more amenable to using effective transit service. It should not be surprising, then, that urban transit systems carry far more riders, and at much lower cost per rider, than their rural counterparts. Under these circumstances, crafting performance measures that treat urban and rural transit systems equitably has proven an elusive goal.

PBAS designers have several options for addressing such concerns about unfairness. One option is to adopt performance measures based on processes (or inputs or structures) that service providers can directly control rather than on outputs or outcomes subject to external influence. Such has been the approach in many health-care PBASs, which might assess such factors as the timely delivery of preventive-care services (e.g., mammography screenings) or the provision of appropriate medication for the patient's clinical condition.

A second option is to develop output measures that assess the *change* in performance over time rather than the *absolute* performance in any single time period. In an educational PBAS, for example, teacher performance might be measured based on improvements in current-year test scores of a cohort of students versus those students' scores in the prior year. This would help to isolate a teacher's current contributions from the quality of instruction that his or her students received in the previous year.

A third option is to adjust or scale output measures to account for relevant factors (e.g., educational capability, health status) that lie beyond service providers' control. In education, for example, a PBAS

might examine the characteristics of students in a school and judge the performance of the teachers and administrators based on how their students' test scores compare with the average scores for students from similar backgrounds. Similar methods have been applied in health-care settings (particularly when measuring mortality, readmission to hospital, and patient-experience ratings) to control for differences in patient characteristics (e.g., race/ethnicity, language spoken, age, severity of illness, presence of other comorbidities) across different hospitals or doctors.

A variation on this option might set different goals for activities when they serve populations with characteristics that impose different burdens on these service activities. Using the examples above, goals might differ for activities serving urban, suburban, and rural populations or for activities treating different kinds of diseases or patients. In practice, efforts of this kind to promote equity by reflecting differences in costs imposed on the activities that PBASs monitor have often led to charges of inequities among different populations served: Setting a lower performance goal for an urban-based activity can easily be perceived as setting a lower performance goal for the urban population it serves. When a PBAS makes such distinctions, it must resolve such conflicts. For example, it might set the same performance standards for all students but set differing goals for the amount the schools spend to achieve these standards in different settings.

A fourth option is to conclude that it is either impractical or infeasible to develop measures that treat service providers equitably and, in response, simply back away from the use of performance-based accountability. It appears that this is what has occurred in many of the transit-funding PBASs that we examined.

All of these options are motivated by the assumption that PBASs should treat service providers equitably. And, indeed, this might prove necessary to garner their support, when necessary, for adopting a system. A counterargument can be made, however, that it is those who benefit from the services, not those who provide it, who should be treated fairly and equitably. This leads to a fifth option. Rather than adjusting measures to promote greater equity among service providers, PBAS designers can adopt output measures that hold all providers

to the same standard so as to promote greater equity among service beneficiaries. This was the approach taken with NCLB, which holds all schools and districts to the same target for academic proficiency regardless of differences in the composition of their student bodies. It also applies to child-care PBASs, based on the argument that all children, regardless of their characteristics, deserve high-quality care.

Designers Seek Measures That Are Resistant to Service Providers' Manipulation

Just as PBAS designers seek to develop performance measures that can be positively influenced through the efforts of service providers, so too do they strive to create measures that are difficult for service providers to manipulate in ways that are incompatible with the PBAS's goals. Such manipulation usually occurs when the set of adopted measures does not completely cover the full range of relevant sectoral goals within a PBAS. As noted in the previous section, limits on the feasibility of measurement and cost considerations typically prevent the creation of a comprehensive set of measures, so some level of manipulation will usually occur. Unless a PBAS is able to perfectly align its service providers' goals with its own goals, the service providers will, in all likelihood, "manipulate" the PBAS rules to promote their own well-being. The PBAS designer should expect that the service providers will always pursue their own interests, possibly to the detriment of PBAS sponsors and designers.

Perhaps the best-known example of manipulation in the cases we examined is the act of teaching to the test in an educational setting. While ethical teachers with knowledge of the contents of a standardized test do not reveal specific questions to their students, they are still likely to adjust their instruction to ensure that students are well prepared for the questions that they will likely face, given the learning standards that the test emphasizes. Several decades of research in education reveals widespread use of coaching and practice on specific item formats, along with efforts to manipulate the test-taking population to maximize scores. Such actions lead to score inflation, a phenomenon in which student performance on standardized tests overstates

the students' mastery of the underlying subject matter.[8] Of course, not all teaching to the test reduces the validity of scores, and, clearly, some degree of focus on these measures is consistent with the goals of test-based accountability systems. Types of test preparation can be placed on a continuum ranging from desirable (e.g., higher-quality presentation of the relevant content) to undesirable (e.g., cheating—see Jacob, 2002), with some ambiguous forms in between (e.g., coaching on specific problem types; see Koretz and Hamilton, 2006, for a discussion of each of these responses).

In the transportation sector, efforts to include performance measures in transit-funding formulas have induced comparable service-provider efforts at manipulation. In cases in which a formula includes total ridership as a proxy for system performance, for example, some local transit operators have lowered their fares in order to enhance ridership and in turn increase their share of regional public funding—in effect, boosting subsidies at the cost of self-sufficiency. Other transit operators have shifted some service from off-peak to peak hours, when it is easier to gain additional riders, leading to greater total ridership but eroding the quality of service for those who must rely on transit service in the middle of the day or late at night.

Comparable, if somewhat more pernicious, effects have also been observed with health-care PBASs. Some health-care providers, for instance, have sought to improve their performance scores by avoiding less attractive patients, such as those having more-serious conditions or those viewed as less likely to comply completely with doctor directions.

While such manipulation provides an obvious cause for concern, evidence of manipulation does indicate that service providers *are responding* to the measures and incentives as currently structured. In other words, manipulation indicates that the broad mechanisms of a PBAS are working as intended, even if the specific structure of the performance measures requires further refinement. And manipulation, viewed broadly, need not lead to undesired effects. Within child-care PBASs, for example, it has been observed that providers often seek to improve their ratings through the least costly options, such as taking

[8] For evidence on and a discussion of this effect, see Koretz (2002).

steps to involve parents to a greater extent as opposed to reducing child-staff ratios. Few child-care stakeholders have expressed alarm over this response. Rather, they have taken the perspective that the PBAS is succeeding in changing provider behavior in the way that it was intended to. Further, they recognize that, if funds are not made available within the PBAS to improve the more costly quality components, such as ratio reduction and staff education, providers have little choice but to focus on the less costly options.

With the potential for manipulation in mind, PBAS designers strive to create a set of measures and incentives that aligns the service providers' interests with the interests of the PBAS—that is, that leads to actions that support rather than undermine programmatic goals. To limit manipulation viewed as undesirable, PBAS can first select measures intended to induce behavior that is compatible with the PBAS's goals. Next, those charged with overseeing the PBAS can carefully audit service-provider response to the measures to detect and characterize any manipulation that might be occurring. With NCLB, for example, comparing student scores on tests that determine a school's adequate yearly progress (AYP) under NCLB with scores by the same students on tests of similar subject matter that have no direct effect on AYP has helped to detect the presence of NCLB-induced score inflation. When it is determined that the measures used within a PBAS have led to undesired manipulation, the measures can be adjusted in specific ways designed to discourage this behavior. With time, PBAS operators progressively learn how to refine the PBAS measures to advance their goals. A PBAS is thus more likely to succeed in the long run if it can maintain flexibility and adaptability in the performance measures that it applies.

Service Providers Want to Understand How Measures Reflect Their Interests and Are Influenced by Their Behaviors

Just as PBAS designers seek measures with certain characteristics (e.g., predict desired outcomes, are hard to manipulate), service providers prefer measures that are clear, verifiable, and operate in a manner they can understand. This perspective is consistent with the broader goal of transparency (i.e., clarity and openness) in governance, designed to

promote fairness and due process in the public sector. In the context of a PBAS, service providers want to understand how the measures that are chosen and the decisions that are made reflect their interests and are influenced by their behaviors.

In the PBASs we examined, service providers believed that better information would help them understand how changes in their behaviors would affect the rewards and sanctions they received and would enhance their ability to improve performance. Yet, the understandability of the measures did not appear to be as important in selecting measures as some of the factors mentioned in previous sections. To some degree, it appears that service providers tend to "muddle through" their internal planning process by inferring from past experience how changes in their behavior are likely to affect future rewards and sanctions.

For example, in child care, although some component factors used to develop ratings, such as child-staff ratios and staff credentials, are easily understood, the translation of these concepts into ratings is not always clear. Higher ratings can be attained for having a particular ratio of staff with bachelor's degrees or for having a particular ratio of staff with an associate's degree in early-childhood education (such as a child-development associate's, or CDA). However, it is not clear whether it is more advantageous to hire a few bachelor-of-arts (BA)–level teachers or to spend the same funds to help a larger group of staff achieve CDA credentials. High levels of staff turnover complicate provider decisions of this kind, because child-care providers cannot predict how much they will benefit from their investments in staff. While some PBASs are clear about how such decisions affect aggregate ratings, others are less clear, making it difficult for a typical provider to anticipate how changes in its staffing behavior will affect the aggregate score it ultimately receives from the PBAS.

We suspect that concerns about understandability of measures will be greater where PBASs combine multiple components into a single rating, as is the case with the AYP designation under NCLB. Educators complain that failure to attain any of more than 40 criteria (depending on the number of subgroups present in the school) can cause the school to miss its AYP target. Yet, districts and schools have become adept at

verifying their status by replicating the AYP calculations on the basis of their reported test scores despite the complications. These computations are usually too complex for individual teachers to replicate. But, over time, as teachers see what is being asked of students on tests, they can figure out what they need to emphasize and what they might be able to deemphasize in order to improve test scores.

In transportation, the models used to predict future emissions that serve as the basis for judging compliance with the CAA are quite complex (Johnston, 2004). It is difficult to predict, prior to running the model, how a particular configuration of planned transportation improvements will affect emissions. As a result, the process might involve some trial and error as planners seek a set of improvements that (1) address the region's transportation needs and (2) simultaneously meet the necessary emission targets.

Attributes of Measures Chosen and Factors Influencing Choices

Table 5.1 summarizes key attributes of the measures used by the PBASs we studied, as well as important contextual factors that influenced the choice of measures. Even a cursory review of the table (particularly the factors that influenced the choice of metrics) highlights the huge variation that exists across the PBASs we studied. Each case is different, and their distinctive contexts seem to affect both the selection and application of measures.

One notable commonality can be seen in the first column of the table; most of the PBASs emphasize the measurement of proximal outputs rather than more-distant outcomes. This observation reflects a fundamental challenge of PBASs that seek to regulate services to the public—the PBAS goals embody aspirations for the future that are both impossible to attain and impossible to measure in the short term. The challenge in designing a PBAS is to identify measurable short-term outputs whose attainment will signal movement toward those long-term aspirations. It should also be pointed out that the health-care P4P adopted measures based on process to ensure a close link between the

performance measured and the behavior of the doctors monitored. The child-care and PHEP PBASs used measures based on processes and inputs because it is too costly to measure changes in young children's knowledge, skills, and behaviors, and it is infeasible to measure outputs relevant to PHEP.

Our cases include threshold, categorical, and continuous measures. Child care uses the only measure defined in terms of ordinal categories (one, two, or more stars); these metrics maintain simplicity, particularly when conveying information to parents. Thresholds suit the CAA PBAS because its mission is to enforce underlying regulations that employ the same thresholds. NCLB uses thresholds to focus attention on achieving certain minimum standards for all students within a state. Continuous measures naturally fit the goals expressed in the other PBASs.

Our sample also displays variety in the complexity of the measures. The simplest one, used in A+B contracting, works well because critical elements of the governance structure beyond the PBAS were already in place when A+B contracting was devised. The NCLB, CAA, and CAFE PBASs all use relatively small numbers of measures to promote ease of understanding. The current version of the PHEP PBAS also uses small numbers of simple measures, but this could easily change as the system matures. The remaining PBASs use less transparent methods to aggregate measures and trigger rewards or sanctions for the parties they monitor.

None of these PBASs appears to collect data on performance measures that are not incentivized. As noted earlier in this chapter and in Chapter Eight, such audit measures can be used to search for unintended effects. In other contexts, public and private organizations often make use of nonincentivized measures (displayed in the form of scorecards and dashboards) to improve their understanding of their own internal operations and their relationships with strategic partners.[9] Such arrangements do not appear to have emerged in any of the PBASs we examined, and it might be interesting to examine further

[9] See, for example, the series of books on the Balanced Scorecard by Robert S. Kaplan and David P. Norton. The most recent is *Alignment: Using the Balanced Scorecard to Create*

Table 5.1
Attributes of Performance Measures Chosen and Factors Influencing Choices

Sector/PBAS	Emphasis of Measurement	Use of Thresholds	Level, Rate of Growth	Complexity	Factors Influencing Choices
Child care, QRISs	Processes, inputs	Some categorical, some continuous	Level	Aggregate	Ill-defined output Professional consensus on specific best practices Costly monitoring Desire for simple information on level of quality
Education, NCLB	Outputs	Thresholds	Levels	Multiple	Professional consensus on testing students to measure outputs Costly testing Emphasis on helping lower performers
Education, P4P	Outputs	Thresholds, continuous	Level, rate of growth	Aggregate	Professional consensus on testing students to measure outputs Costly testing Emphasis on helping lower performers Limits of teacher's control over measurable outputs Need to integrate PBAS incentives with preexisting personnel-management processes
Health care, P4P	Process	Continuous	Level	Aggregate	Professional consensus on desired standards of care Costly monitoring Limits of doctor's control over measurable outputs Need to integrate PBAS incentives with preexisting personnel-management processes
PHEP	Processes, inputs	Unknown	Levels	Multiple	Infeasibility of measuring outputs Consensus on relevant processes and inputs still slowly forming Desire for simple information on levels of capacity that are considered relevant

Table 5.1—Continued

Sector/PBAS	Emphasis of Measurement	Use of Thresholds	Level, Rate of Growth	Complexity	Factors Influencing Choices
Transportation, A+B	Output	Continuous	Level	Single	Level of calendar time continuous, easy to monitor Given structure of governance for contracts, only one metric required
Transportation, CAA	Output	Thresholds	Level	Multiple	Political consensus on multiple acceptable ambient air emission thresholds Lack of consensus on how to integrate No simple way to value noncompliance Ambiguity in identification of parties responsible for noncompliance
Transportation, CAFE	Output	Continuous, threshold	Level	Multiple	Political consensus on two desired fuel-efficiency levels and methods in place to measure efficiency Ease of measuring continuous degree of noncompliance
Transportation, transit	Output	Continuous	Level	Aggregate	Ultimate emphasis on equity, not performance Professional consensus that levels of many continuous metrics affect equity Need to integrate to allocate funds to budgets

why this has occurred and what the benefits might be of moving in that direction.

Among the factors that influence the choice of measures (final column), professional consensus plays an especially important role. In most cases, considerable latitude exists on what to measure. Analytic evidence on what matters in child care is limited. Most agree that reading and math skills are important in education, but many also believe that history, civics, and science are as well. Even in health care, the field that views itself as the most heavily science-based, professional judgment about which practices deserve the greatest focus has changed repeatedly and is expected to change as the state of knowledge in the field changes. PHEP is a young field still feeling its way toward real agreement on what matters. And, in the most extreme case, the utter failure to achieve and sustain any professional consensus on what external factors affect cost and performance in transit systems ultimately doomed the success of the PBASs attempted in this sector. When knowledge about what really matters is limited and a PBAS must focus its measurement on a short list of outputs to move forward, broader professional consensus helps the sponsors achieve and sustain consensus among themselves and convey the legitimacy of that consensus to those that the PBAS monitors.

Chapter Summary

Performance measures are used to quantify performance for the purposes of implementing the incentive structures that PBASs use to change behavior. These measures can vary in many dimensions, and the measures used by the PBASs we examined display much of the variation that is possible. This variation arises because PBAS designers balance a number of factors when choosing measures, and the factors they appear to consider differ widely across the PBASs we examined.

Corporate Synergies (Kaplan and Norton, 2006). The Palladium Group maintains an online community of interest on this topic (Palladium Group, undated).

Six factors appear to be especially important in the cases we examined:

- the feasibility, availability, and cost of measures
- the context within which a PBAS operates
- the alignment of measures with PBAS goals
- the degree of control of the monitored party
- resistance to manipulation by the monitored service activity
- understandability.

PBAS designers can include only things they can actually measure, and they seem to prefer measures that can be collected at low cost or that already exist external to the PBAS. The context within which PBASs operate significantly narrows the range of low-cost measures that are acceptable. To choose among potentially acceptable measures, a PBAS tends to balance two major considerations—the alignment of a measure with the PBAS's goals and the extent to which a party monitored by the PBAS has the ability to control the value of that measure. A natural tension arises from efforts to achieve balance between these objectives. Over time, the parties that a PBAS monitors might find ways to game the system, increasing their standing on a measure in ways that are not aligned with the PBAS goals. Continuing vigilance and flexibility can help a PBAS to manage this tension and maintain the balance between its priorities and the capabilities of the parties it monitors. Such a balance tends to be easier to achieve when the measures the PBAS uses are clear and open and the parties it monitors understand how their behavior affects the values of the measures. Such transparent metrics help avoid behaviors that are not of mutual advantage of the PBAS and the parties it monitors.

Consideration of these six factors seems to explain the choices of measures in the PBASs we examined. Further, improvement in any of the PBASs would have to reflect improvement in the balance among these factors. This suggests a constructive method for seeking improvements, i.e., testing the balance among these factors for potential weakness. Our own attempt to do exactly this with the PBASs we have examined suggests only a few opportunities for improvement.

CHAPTER SIX

Implementation and Monitoring

Between the design of a PBAS and its full operation, there is a period of implementation when a number of important activities must take place, including establishing rules governing how the system will operate in its particular setting, creating systems to collect data to support necessary measures of performance, installing incentive structures, designing and operating a data reporting system for providers, and possibly creating an audit system to monitor the ongoing functioning of the PBAS. Because these implementation decisions will affect the operation of the PBAS, they deserve attention. This chapter examines the implementation process associated with a PBAS, including both the steps needed to bring the system into operation and procedures to gather data for "formative evaluation," which allow PBAS sponsors to improve the operations of the ongoing PBAS.

The nine cases we studied vary considerably in terms of implementation. In education and health care, PBASs have existed for a decade in a number of different forms. Similarly, in transportation, a long history of performance measurement exists, although tying performance explicitly to financial incentives is still rare. PBASs in PHEP are fairly new; the accountability provisions of PAHPA were rolled out in 2009. PBASs in child care are also recent: As of March 2009, only 19 states had implemented statewide QRISs. The oldest is ten years old; however, most are much more recent.

Policymakers are often impatient to see their efforts bear fruit in a form that is recognizable to constituents, and nowhere is this impatience more apparent than in policy implementation. The desire to

have policy change occur rapidly can result in giving the implementation process short shrift, despite the importance of this process to the ultimate success of a PBAS. Even a well-designed PBAS might not yield the desired results if it is not executed effectively. Because very little research has been done on the PBAS implementation process, we cannot provide much empirical evidence about how the process works in practice. But we believe that the implementation phase is important enough to the eventual success of a PBAS that we have mined both the implementation literature and implementation experiences in our cases to lay out some of the problems that have emerged, and we provide lessons learned where we can.

This chapter addresses the following questions:

- What are common pitfalls in implementing a PBAS?
- What are some potential strategies by which to address these pitfalls?

We explore issues involved in building the system (i.e., operationalizing the design of the PBAS) and in communicating with PBAS stakeholders, including the service providers held accountable by the PBAS and the clients or users intended to benefit from the PBAS. Communication often plays a key, yet undervalued, role in determining the success or failure of a PBAS. Additionally, we consider how formative monitoring can be used to update the system based on early feedback in order to address problems early and fine-tune the PBAS so that each element of the system works efficiently and as intended.

Common Pitfalls in Implementing a Performance-Based Accountability System

PBASs can take a number of years to set up, depending on the context and the complexity of the PBAS. There are a number of pitfalls that can arise during this process. Understanding what to expect and preparing to address these issues is critical to ensuring the long-term

viability of the PBAS. Our review of the literature and the nine cases identified five key pitfalls that can affect the implementation process:

- lack of PBAS experience and infrastructure
- unrealistic timelines
- complexity of the PBAS
- failure to communicate
- stakeholder resistance.

Lack of System Experience and Infrastructure Can Pose Operational and Capacity Challenges

The ease of PBAS implementation depends on whether it is possible to build on an existing infrastructure or whether it is necessary to put new architecture in place to support the activity of the PBAS. As discussed in Chapter Five, choosing which measures to include in a PBAS might be easier when there is a history of measurement in the sector. Similarly, once the design phase is complete, implementing the system might also be easier when there is an established history of performance measurement, which can provide structures for collecting, tabulating, and reporting the measures. On the other hand, where measurement is unfamiliar, not only is it more difficult to design appropriate structures but the system is also less likely to get it right the first time. Under those circumstances, significant effort and resources might be required to create the data definitions, collection instruments, summary reports, and other system elements, thus stretching the capacity of those responsible for building the PBAS.

Given the complexity of some PBASs, operational and capacity challenges can occur even when there is a history of performance measurement. For example, as described previously, NCLB required states and school districts to adopt new standards, align tests, set performance standards, enumerate targets for annual improvement, create systems for support, and construct a parallel system of English-language proficiency standards and tests for students with limited English proficiency. By 2007, the state systems were operating largely as planned: collecting and analyzing data, making judgments about schools, and

providing support systems, with mandated interventions taking place (J. Taylor et al., forthcoming).

However, there were many operational challenges. For example, many states had to create systems and procedures where none previously existed (e.g., add tests in grades in which students were not previously tested, aggregate data in ways that were not previously done). Similarly, some states had to modify accountability systems they had spent years developing. In some cases, laws had to be changed or long-standing procedures altered (e.g., determining which subjects were to be tested or which student demographic information was to be collected). Communicating the changes to district administrators, school principals, teachers, parents, and students was also a major task. There were also capacity challenges; in many cases, financial resources were insufficient to meet the demands (e.g., district budgets were strained by requirements for additional parental notification), human resources were lacking to develop the procedures (e.g., state departments of education did not have psychometric expertise to address the requirements for testing limited-English-proficient students), or knowledge was insufficient to provide the desired assistance (e.g., states did not have the experience turning around low-performing schools to design effective systems of school support).

Unrealistic Timelines Can Create False Expectations Among Stakeholders

Cases from transportation and child care highlight another common problem with PBASs: unrealistic implementation timelines established by policymakers. For example, in transportation, although CAFE standards were enacted with the EPCA and originally slated to apply to vehicles beginning in model year 1975, implementation was delayed for several years (until 1978 and 1979 for cars and trucks, respectively). This was due in part to a failure to recognize the amount of time that it takes for auto manufacturers to plan, develop, and build vehicles that would meet the standards. In retrospect, the goal of applying CAFE standards a year after passage of EPCA appears naïve. Similarly, in child care, the California QRIS was originally supposed to be fully designed within one year, despite the sector's relative inexperi-

ence with quality measurement and reporting. Program planners managed to convince lawmakers to extend the timeline to 24 months, but, well into the second year, many stakeholders saw that two years is not nearly enough time to assess and align existing infrastructure, develop a rating scale, and design a funding model.

Establishing hurried or unrealistic timelines can set a PBAS on a course for failure even when the program is making good progress in changing the status quo. While accelerated deadlines might serve to force action, especially from reluctant service providers, failure to provide adequate time for implementation might lead to discouragement. The result might be undeserved criticism from policymakers, who expected that the PBAS would be able to meet the timeline, or reduced morale among service providers, whose best efforts to implement the PBAS still do not result in success as defined by the policymakers.

System Complexity Can Create Confusion in Implementation

PBASs can be extremely complex, and they usually include features that are new to providers, such as measures, rules for determining whether measured performance meets targets, and the awarding of incentives. In many cases, a PBAS requires a significant change in long-established practice. Thus, there are numerous places for confusion to arise. For example, health-care providers might be confused about the specifications of the measures; other stakeholders might be confused about the thresholds for participation and incentive-structure rules. The incentive structure has caused considerable confusion in health care, particularly when there are multiple PBAS sponsors operating in the same market with different rules. In general, there is more likely to be lack of understanding when a PBAS is new, and confusion can result if the new practices, including measures (and their specifications), score rules, and timetables, are not well documented at the outset.

Failure to Communicate with Stakeholders Can Limit System Effectiveness

Failing to communicate effectively with providers and other stakeholders is a common barrier to implementation. In education, for example, communication is one of the implementation challenges associated

with P4P systems. It can be difficult to explain the rules to teachers and administrators, to help them understand the complex formulas that are used to compute estimates of effectiveness and bonuses. Communication is not the only challenge; P4P systems have also confronted problems of distrust and organized opposition by teachers' unions, but part of the distrust might well be related to misunderstandings growing out of inadequate communication.

Sometimes, PBAS implementation can falter because service providers and other stakeholders simply are not aware of the program. Failing to communicate the existence or details of a PBAS can severely limit the effectiveness of such programs. In health care, for example, communication problems have abounded. In some cases, physicians have failed to realize that a PBAS exists or have failed to notice the receipt of performance reports or even award checks in the deluge of regular print and electronic correspondence. The communication process must be an ongoing effort, particularly if program design features change over time.

In child care, many parents do not know about the existence of a QRIS; thus, sponsors need to find a way to inform parents and encourage them to use the ratings to make child-care choices. A number of states rely on resource and referral agencies (R&Rs) to inform parents about the QRIS when parents contact them for child-care referrals. But many R&Rs are unnecessarily worried about divulging quality-relevant information to parents (Gormley, 1995).[1] To avoid the appearance of recommending centers that might later prove to be inadequate, one state barred R&Rs from telling parents whether a provider was participating in the QRIS or what rating it had received. This created a substantial implementation problem, since parents are most interested in learning about child-care quality when they are contacting an R&R.

Common mistakes made by some health-care PBASs include insufficient communication, communication through ineffective channels (e.g., mail or website that required doctors to log in despite the fact

[1] Gormley (1995) notes that R&Rs need to be educated about the difference between providing a useful public information source to parents and appearing to recommend a particular center.

that many do not have Internet connections or even email addresses), and communication that occurred late in the implementation process. In contrast, in transportation, where the A+B program is optional, significant outreach—in the form of conference presentations and publications, for example—helped to ensure that states were aware of this contracting innovation and had the opportunity to make use of it in their own projects.

Stakeholder Resistance Can Undermine System Credibility and Create a Barrier to Change

A related problem is that stakeholders might be aware of a PBAS but not endorse its goals or buy into the program. Getting stakeholders, particularly service providers, to endorse the goals of a PBAS can be extremely important when it comes to motivating behavioral change. Ironically, providers in service-oriented sectors, such as health care, education, child care, or transit, often believe that they are doing their best to provide effective service to their clients and that they do not need extra motivation to perform better. In some cases (e.g., child care), providers might want to provide good services but recognize that they are not currently doing so and *welcome* the PBAS incentives (and the subsidies and support that might come with it). In other cases, however, service providers might resent the implication that they need additional incentives to provide high-quality services. Where PBASs are voluntary, as is the case in child care and many health-care PBASs, getting buy-in from providers and encouraging participation in the program are critical. When participation is not coerced, PBAS designers can rely on higher levels of cooperation among participating providers—but that does not always ensure high levels of provider participation in the program.

Health care provides an example of a case in which PBASs have encountered significant resistance. Physicians sometimes try initially to ignore the existence of the PBAS, and, when that is no longer possible, some express anger at the system and resist participation. The anger typically emerges when providers are shown their performance scores, and many claim that the data are wrong or the program is flawed. While, over time, service providers might come to feel that a

PBAS is inevitable and thus come to accept the program, it is unclear whether many programs have reached the acceptance stage with service providers.

Voluntary PBASs are not the only ones that depend on provider buy-in for success. Such sectors as PHEP and transit, for example, are highly decentralized systems in which the PBAS sponsor (e.g., the federal or state government) has limited direct influence over providers. As a result, provider buy-in can be important to the success or failure of a PBAS. For example, transit operators have successfully opposed the use of PBASs in making funding allocations; they argue that no two transit providers operate under the same conditions (i.e., some markets and development patterns are more difficult to serve), that operators must meet a variety of local objectives, and that competing on a standard performance measurement is patently unfair. As a result of these protests, PBASs are not used in allocating funds to transit operators.

Potential Strategies for Successful Implementation

While the success of a PBAS implementation depends in part on the specific interactions between the PBAS design and the context in which the PBAS arises, some common strategies for success were observed in the cases we examined. In this section are some lessons relating to the implementation of the PBAS.

Exploiting Existing Infrastructure and Securing Other Resources Can Help Shorten Implementation

As noted in the previous section, building a PBAS on top of an existing infrastructure can facilitate the quick and accurate implementation of the operational components (e.g., data collection, reporting, accounting) needed to run the PBAS. This has been done, for example, in health care, where many health plans have adopted or adapted HEDIS—a set of explicit quality-of-care measures that were developed by NCQA—for use in PBASs. As in the health-care sector, performance-based transit-funding allocation relied on a data reporting and collection system that was in operation at the time of the PBAS creation.

In settings in which local measurement is not already in place, development of these operational systems requires additional time (typically a year or more, depending on the number of providers to be measured and the scope of what is being measured). Generally, the more complex and large the PBAS is, the longer the amount of time to construct the operational system. Operational systems typically require ongoing modification and adjustment to correct problems and handle new data sources and types of measures as the PBAS evolves.

In child care, for example, policymakers have invested substantial time and effort in, for example, finding and adopting measures and training observers to rate providers. However, problems arose when the incentive structure assumed that certain services or resources would be provided outside of the child-care system. In one state, for example, QRIS incentives included scholarships for teachers working in provider organizations that achieved a particular rating level. But the classes were not offered statewide, and the community colleges were overwhelmed with applicants. A different approach might have engaged other institutions in advance to ensure that they function as part of the QRIS as the QRIS planners envisioned.

Generally, it is important to assess upfront whether providers have sufficient resources to do what is required of them and to ensure that the agency responsible for overseeing the PBAS is adequately staffed and capable and has the necessary resources. Some of this information might be obtained from targeted pilot studies of the PBAS implementation process or from careful ongoing monitoring of the implementation process.

Allowing Systems to Evolve Incrementally Can Reduce Mistakes and Increase Buy-In

Many of the PBASs we examined represented the final stage in an evolutionary process, which allowed program sponsors to gradually add performance reporting and incentive components one step at a time. In some cases, early planners might not have had a PBAS in mind as the end result of this process; rather, the PBAS later came to be seen as the next natural step in the evolution of performance management. When feasible, allowing the PBAS to evolve incrementally can be advanta-

geous. For example, the gradual evolution of the CAA established all the necessary groundwork for PBAS metrics, measurement techniques and systems, and standards by the time the accountability element was added in 1990. The Federal-Aid Highway Amendments Act of 1963 (Pub. L. 88-157) first required urban areas to use a continuing, comprehensive, and cooperative planning process to qualify for federal funding; the CAA Extension of 1970 (Pub. L. 91-604) required all states to adopt a state implementation plan (SIP), including an emission inventory for each region in the state and a plan for attainment of all federal ambient air quality standards; and the CAA Amendments of 1977 (Pub. L. 95-95) and subsequent Federal-Aid Highway Act of 1978 (Pub. L. 95-599) required all regional transportation plans to show attainment of vehicle emission reductions specified in the SIP based on EPA-governed modeling. Finally, the CAA Amendments of 1990 (Pub. L. 101-549) provided that federal funds could be withheld from regions that adopt transportation plans that fail to show attainment of modeled emission reductions. The layering of steps in the accountability process over time demonstrated that the explicit accountability element of the 1990 act was needed, as the earlier steps were clearly insufficient by themselves to achieve the desired reduction.[2]

A common pattern in PHEP and public health care has been to tie initial funding to performance reporting and to move from there to explicitly rewarding performance itself. In PHEP, CDC's decision to tie funding to reporting seems clearly designed to ensure state buy-in for the PBAS. CDC does not have a large data-collection capacity, so it must rely heavily on self-reporting of data from grantees.

In health care, although private-sector programs have been well under way for some time, the federal government has been taking a

[2] It is worth noting that there might be downsides to such an incremental approach. Some critics of the CAA might argue that, by building on existing requirements and models, the program accepted some poor surrogates and inadequate measures from the earlier programs. For example, the models that must be used under the CAA guidelines are, in many ways, inadequate to the task to which they are applied. Had EPA been required to develop entirely new models for the purpose of implementing the amendments, environmental planning and possibly all of transportation planning could be more effective. The cost of that would have been staggering, but, by not doing it, policymakers have stifled advancement in the state of the art of assessing impacts of transportation control measures on air quality.

slower approach focused on learning as it goes and building political consensus. It started with voluntary measurement programs, to help providers become accustomed to measures and build systems and processes to capture and submit data. It then changed to pay-for-reporting (P4R) when the voluntary measurement programs typically yielded low participation rates. P4R has led to higher participation rates, but the response depends on the size of the payment relative to the provider's cost to participate. The P4R programs typically involve public reporting, but no financial incentives beyond these are given to providers for data submission. Many view P4R as the foundation on which Medicare can build a PBAS with financial incentives for performance.

A typical strategy of PBASs in the private health-care sector has been to start with only a few measures and to expand the measure set over time as there is more experience running the program and new measures are being tested and deemed ready for implementation. The initial measures were often already being collected and reported internally or at a higher level of the system (e.g., the health plan level). There is tension among key PBAS stakeholders about how many metrics should be included, with sponsors typically wanting to expand the set faster and those exposed to the PBAS wanting to go slower. Thus, ongoing implementation of these programs sees minor year-to-year adjustments in the number and types of measures.

Incorporating a Pilot-Testing Phase Can Head Off Unexpected Problems

In a similar vein, pilot-testing the PBAS might provide a valuable opportunity to iron out important kinks before the real program debuts. For example, in transportation, the evolution of A+B contracting for road construction began with carefully monitored pilot-testing. Under Special Experimental Project 14 (SEP-14), initiated in 1990, the Federal Highway Administration (FHWA) allowed state departments of transportation to evaluate, on a limited basis, a range of nontraditional contracting techniques. SEP-14 focused specifically on contracting options, including A+B and several other innovations, that are competitive in nature but do not fully comply with the requirements in Title 23 of the U.S. Code—that is, options that do not focus solely

on the minimum-cost qualified bid but rather incorporate additional measures of value (faster construction time, in the case of A+B). Goals of the evaluation phase varied by contracting mechanism; for A+B, the objectives were to determine whether the technique would lead to substantial reductions in the time required to complete construction, whether it would do so in a cost-effective manner (accounting for the benefits of reducing construction-related traffic-congestion delays), and whether efforts to reduce construction time would lead to adverse unintended consequences (e.g., lower-quality construction or greater safety risks for motorists or work crews during construction). In short, the pilot stage embodied in SEP-14 allowed time *prior* to full implementation to gather evidence that the technique was effective. Informed by the generally positive experience of state departments of transportation, FHWA ruled A+B suitable for operational use in 1995, giving states the latitude to employ this form of contracting for federally funded projects in suitable contexts.

Pilot-testing can help head off a number of potential problems. For instance, it can be difficult to define the right measures and set the right performance targets if there is no previous experience on which to draw. A common problem with new measures used for the first time is that not enough variation in performance exists to reward providers differentially for their performance. This can be a problem when results cluster at the top of the scale, so performance has "nowhere to go," and when results cluster at the bottom, so targets appear to be "out of reach." Pilot-testing, or implementing the PBAS following an observation period of performance measurement, can help to avoid "topping out" and "bottoming out." For example, the Oklahoma child-care QRIS, Reaching for the Stars, emphasizes the quality of staff and the learning program. Rating components include compliance with licensing requirements, teacher and director training, and teacher credentials. The original QRIS had just two levels. One star was awarded automatically with licensing. A second star required that a program meet internal quality criteria or achieve NAEYC accreditation, the latter of which was a very high standard attained by less than 10 per-

cent of the programs.[3] Very few programs were able to move from one to two stars, and, eventually, a third star was added, making the criteria to reach two stars less onerous. However, even this was not enough, and, the following year, a "one-star-plus" level was added because so few programs could reach the two-star level. Eventually, a time element was added, so that a program at this new level must move up to two stars within two years or drop back to one star. Some of these problems might have been avoided had the measures been pilot-tested before implementation.

Communicating Early, Often, and Through Multiple Channels Facilitates Understanding

Communicating effectively with stakeholders, including consumers of the service, is especially important in PBASs that rely on public reporting instead of or in addition to explicit financial incentives. In child care, for example, parents play a crucial role in choosing providers for their children; thus, sponsors need to find ways to inform parents and encourage them to use the ratings to make child-care choices. Another important lesson from the child-care sector is that parents are interested in child care only when they need it, and the window of time in which this is the case is small. One cannot assume that parents will retain knowledge of a QRIS to be used several years later. Consequently, efforts to inform parents (or, in the case of health care, patients) must be ongoing and, to the extent possible, targeted to those with immediate interest in the information. In health care, numerous consumer report cards have been created to help consumers make more-informed choices among providers; however, all too often, patients have not used these reports (Fung et al., 2008). To date, providers, employers, and health plans have been the primary consumers of the report cards. That said, Hibbard, Slovic, et al. (2002) and others are leading efforts to improve the design of consumer health-care report cards to make them easier for consumers to evaluate and use.

[3] An exception is in military child-development centers, where eligible programs are expected to achieve accreditation and accreditation rates are very high (see Zellman and Johansen, 1996).

The challenge of communicating effectively and engaging stakeholder support might also be affected by the design of the PBAS. For example, engaging providers might be a significantly easier task if the financial incentives are large enough to command their attention. In education, the possible loss of funding under Title I of the Elementary and Secondary Education Act (Pub. L. 89-10, 1965) (approximately 8–10 percent of total school education resources) made all stakeholders interested in understanding the NCLB regulations and eager to receive information about them. By contrast, some early attempts at P4P in education were seen as providing insufficient rewards, which might have reduced teachers' concern and hampered their acceptance. In transportation, the incentives associated with PBASs are uniformly large enough to matter to providers (e.g., auto manufacturers, contractors, metropolitan planning organizations). Thus, the providers had ample incentive to ensure that *they* learned about the system.

Communicating performance effectively to providers and other stakeholders also requires avoiding ambiguity and presenting results in a simplified, easy-to-understand way. In health care, there is some evidence that public reports that display data in a format designed to facilitate consumer understanding and choice are more successful in motivating quality improvements. In child care, QRISs were explicitly designed to communicate to parents, and star ratings have been an effective vehicle for doing this: Three stars is clearly better than one star. At the same time, care must be taken to avoid ambiguity about what a star really means. For example, in one state, the decision to provide every licensed provider with a one-star rating created ambiguity about the provider's quality: One star could mean that the provider was licensed but had not chosen to participate in the QRIS, or it could mean that a licensed provider was participating and delivered low-quality care.

Engaging Stakeholder Support Is Key to System Success

Effective communication is also needed to engage the support of stakeholders in the PBAS. Generally, to win providers' support, it is important to convince them that the PBAS is fair, particularly with respect to the measurements that determine their rewards or sanctions.

In health care, as mentioned in the previous section, this has been a major impediment. P4P programs in education have also struggled to gain educator acceptance. Denver's teacher organization rejected early efforts there, and only after years of hard bargaining with the teachers could the district implement a program that received its formal endorsement and general support.

Formative Evaluation Can Be Used to Identify and Correct Implementation Problems

The implementation process need not end when the system is first operational; instead, monitoring of the PBAS is an important tool for identifying and correcting potential problems and for improving the effectiveness of the PBAS. Chapter Eight examines the effectiveness of fully formed PBASs; here, we consider *formative-evaluation* activities that can be used to identify and address problems after the system is first operational. Formative evaluation can be contrasted with summative (ad hoc) evaluations in that formative evaluation encompasses monitoring program operations as part of an ongoing process. Both formative and summative evaluations are important and serve different purposes (understanding and improving the day-to-day functioning of the PBAS and evaluating its effectiveness as a tool for motivating service providers to improve, respectively).

It is important to monitor program operations both to ensure that the program is fully implemented and to provide information that might be used to revise the structure of the PBAS over time. For example, if a PBAS represents a significant change in practice, it might be important to assess stakeholders' knowledge of and participation in the system. Toward that end, useful measures might include the percentage of providers who have volunteered to be rated, the number of hits on the PBAS website, the number of providers or other stakeholders (such as parents) who have heard of the PBAS, the number of radio ads delivered, and the amount of newspaper coverage.

For example, in the early years of a child-care PBAS, the focus might be on the communication of PBAS goals and processes to parents, who are key PBAS stakeholders. Parents' use of the PBAS ratings in making child-care choices is regarded as critical to motivating

providers to improve their programs. Parents are intended to use a key PBAS system output—provider ratings—as they make decisions about where to seek care for their children. As a result, these ratings will become important to providers who want to attract new enrollments. However, it is not reasonable to expect providers to care about ratings if parents are not aware of them, do not understand them, or do not have easy access to them.[4] Similarly, it would be unreasonable to expect that providers would improve the quality of care they provide if the quality-improvement infrastructure (e.g., rating reports and trained coaches to work with providers to improve those aspects of their program that were rated lowest) are not yet in place.

The child-care case illustrates another interesting principle: Formative monitoring evolves to match the implementation of the PBAS. In the case of child care, the desired changes in the sector are incremental, evolving through a logical path. Consequently, it might be effective to have monitoring evolve as well. Initially, monitoring might focus on parents' access to and understanding of performance reports. As the system becomes more widespread, more effort could be devoted to assessing child-care providers' responses to signals from parents and providers' engagement with the improvement infrastructure.

Similarly, in a health-care PBAS in its early stages, it might be best to monitor whether groups targeted by the PBAS, such as physicians, know about the PBAS or whether proffered incentives are viewed as adequate to motivate them to pay attention to the PBAS and change their behavior in response to it. If incentives are too low to attract physicians' attention, any evaluation of PBAS outcomes is unlikely to show progress, and it might lead to the erroneous conclusion that the whole PBAS is flawed. Indeed, a monitoring study might reveal that simply increasing incentive levels could produce the desired behavioral change.

---

[4] In a personal communication to the authors in 2010, William Gormley pointed out that the existence of a public rating might lead providers to assume that parents will notice and use them. Unless there is evidence to the contrary, providers might act *as if* parents are attending to and using the ratings even if they are not.

In situations in which a PBAS is a new phenomenon, such as PHEP, it is important to know not only the incentives used but also the skills and capacities in place to close performance gaps revealed through measurement. RAND research on quality improvement suggests that the ability to close gaps (e.g., corrective-action planning, implementation and execution, follow-up and retesting) is very thinly and unevenly spread (see, e.g., Seid et al., 2007). As a result, it seems likely that observed variations in preparedness over time would be due, in large part, to the internal capacity of health departments as much as to variations in incentives introduced by the PBAS. Thus, it is important to interpret output measures in light of the developmental stage of PHEP.

Ongoing monitoring can also help identify and correct problems that might become apparent only with sustained attention. For example, measures and targets might need to be adjusted to accommodate changes in the distribution of performance *over time*. Health care provides an example of topping out occurring as the PBAS operates and performance improves. As providers move toward higher performance, measures might become highly compressed; P4P program sponsors have typically flagged such measures as having topped out. There has been much discussion about what to do with topped-out measures: Should they be retired so that providers can focus resources elsewhere? Should they continue to be monitored or become a prerequisite for participating in the PBAS, to reduce the likelihood of backsliding? Some sponsors require continued data collection, reporting, and maintenance of performance as a condition of being eligible for incentives. Much of the resolution of this tension depends on how PBAS sponsors weigh the two goals of encouraging high performance and encouraging improvement in comparison with one another.

Chapter Summary

Even a well-designed PBAS might not yield the desired results if it is not executed effectively. In this chapter, we described many of the key implementation issues and pitfalls that policymakers need to address,

identified some of the implementation problems that have emerged in our cases, and extracted some lessons learned.

When building a PBAS, exploiting the existing infrastructure and implementing in stages can minimize both the time needed for implementation and potential mistakes before they can compound. To the extent possible, PBAS sponsors should capitalize on existing infrastructure and make sure that what is required of the individual actors is possible, given current constraints. Implementing the PBAS in stages can also improve buy-in among key players and thereby increase the success of the program. Potential strategies for doing so include tying funding initially to reporting to build capacity for measurement and starting small and expanding measures and incentives over time. Incorporating a pilot-testing phase can head off a number of problems early, before the PBAS is rolled out in full, saving a great deal of time, effort, and resources.

Communicating with stakeholders is also integral to the success of the PBAS. PBAS sponsors should communicate with key stakeholders early, often, and through multiple channels. They should aim to make materials unambiguous and understandable and consider carefully whether the medium (e.g., email, website) is appropriate. Investing resources in engaging the support of stakeholders can be well worth the effort. Stakeholders need to be convinced that the measures are fair and tied to outcomes about which they care. Additionally, it is often useful to invest resources in educating providers with strategies for *how* they can improve.

Finally, formative monitoring can be an important tool for getting the most out of a PBAS. After the PBAS is rolled out, ongoing monitoring can help identify and correct problems that might become apparent only with sustained attention. Collecting data to track implementation progress can identify problems, such as earlier-than-expected topping out, to keep the system operating smoothly and efficiently.

CHAPTER SEVEN

Effectiveness of Performance-Based Accountability Systems

As noted in previous chapters, PBASs arise out of a desire to improve services that, if well-delivered, are posited to lead to the attainment of desired goals. But how effective have PBASs been in achieving these goals? This chapter examines empirical findings concerning the degree to which PBASs in the five sectors have been effective in improving services and whether those enhanced services have led to the achievement of desired system goals. As with so many other aspects of PBASs, the nine cases we investigated differ in terms of the amount of data available. We also draw on the relevant literature concerning PBASs in the sectors studied.

For the purposes of this analysis, *effectiveness* refers to the long-term correspondence between the espoused goals of the PBAS and the actual outputs and outcomes of the service-delivery activity to which it applies.[1] This relationship seems simple enough, but, as we will discuss, researchers face serious challenges in exploring these issues: Goals are not always well articulated; relevant outputs are not always well measured; the relationship between outputs and goals might not always be clear; other simultaneous interventions might make it difficult to discern the discrete effect of the PBAS; and tracking changes over time might be fraught with obstacles.

The chapter is organized around three questions:

[1] Again, we distinguish *outputs*—effects that a PBAS can observe, measure, and use to hold an activity accountable for service—from *outcomes*—effects that the PBAS cannot reliably link to the service provider's behavior.

- How effective have PBASs been in achieving their goals?
- Have any unintended consequences been associated with PBASs?
- What do we still not know about the effectiveness of PBASs?

Evidence of System Effectiveness in the Five Sectors

To date, there have been very few rigorous evaluations of the impact of PBASs in these sectors; the weak evidence that exists does not support a clear conclusion about the effectiveness of these systems. The data suggest that the implementation of a PBAS influences provider and user awareness and behavior to some degree and that PBASs often achieve some of their goals in terms of processes and outputs.[2]

Performance-Based Accountability Systems Have Captured the Attention of Service-Delivery Providers and Users

PBASs have captured providers' attention across the cases we studied. Changes in provider behavior represent evidence of this fact. For example, in health care, PBASs have led providers to dedicate new resources to incentivized activities (see, e.g., Sorbero, Damberg, and Shaw, 2006; Damberg, Raube, et al., 2009; Damberg, Sorbero, et al., 2007). There has also been much more attention to building and refining data systems as providers increasingly recognize that such systems represent a quality-improvement tool that managers can use to track provider performance (Damberg, Raube, et al., 2009). P4P programs have also resulted in greater management willingness to allocate resources targeted to quality improvement (Damberg, Raube, et al., 2009). In another example, PBASs have captured parents' attention in the child-care sector; in some states, the highest-rated providers have long waiting lists as parents increasingly seek out better care (Zellman and Perlman, 2008).

[2] Rather than a comprehensive listing of all findings in all sectors related to processes, we provide illustrative examples of results in this area. We are more comprehensive in reporting findings with respect to outputs and outcomes.

Performance-Based Accountability Systems Have Been Effective in Motivating Behavior Change

The available evidence suggests that, at the very least, in most of our cases, PBASs have been effective in motivating change in service providers' behavior, particularly if data about the PBAS are made public. Behavioral changes have been most notable in education, child care, health care, and transportation. In education and child care, the evidence suggests that these changes are motivated by several factors, including the fact that quality indicators are made public and are presented in a way that makes comparing service providers fairly easy (indeed, in child care, the ratings are designed to promote easy comparability across child-care providers). Studies by Lindenauer et al. (2007) and Hibbard, Stockard, and Tusler (2003, 2005) also show that, in the health sector, public reporting alone has been shown to change provider behaviors in ways that raise quality.

Overall, the evidence on PBASs in the education sector suggests that systems with consequences influence behaviors quite dramatically and that even just public disclosure of outcomes can be a strong motivator for some educators. Since 2001, NCLB has been the dominant PBAS in education, but, prior to its passage, there were other instances in which states adopted elements of a PBAS to try to improve public-school performance. These included various efforts to attach consequences (either for schools as a whole or for individual teachers) to student test results. During the past two decades, a number of studies have looked at the impact of such high-stakes testing in selected states. These policies, which embody many of the elements of a PBAS, have usually focused on schools as the unit of accountability, although some, such as minimum-competency testing for graduation, focused on students. Few of these state programs carried financial rewards (although California made bonuses of up to $25,000 available to all teachers in a small number of high-gain schools in 1999–2001). Instead, most either published summary scores for schools or coupled the public release of school data with ceremonial recognition for high performance or with interventions for poor performance.

The findings from evaluations of these education programs were nearly unanimous that attaching consequences to student test scores

influences teacher and administrator behavior and improves student outcomes (see Hamilton, 2003; Stecher, 2002, Hanushek and Raymond, 2005). On the positive side, these studies suggested that, in response to high-stakes testing, schools and district staff have taken steps to improve the quality and rigor of their curricula, increased programs and resources for low-performing students, incorporated data-driven decisionmaking into their practice, and provided professional development to help teachers improve their practices (Center on Education Policy, 2006; Hamilton et al., 2007; Lane, Parke, and Stone, 2002; Stecher et al., 2008).

In the health sector, the evidence is mixed about the impact that incentives have on provider performance on the incentivized measures. In many cases, the incentivized measures focus on process (e.g., behaviors believed to be necessary to achieve the PBAS's ultimate outcomes) (e.g., Damberg, Raube, et al., 2009), although, in most cases, these process and behavioral changes have been associated in other research with the desired performance outcomes. Overall, in health care, some studies have shown positive effects—in some cases, modest in size (e.g., Rosenthal et al., 2005; Pearson et al., 2008)—while others have found no discernible impact (e.g., Campbell, Reeves, Kontopantelis, Middleton, et al., 2007; Campbell, Reeves, Kontopantelis, Sibbald, et al., 2009; Curtin et al., 2006; Lindenauer et al., 2007; Reiter et al., 2006).

Performance-Based Accountability Systems Have Helped Providers Focus Attention on Aspects of Service Needing Improvement

In child care, health care, and education, PBASs have set standards that help providers focus attention on those aspects of service delivery considered to be most closely associated with improved quality and those that are therefore incentivized. This focus is also apparent in transportation, in which large potential payouts or penalties have led metropolitan planning organizations to drastically alter planned transportation investments to comply with federal air quality regulations (McCarthy, 2004), contractors to significantly reduce the time required to complete highway maintenance or improvement projects (AASHTO, 2007; Strong, 2006), and auto manufacturers to increase

the average fuel efficiency of their vehicles (BEES, 2002, T. Martin, 2005).

Some Evidence Links Performance-Based Accountability Systems to Improvements in Long-Term Outcomes

PBASs are designed to achieve long-term performance goals, but the evidence about the effects of PBASs on outcomes is limited. One reason is that performance-measurement systems are often limited to tracking clients for, at most, one year postservice; longer-term outcomes are often left to in-depth program evaluations. But assessment of short-term outcomes might have considerable value to both clients and policymakers. Some sectors have seen more outcome-focused research than others. There is also variation in the amount and type of evaluation data available within sectors that include more than one PBAS. In this section, we review the outcome evidence in each of the five sectors we studied.

Child Care. There have been only two systematic studies of the effects of a statewide QRISs on child outcomes, although newer QRISs are planning such studies. These studies have shown mixed results. Zellman et al. (2008) found no relationship between QRIS quality and child outcomes in their Colorado assessment, although their study was burdened by very high child attrition, which limited the value of its outcome analyses. In addition, research undertaken since the development of this QRIS suggests that some important, outcome-relevant components, such as use of an evidence-based curriculum, should be part of both the QRIS standards and the rating process. A recent study by Thornburg et al. (2009) of Missouri's QRIS found statistically significant gains over the course of a school year in overall social and behavioral skills, motivation, self-control, and positive adult relationships among children when improvements over the course of a school year were compared among children attending the highest- and lowest-quality providers. However, it is not possible to disentangle the effects of selectivity in this study: Children in the higher-ranked programs might come from backgrounds that better support their development and therefore have parents who seek out and pay for higher-quality care.

Given the high costs of assessing child outcomes (data must be gathered through individual testing of young children), many QRISs have focused their evaluation activities instead on demonstrating gains in participating-provider quality. Provider quality is associated in other studies with improved developmental outcomes, including improved language development, cognitive functioning, social competence, and emotional adjustment (e.g., Howes, 1988; NICHD ECCRN, 2000; Peisner-Feinberg et al., 2001; Burchinal et al., 1996; Clarke-Stewart et al., 2002). The dearth of evaluations of the effects of PBAS implementation on child outcomes might also reflect the strong professional confidence in the ability of standards, ratings, incentives, and support to improve outcomes.

Limited evaluation data suggest that the combination of public reporting, provider-level quality incentives, and quality-improvement support have been effective in improving provider quality, a key system output (e.g., Zellman and Perlman, 2008; Zellman et al., 2008; Thornburg et al., 2009). However, at least some of these studies have overstated the success of their systems because of problems in the research design: Success is determined by correlating summary ratings with a global quality indicator, often the Early Childhood Environment Rating Scale, Revised (ECERS-R), but the summary ratings in many QRISs include the ECERS-R. Hence, the level of success is inappropriately inflated.

Education. Considerable research on education PBASs has examined student performance, a key education PBAS outcome. Several correlational studies have sought to link variations in accountability policies (such as NCLB) across states or countries with the performance of students on statewide or national tests (Amrein and Berliner, 2002; Bishop, 1998; Bishop, Mane, and Bishop, 2001; Carnoy and Loeb, 2002; Dee and Jacob, 2009; Hanushek and Raymond, 2005). The results of these studies were mixed, although the predominant finding was a positive association between strong accountability systems (with goals, measures, and incentives) and improved student performance. However, although most of the studies attempted to control for factors that might influence the relationship between accountability and

performance, all of these findings might be influenced by unmeasured factors.

P4P programs are relatively new in education but appear to have strong effects on student achievement. Podgursky and Springer (2007), although reviewing only a modest number of studies, report that positive individual teacher incentives were effective in raising the level of the variable being incentivized and that, in most cases, the incentive regime resulted in positive student achievement.

Health Care. PBAS evaluation in the health sector has focused less on outcomes than has evaluation in other sectors because key outcomes, such as long-term health, take a long time to emerge; linking them to a particular PBAS or any other discrete intervention is difficult. However, some work by NCQA has tried to translate improvements in processes to anticipated benefits. For example, using the clinical literature and extrapolating the effects of changes in processes, NCQA (2009) projected longer-term outcomes, e.g., "XX fewer cases of childhood diseases, or XX heart attacks prevented." Another study (Reiter et al., 2006) estimated the impact of a PBAS in terms of the number of quality-adjusted life years gained by virtue of the improvements they saw in the incentivized indicators.[3]

For a variety of reasons, the health-care sector is further along than other sectors in assessing the effect of its P4P PBASs. There is a long history of performance measurement in health care and considerable evidence that, among the various incentives that have been deployed to shift the behavior of health-care providers, public reporting of health outcomes for provider organizations leads to those organizations trying to improve services (e.g., Lindenauer et al., 2007; Hibbard, Stockard, and Tusler, 2003, 2005). There is also some evidence that financial incentives have led to shifts in provider behavior (Damberg, Raube, et al., 2009; Rosenthal et al., 2005). The nature of the health-care sector, which includes public- and private-sector payers (the government, private employers, and private health plans) and private providers, complicates the implementation and evaluation of PBASs, since a given

[3] A quality-adjusted life year is a year of life statistically adjusted for quality-of-life differences arising from the presence or absence of disease conditions or functional limitations.

provider whose behavior is the target of a PBAS might be the target of several other PBASs as well, offering varying levels of incentives.[4] This creates challenges in understanding what factor or set of factors might have been responsible for the observed changes in behavior.

Despite the history of assessing P4P initiatives in the health-care sector, the evidence to date on their effectiveness at improving health-care quality is limited and shows mixed results. Providers often note anecdotally (and some studies support their view) that public reporting has not yet led to changes in patient behavior—that is, using public reports to decide whether to stay with a current physician or to make a change (Damberg, Raube, et al., 2009, Fung et al., 2008). The seven most rigorously designed studies (i.e., those using randomized controlled trials, or RCTs) provide an ambiguous message: Four show mixed results (Fairbrother, Hanson, et al., 1999, Fairbrother, Siegel, et al., 2002; Kouides et al., 1998; Roski et al., 2003), and three report no effect (Grady et al., 1997; Hillman, Ripley, Goldfarb, Nuamah, et al., 1998; Hillman, Ripley, Goldfarb, Weiner, et al., 1999). The least rigorously designed studies (e.g., quasi-experimental or correlational) tend to report positive results for at least one aspect of the programs examined (Francis et al., 2006; Greene et al., 2004; Amundson et al., 2003; Armour et al., 2004; Fairbrother, Friedman, et al., 1997; Morrow, Gooding, and Clark, 1995).

Drawing conclusions from these studies about how P4P affects health-care quality is problematic, given the small scale and brief duration of these interventions, many of which were mounted at a single location having selected characteristics (e.g., Medicaid providers). Many of the studies lacked control groups, making it difficult to distinguish the effects of P4P from the effects of other factors in the environment (e.g., medical-group quality-improvement interventions, public reporting of performance scores). CMS has also studied the effects of physician P4P programs; however, because providers volunteered to

[4] Typical quality-based incentive payments offered to U.S. physicians are in the range of $1,500–$5,000; incentives to reduce the utilization of health-care services tend to be much higher than incentives to improve quality, often representing three times the amount allocated for quality improvement.

participate in the P4P regime, it is difficult to know whether changes in outcomes are due to the PBAS or to other conditions that might encourage some providers to volunteer rather than others (Lindenauer et al., 2007).

The more recent health-care P4P evaluation literature focuses on larger-scale interventions that were longer in duration and typically offered larger rewards. Across these studies, the results have shown either modest improvements or mixed results (e.g., Pearson et al., 2008; Damberg, Raube, et al., 2009; Campbell, Reeves, Kontopantelis, Sibbald, et al., 2007; Curtin et al., 2006; Rosenthal et al., 2005; Lindenauer et al., 2007). Although these studies evaluated more-robust efforts, a key limitation of the studies is that they provide no information about the various design features that might have played a role in an intervention's success or failure, such as the level of engagement and communication with providers and what share of a physician's practice the intervention represented (i.e., the dose effect).

Some health-care sector evaluations have been seriously flawed. As in child care, some studies have been launched too soon after implementation to fairly measure PBAS effects. For example, it can take a couple of cycles of performance reporting for providers to submit complete data, so changes measured during those initial periods reflect more about the completeness of data than about the quality of care. Other problems observed in evaluations of health-care PBASs include lack of a comparison group, which makes it hard to discern the effect of the intervention as opposed to other changes occurring concurrently; bias among the volunteers in the PBAS such that participants are substantially different from nonparticipants; and the absence of preintervention-period data to assess before-and-after effects (Sorbero et al., 2006; Damberg, Sorbero, et al., 2007). These research-design problems make it difficult to attribute observed changes in performance to the PBAS.

PHEP. Not surprisingly, little evaluation activity has occurred to date in PHEP because PBASs in this sector are so new. Moreover, the very nature of the sector might make it impossible to conduct the sorts of evaluations possible in the other sectors. The nature of preparedness precludes many sorts of studies; relevant events occur rarely, and, when

they do, by definition, they are rarely exactly the kind of emergency that was anticipated. Moreover, since services are not delivered on a regular basis (emergencies do not occur on a regular basis), evaluations necessarily must focus on inputs, structures, and processes. However, professional consensus is building concerning the value of drills and simulations, and applying resources in selected simulated potential contingencies might be a way to assess preparedness. However, the relationship of these efforts to preparedness in the event of an emergency is not clear.

Transportation. Transportation PBASs have been found to change provider behavior as well as key outcomes—due at least in part to the high level of incentives. P4P initiatives in the transportation sector, particularly A+B contracts, have been successful in meeting PBAS goals. Studies of A+B contracting, for instance, reveal that this approach, by enabling contractors to earn more for faster project delivery, has consistently reduced the calendar time required for highway construction by substantial amounts while promoting innovations among contractors (AASHTO, 2007; Strong, 2006). The use of large incentives motivates behavioral change; the use of simple metrics makes desired outputs clear and conclusions easy to draw. With A+B contracting, the exclusive focus is on time to completion; such a narrow focus is possible because highway construction projects are already governed by regulations that address other important concerns, such as safety and engineering standards.

CAFE standards have also led to changes in provider behavior that supported their intended outcomes. These standards, too, focus on a relatively simple measure: a manufacturer's sales-weighted mean fleetwide fuel economy levels. Compliance among major auto manufacturers (with the exception of such performance-oriented producers as Porsche, Daimler, and BMW) has been high (Martin, 2005), and a recent examination by the National Academy of Sciences (BEES, 2002) estimated that total U.S. fuel consumption, as of 2002, would have been 14 percent greater in the absence of CAFE standards or other fuel-consumption reduction policies. The program has not, however, been an unqualified success. Several analyses have suggested that alternative approaches to reducing fuel consumption—for instance,

increased gas taxes, "fee-bate" programs (in which consumers pay a fee when purchasing less fuel-efficient vehicles and receive a rebate when purchasing more fuel-efficient vehicles), or carbon cap and trade—might achieve the same outcomes more cost-effectively (BEES, 2002; CBO, 2002, 2003).

Accountability provisions in the CAA Amendments of 1990, which threaten to withhold federal transportation funds from regions that fail to demonstrate adequate progress toward compliance with applicable air quality standards (Johnston, 2004), have proven effective in motivating changes on the part of metropolitan planning organizations. To avoid the loss of federal funds, several metropolitan planning organizations have made significant changes to proposed transportation improvements—for instance, replacing highway expansion plans with transit investments—in order to reduce forecasted pollutant emissions (McCarthy, 2004). It is less clear, though, that the CAA's accountability provisions have been sufficient to achieve the intended outcomes. Though urban air quality has improved significantly over the past 50 years (Bae, 2004), there were still, as of 2004, more than 100 urban areas, with a combined population of more than 160 million residents, that had not achieved compliance for one or more of the criteria pollutants regulated under the CAA (McCarthy, 2004). Particulate matter and ozone, in particular, remain problematic in many regions (Bae, 2004). There are at least two possible explanations for these shortfalls. First, the modeled projections of future emissions on which the accountability provisions in the CAA are based might be prone to error. Second, there are many additional contributors to air pollution beyond motor vehicles—including stationary sources, such as factories or refineries; airplanes; ships; off-road vehicles; and gas-powered lawn mowers and leaf blowers—that fall beyond the purview of the metropolitan planning organization.

The evidence on the effects of PBASs in transit has been mixed. This might reflect the greater complexity and higher numbers of potential stakeholders, which renders these PBASs more similar to those in health care, education, and child care. While A+B contracting is clear and straightforward, transit is complex, characterized by multiple and often competing objectives among different stakeholders. Consider, for

example, the twin goals of providing greater accessibility and maximizing ridership. The former might lead to increasing transit service in suburban areas to increase geographic accessibility, though most suburban lines are only lightly patronized. The latter, in contrast, might argue for concentrating even more service in dense urban areas. Given the large amounts of money to be allocated, continuing disagreements about fundamental measures have complicated both the implementation and evaluation of transit PBASs.

Information About Unintended Consequences and Costs

PBASs have the potential to cause unintended consequences, such as incentivizing the wrong kind of behavior or encouraging undesirable effects. A PBAS might also be less cost-effective than other potential improvement initiatives. In this section, we discuss some of the unintended consequences of PBASs and some cost-related issues found in our review of evaluations in these sectors.

Unintended Consequences Vary Across Sectors

Reengineering complex systems and attaching incentives to specific behaviors inevitably change the system; some of the changes might not be intended or positive. In education, attaching public reporting and other incentives to test scores has led to unintended behavioral changes that might be considered undesirable. Most of these changes stem from the fact that the tests that assess student performance in PBASs measure only a portion of what schools are expected to do. Probably the most common response among teachers to the reliance on test scores in the PBAS is to shift instruction away from nontested content toward tested content. These changes can take many forms, ranging from shifting instructional time across different subjects to more-subtle changes in emphasis within a subject, such as having students read more short passages and fewer novels. In systems that are intended to promote instruction that is aligned with content standards, these responses can be problematic when tests are not perfectly aligned with standards. Research suggests that tests tend to focus on standards

with lower cognitive demand and underrepresent the more challenging standards (Rothman et al., 2002). So, while some educators have characterized teaching to the test as a good thing, the kind of test-focused instruction described here might result in insufficient attention being paid to important problem-solving and reasoning skills.[5]

There might also have been some unintended effects from PBASs in the transportation sector. With CAFE standards, one common strategy for boosting fuel economy to meet the fuel-efficiency standards is to make cars lighter, and this has introduced greater safety risks. Although this problem has been mitigated to some extent through improved engineering designs, it is still the case that, all else equal, lighter cars face greater safety risks than heavier cars, so the PBAS associated with CAFE standards tends to increase safety risks in unintended and unanticipated ways (BEES, 2002; Gayer, 2004; Viscusi and Gayer, 2002). CAFE standards might also have led auto manufacturers to subsidize their most efficient vehicles to boost average fleetwide economy; this results in more total vehicles on the road, working counter to the goal of reducing aggregate fuel consumption (T. Martin, 2005). The decision to create separate standards for passenger vehicles and light-duty trucks has also undermined the goal of reducing total fuel consumption. Between 1979 and 2000, the market share for light-duty trucks (including minivans and sport-utility vehicles, or SUVs) increased from 10 percent to 44 percent. Thus, while vehicles in both classes became more fuel-efficient over this period, the sales-weighted average fuel efficiency for the two classes taken together exhibited a much more modest increase (BEES, 2002; T. Martin, 2005).

While there have been concerns about possible unintended consequences occurring in health-care PBASs (Casalino et al., 2007), to date, there is an absence of empirical evidence showing such effects (Friedberg, Mehrotra, and Linder, 2009; Doran et al., 2008). This lack

[5] In the early years of test-based accountability, school and district administrators took such steps as moving the most highly qualified teachers to tested grade levels and subjects and encouraging low-achieving students to stay home on testing day (see Hannaway and Hamilton, 2008, for a review) as a way to improve test scores. However, newer accountability policies have made it more difficult to engage in these kinds of activities (e.g., by mandating that nearly all students take the test).

of evidence might reflect the difficulty of studying the problem and, within the United States, the low levels of incentives (in contrast, high incentives might cause physicians to drop more-difficult patients to achieve higher scores). A related concern in the health-care sector is that PBASs include a narrow set of performance markers; this might increase physicians' focus on what is measured and reduce their attention to unmeasured behaviors or procedures. However, a recent study found that providers who improved slightly on incentivized behaviors did not significantly decrease performance in unrelated areas of care (Mullen, Frank, and Rosenthal, 2010). In another study, areas that were related to the targeted measures also saw similar improvements, but unrelated areas showed none (Asch et al., 2004). Tracking non-incentivized measures in addition to incentivized measures can help shed light on whether providers are responding to P4P in adverse ways, as noted earlier in our discussion of education.

A Performance-Based Accountability System Might Not Always Be the Most Cost-Effective Option

While data are generally lacking on the cost-effectiveness of the PBASs we studied, a few health studies address this issue. For example, Curtin et al. (2006) used data from 2003 and 2004 to evaluate a private-sector P4P program and found an average net savings of $2.4 million per year compared with the projected spending trend for diabetes care associated with providing more-reliable diabetes care (a return on investment of 1.6 to 1 in year 1 and 2.5 to 1 in year 2 of the program). The savings estimates accounted for new spending to provide underused services for managing patients with diabetes. The largest savings came from reducing hospitalizations; physician costs, pharmacy, and outpatient spending were also reduced. However, this study has not been replicated in the same or a different setting, so it is unclear whether the results can be generalized to other settings.

Wheeler et al. (2007) suggest a more-refined approach to considering cost-benefit calculations in health P4P initiatives. Their estimates of net cash flow create separate estimates for payers and providers. They argue that it is important to calculate the sometimes very different net costs and benefits for the key players in order to understand each play-

er's incentives. Once done, it might be possible to adjust incentives to ensure that each player sees the advantage of participation and thus helps ensure the long-term stability of the initiative.

In the transportation sector, some argue that other policy options, such as fuel taxes or fee-bate programs, could have reduced fuel consumption more cost-effectively than CAFE standards (BEES, 2002; CBO, 2002, 2003).

Gaps in Our Knowledge

As the preceding review suggests, the absence of studies of many PBASs and the often weak study designs used when evaluations are conducted leave fundamental questions unanswered. For example, policymakers in specific sectors need to better understand whether PBASs are effective in achieving their ultimate goals (e.g., improved school readiness in child care, better care and improved intermediate clinical outcomes in health care). In general, policymakers could also benefit from clearer understanding of the circumstances under which PBASs are most effective. There is a dearth of studies about the cost-effectiveness of PBASs or about whether other approaches to quality improvement might be equally or more effective. Yet, such information is crucial to knowing how best to allocate scarce resources. We also need more information about the ideal size of incentives and how many measures on which to concentrate a PBAS. Without such information, we might be designing suboptimal PBASs, which lead to false impressions about whether a PBAS is the best policy approach. And, as Wheeler et al. (2007) note, this information must be sensitive to the systems in which PBASs operate. In health care, for example, PBASs that do not financially benefit *both* payer and provider are less likely to flourish, even when substantial health improvements result. Indeed, cost-benefit questions are critical ones, as the cost of implementing PBASs can be substantial, particularly in such sectors as child care, in which few data are normally collected.

Chapter Summary

Rigorous studies of PBAS effectiveness are more common in some sectors than others; yet, in general, PBASs have not been subject to rigorous scrutiny. For example, PBASs in education and health care are being studied regularly, but the study designs could be improved to strengthen the ability to draw conclusions about impacts and to understand the degree to which specific design elements are affecting a program's success in meeting its goals. Few output- or outcome-focused evaluations have been conducted in child care; PHEP evaluations are even more limited. Overall, the bulk of the evidence about PBASs in these sectors is descriptive, indicating how well the program was implemented; there are also findings relating to changes that have occurred in service-delivery practices, which represent intermediate effects of the PBAS. These studies do show changes in provider behavior and some improvements in service delivery. The number of outcome studies is small, and the number of cost-benefit studies is tiny.

The evidence that does exist leads to somewhat different conclusions by sector. In education, it is clear that NCLB and other high-stakes testing programs with public reporting and other incentives at the school level lead to changes in teacher behavior; however, teachers seem to respond narrowly in ways that improve measured outputs (i.e., the measures) with less attention to long-term outcomes (i.e., the goals). Interestingly, the absolute size of incentives does not seem to matter much, with public scrutiny being very salient in shaping behavior. While student test scores have risen, there is uncertainty whether these reflect more learning or are to some degree the product of teaching to the test or other approaches to generating apparent improvement. In health care, in contrast, relatively small financial incentives (frequently combined with public reporting) have had some modest effects in improving the quality of care delivered. The transportation examples suggest that large financial incentives can lead to creative solutions, as well as to lobbying designed to influence the components of the PBAS, as in the case of CAFE standards. Child-care PBASs seem to be increasing the quality of care among participating providers, but the limited studies that looked at longer-term outputs have reported mixed

effects on child outcomes. It is too soon to judge the effectiveness of PBASs in child care and PHEP.

Obviously, we need more and better evaluation data and the commitment by sponsors to evaluate their programs. In the next chapter, we discuss the many reasons to better evaluate PBASs. We also discuss why PBASs have not been rigorously studied and identify ways to make evaluation both more appealing and less challenging. We end the chapter by highlighting three key aspects of evaluation that are particularly relevant to the study of PBASs.

CHAPTER EIGHT

Motivating and Improving System Evaluation

Despite the recent increase in the use of PBASs to manage public service programs, there has been little rigorous research about their implementation or impact in the cases we examined, as documented in Chapter Seven. Given that PBASs are designed to promote accountability through performance measurement, it is somewhat ironic that there are very few studies across the sectors that measure the performance of PBASs themselves, either in terms of their implementation and processes or in terms of their outcomes. As we explored PBASs in this study, we developed a number of hypotheses about why PBASs are implemented more often than they are evaluated. We describe those hypotheses in this chapter. We also argue strongly in favor of PBAS accountability by describing the benefits of PBAS evaluation and ways to make the evaluation of PBASs more feasible. In this chapter, we focus on the following questions:

- Why evaluate a PBAS, and what sorts of questions could such an evaluation answer?
- Why is there currently so little evidence about the effectiveness of PBASs?
- How could evaluation be made more appealing and more effective?

In addition, Appendix B briefly presents a range of design alternatives for evaluating PBASs and refers interested readers to more-detailed and technical discussions of those designs.

Reasons to Evaluate

The overriding motivation for evaluating any policy or system is to find out whether it "works" and, "if not, why not?" *Working* might include a wide range of issues, including these:

- Are all the pieces in place? Is the system achieving its goals? These questions are generally answered through process evaluations
- Does the system produce better outcomes than another policy with the same goals or than business as usual? Along with the first pair of questions, this is generally answered through outcome evaluations.
- Given what it does, is the system worth the cost?

Without an evaluation, it is not possible to know whether a policy is living up to its promise, nor how to modify it so that it can. Exploring whether a policy works is important because the implementation of any policy imposes a series of costs on a system; policymakers want to make sure that these costs produce commensurate benefits. In addition, the implementation of a given policy often precludes the adoption or implementation of other policy options; if the selected policy is not meeting its goals, policymakers might be wise to replace it with another approach. A given policy or approach might also produce unintended consequences that might overshadow its positive outcomes. Evaluation can provide essential information that enables program sponsors to adjust the PBAS to maximize the positive impact and minimize the negative.

An evaluation, correctly done, might also help policymakers understand whether outcomes are the result of a particular policy or of other co-occurring events. For example, school leaders in Atlanta told RAND interviewers about recent findings that could have been interpreted incorrectly if external factors had been ignored (Zellman, 2010). A major school-reform effort in the Atlanta public schools led to apparent improvements in student outcomes. However, while the reforms were being instituted, several public housing units were demolished as part of an urban renewal program; the demolition forced large

numbers of struggling students to move out of the district into public housing in another district. Both the reforms and the changes in student enrollment might have contributed to improvements in aggregate student achievement, and it would take a careful analysis to tease out those competing explanations for improved aggregate scores. Similarly, in health care, observed improvements in quality might be a function of the PBAS's financial incentives or might be due to other factors, such as investments in health information technology, investments in quality-improvement interventions, or public reporting.

Careful evaluations of PBASs can also help policymakers determine whether the costs associated with implementing and operating the system are producing commensurate benefits. A well-designed evaluation can help to pinpoint implementation problems that need to be corrected before it is reasonable to assess whether the PBAS is achieving its ultimate goals. An evaluation can also help to identify unintended consequences, such as gaming the system, weaknesses in the measures used to assess key outcomes, and flaws in system implementation and operation.

Of course, there are limits to what can be learned from an evaluation even under the best of circumstances. Hess (2005) notes that management reforms do not take place in controlled circumstances and that the results of evaluations will, of necessity, be context-dependent and have limited generalizability beyond the situation in which the evaluation were conducted. Furthermore, circumstances often conspire to limit how well evaluations can be conducted. Compromises dictated by local conditions will lessen the scope and rigor of conclusions that can be drawn from a given evaluation.

What Questions Could Evaluations Answer About Performance-Based Accountability Systems?

PBASs have become popular in some sectors despite a lack of strong evidence that they work well in achieving system goals or that they are superior to other policy approaches. A good example of this proliferation without empirical support might be found in the child-care sector,

in which, in only 10 years, statewide PBASs have been established in 19 states, with several more states poised to launch them in the near future. Yet, states that are just now designing statewide child-care PBASs have little in the way of rigorous design guidance. While PBAS designers from the states that have adopted these systems are generous in sharing their thoughts and experiences, they cannot provide more than anecdotal evidence about what specific elements, in which combinations, appear to be most effective in motivating desired behaviors or supporting higher-quality care. More-frequent and more-rigorous evaluation could help policymakers decide whether to adopt a PBAS or to favor other policy alternatives.

Similarly, the absence of evaluations undermines one of the key potential advantages of PBASs—that the wealth of data they produce about system performance could be used for both system improvement and PBAS improvement. This seems to be one of the areas in which PBASs in the public sector function differently from PBASs in the private sector. Private-sector PBASs are often explicitly designed to facilitate organizational learning and evolution over time. Continuous monitoring and analysis of data lead to refinements in the monitored services and in the PBAS itself.

Careful evaluations could be extremely useful to PBAS designers, as well as to others who might be considering the implementation of a PBAS or similar performance-management system. Even less rigorous designs have considerable value; simply looking at what has happened in a systematic manner and examining the reasons can provide extremely useful information about a performance-measurement system. While evaluation experts rightly argue that more-rigorous outcome evaluations provide better information, less powerful approaches should not be ignored if more-rigorous approaches are not feasible. Process or outcome evaluations could provide answers to a number of questions, which would be extremely valuable to policymakers and PBAS designers. A few of these questions are discussed in the next paragraphs.

What Does It Cost to Design and Run a Performance-Based Accountability System?

Information about what it costs policymakers and designers to design and implement the PBAS and what it costs the service provider to respond would help both groups better use available resources. For example, states that operate child-care QRISs only rarely calculate the cost of the rating process, even though ratings constitute a major PBAS expenditure. Yet, high rating costs threaten to limit resources for other key system activities, such as quality improvement. Learning early on how much ratings and quality-improvement efforts cost would enable system designers to consider, in a way they rarely do now, possible trade-offs across system elements.

How Cost-Effective Are Particular Systems or System Designs?

To answer this question, it is necessary to collect both cost and output data. Ideally, these data would be available for a range of PBAS designs so that the impact of different design elements could be assessed. It would be even better if cost and output data were available for other policy options as well, so that PBASs could be considered as one of several policy alternatives, as discussed in Chapter Seven. Cost-effectiveness studies of PBASs are rare at present, but more-widespread implementation of these techniques not only would benefit specific programs but would also create a potentially powerful database to improve future comparisons.

To What Degree Have Performance-Based Accountability Systems Become Learning Organizations?

Data from a PBAS evaluation could also be used to determine the extent to which PBASs function like learning organizations—i.e., systems that are *designed* to evolve and improve over time. Evidence from our cases reveals, at best, limited efforts to date to design PBASs in this way.

Short-term implementation and output data are needed to conduct this type of analysis. The child-care sector offers a good example of a system that learned from evidence and responded by making positive changes. Oklahoma designed the first statewide child-care quality

rating system. Launched in 1998, Reaching for the Stars had two levels of quality. One star was awarded automatically to state-licensed providers. A second star set a very high bar: attainment of internal quality criteria or accreditation by NAEYC. Designers soon learned through provider and rater feedback that providers were having trouble moving to two stars because the amount of quality improvement required was so great. In 1999, a third star was added at the top that enabled the distance from one to two stars to be reduced. In 2000, a one-star-plus level was added because so few programs were reaching the two-star level.

Education offers an opposite example. NCLB includes increasingly rigorous student performance benchmarks over time, but these increasing targets were established at the outset and have not been modified in light of states' progressively great difficulties in achieving them. On the other hand, some of the rules have been modified over time in response to concerns expressed by the states. For example, the guidelines for excluding students with disabilities were modified when it became clear that the strict initial requirements were inappropriate. Changes have also been made in response to policymakers' interest in new alternatives. For example, the U.S. Department of Education undertook a growth-model experiment to test a modified metric and target system based on student growth toward proficiency. However, neither type of change was anticipated in the development of the system, and neither occurred as the result of a planned evaluation.

In health care, some PBASs are closer in design to learning organizations, and those that are typically have embraced evaluation as a core component of their ongoing operations. However, for many PBASs, midcourse corrections represent ad hoc responses to something that is not working (e.g., in response to provider complaints) rather than systematic monitoring processes established at the front end of the program to capture data and provide routine feedback for improvement.

Are Performance-Based Accountability Measures Adequate Proxies for Long-Term System Goals?

As we discussed in Chapter Seven, little evidence is available to evaluate the long-term outcomes of the PBASs. In most sectors we examined,

the outputs (or, in some cases, the processes or inputs) that are monitored and incentivized are short-term indicators of desired long-term social outcomes that (1) are more general and more difficult to quantify or (2) unfold over a much longer timeframe and cannot be measured during the evaluation cycle of a PBAS. So, for example, in education PBASs, student annual test scores serve as a proxy for the ultimate goals of education, including employability, civic involvement, and a fulfilling life. In health-sector PBASs, discrete physician behaviors, such as conducting foot examinations on diabetic patients or ordering regular mammograms for women over 50, are measured because the scientific evidence has shown that doing so leads to better overall health outcomes, e.g., fewer amputations and earlier detection of cancer. In the case of performance-based transit funding—one of the less successful examples in the transportation sector—relatively crude measures of the quality or quantity of service are used to represent the broader goals of promoting greater mobility and accessibility among the populace. For CAFE standards, it is assumed that the measure of the average fuel economy of vehicles in the current model year should have some bearing on total fuel consumption in the coming years.

Evaluation designs that examine the link between long-term outcomes and near-term outputs would help to improve the design of PBASs and might reduce the sometimes overblown claims made for them. For example, child-care PBASs have been touted as a means of improving school readiness. Such claims are based on rigorous RCTs comparing no child care to high-quality care provided using a standardized curriculum; these studies show that better care is associated with improved child outcomes (e.g., Campbell and Ramey, 1995; Weikart, Bond, and McNeil, 1978). However, these conditions—high-quality care with a standardized curriculum—do not apply in usual child-care settings (Ramey and Ramey, 2006), and many in the child-care sector worry that this promise, if not realized completely, will undermine PBASs that are producing important outputs, such as higher-quality care, in many settings.

How Well Do the System's Incentives Work to Alter Behavior?

Sectors differ markedly in the degree to which service providers face multiple, perhaps competing, incentives. In health care, for example, doctors might be pulled in many directions by multiple incentive programs and different types of measures that might conflict. In some cases, substantial incentives for reduced health-care utilization might eclipse incentives for improving quality. In contrast, in a resource-constrained sector, such as child care, small incentives might not have much impact on behaviors because resource constraints impede the changes that are needed to deliver higher-quality care.[1] Incentives would need to be sufficiently large to cover the cost of potential improvements before changes were likely to occur. For example, a PBAS might offer child-care providers a bonus for reducing child-to-staff ratios, but, if the incentive level does not cover the cost of hiring more staff, ratios will not decline, even among highly motivated providers. Conversely, PBAS designers would benefit from knowing how low the incentive level can be while still producing the desired change. That way, system resources can be used more efficiently to effect change.

What Unintended Consequences Are Likely to Emerge?

Unintended consequences, by their very nature, are hard to anticipate, but they are often easy to identify once a PBAS is implemented. Usually, service providers are happy to inform program sponsors about the negative consequences they are experiencing. Early evaluation might enable system planners to make changes that reduce the likelihood or extent of unintended consequences. While it might be impossible to eliminate all such effects, learning about unintended consequences early in the implementation period can help to mitigate them.

---

[1] William Gormley told us that financial incentives do need to be large enough to matter: In an unpublished paper on child-care reimbursement schemes, he found that states that offered a quality premium of 15 percent or more were more successful in encouraging NAEYC accreditation than states that offered only 5 percent more to providers.

What Contextual Factors Might Hinder the System's Implementation?

Although PBAS designers are usually well acquainted with their sector, they might lack the more refined knowledge that would enable them to predict how well a PBAS might fare in different settings (e.g., in child-care centers versus homes; in rural versus urban transit settings; in states with strong teacher unions versus states with weaker teacher organizations; in communities or regions with large, strong physician organizations versus those with small, unorganized physician practices).

Possible Reasons for the Limited Evidence About System Effectiveness

We have just described many insights that could be provided by PBAS evaluation; yet, as explained in Chapter Seven, the incidence of PBAS evaluation is limited. Our examination of PBASs and their evaluations led us to formulate a number of hypotheses about why there are so few studies. In this section, we discuss several potential barriers to evaluation; virtually all apply in varying degrees to evaluation of any public policy—not just a PBAS—and many have been commented on in policy literature more generally. Later in this chapter, we explore ways in which PBAS evaluation might be made more feasible.

Political Climate Might Discourage Close Scrutiny of Performance-Based Accountability Systems

In order to marshal the necessary political support to adopt a new policy, particularly a policy like a PBAS that might require substantial resources and often will subject previously unmeasured processes to public scrutiny, a strong argument must be made for its value. To line up the necessary political support, advocates might feel that it is necessary to assert that the policy *will* result in the achievement of the goals it is designed to address, rather than presenting the policy as a logical approach, a good bet, or a best-available option. In the process of winning support, there might be little room for expressing doubt or uncertainty.

Such unwillingness to admit doubt is unfortunate because new policies, particularly policies designed to produce change, should be thoroughly scrutinized to ensure that they are as effective as possible. Indeed, healthy skepticism is probably an appropriate and productive stance to take toward any new policy: It creates an environment that supports careful and unbiased assessment of all aspects of the new policy and its implementation.

When the adoption of a PBAS represents the successful conclusion of a contentious and hard-fought political battle, the decision-makers (i.e., politicians) might come to associate "victory" with the mere creation of the PBAS, leaving little enthusiasm for evaluating its long-term impact. Even when a PBAS can be implemented without a political battle (as is sometimes the case with private-sector health-care initiatives), the substantial resources required to design and implement a PBAS might leave little appetite for evaluation. If the PBAS succeeds over time, there might be positive acclaim for its sponsors in response to a positive evaluation, but a positive outcome is uncertain. If it fails, the efforts of PBAS proponents risk appearing both ineffective and wasteful. Under such circumstances, the tendency to "declare victory now" is strong. In some cases, the PBAS is simply not viewed as a testable policy. For example, in education, NCLB represented a hard-fought change in the underlying philosophy about the federal role in educational governance; it was not perceived as an experiment to test a new approach. Under such circumstances, evaluation can be an afterthought. These examples demonstrate the reasons that policymakers might be uninterested in funding evaluations of their own programs; once the programs are established, policymakers have little interest in finding out whether, in fact, they "worked."

Call for Evaluation Might Be Considered a Sign of Failure

Bringing evaluation to bear on a new policy, such as a PBAS, acknowledges that the current approach is but one of many possible alternatives (both different permutations of a PBAS and completely different approaches) for achieving the system's goals; a particular PBAS was selected because it seemed the most efficient or productive approach, but there could have been many others. This perspective is important

because, if the evaluation shows that the adopted approach is not working well, there will be less inclination to reject the entire idea out of hand and more willingness to reexamine which aspects were less effective, in an effort to refine the policy or its implementation.

However, policymakers might be unable to appreciate the value of learning which elements are successful and which are not. Instead, they might view evaluations as inherently risky because programs might not succeed as planned. In health care, there has been a strong national push to align provider incentives with the delivery of better care; however, to date, the evidence has shown relatively modest positive effects. While advocates claim that they want to improve the effectiveness of their programs in the hope of achieving more-robust results, few health-care PBASs are evaluating the impact of modifying various design elements in pursuit of this goal. Politicians who have put their prestige and power behind a new policy might not want to learn that it is not working, even if an evaluation might help to pinpoint places where changes might alter the policy's trajectory and improve its ultimate outcomes.

This reluctance to uncover failure might reflect fears that stakeholder buy-in, which often is very difficult to achieve, might be lost. It also might be exacerbated in some sectors because the PBAS represents a response to other, failed approaches. Under these circumstances, policymakers might be even more failure-averse than usual. However, experience in the health-care sector suggests that the wish to avoid failure can be overcome; in the Medicare program, the failure of an earlier investment in quality-improvement technical assistance to physicians led to the adoption of P4P and to support for its evaluation. While researchers might argue that there is no shame in studying the implementation of a policy and discovering that its effect was neutral or even negative, policymakers and politicians often do not share that view.

Stakeholders Might Be Comfortable with the Status Quo

Evaluation might also be rejected when evaluation findings might undermine a carefully negotiated division of power or resources among stakeholders or call into question an approach that garners more political support than others. For example, Congress recognized that the

goal of greater fuel efficiency might be achieved through higher gas taxes or new CAFE standards. The latter approach was far more politically acceptable because it did not require new taxes; no one wanted to find out that it did not work, because the alternative could not achieve political traction. In transit, substantial sums are up for grabs in the design and implementation of PBASs. Once agreements are reached about how these funds are to be distributed across communities and entities, providers are reluctant to undertake a study that might call into question a funding allocation scheme that required much discussion and compromise. In health-care PBASs, the bulk of incentive payments frequently go to historically high-performing providers who often have made little or no improvement but are the largest and most influential players; evaluations have shown that providers who made the largest gains were rewarded least or not at all (Damberg, Raube, et al., 2009; Rosenthal et al., 2005).

Such findings have raised questions about the merits of design structures that reward top performers at the expense of those that demonstrate improvement. As a consequence, some incentive structures have been modified to reward improvement, changing the distribution of resources among providers and upsetting the status quo. When service providers become comfortable with an existing resource-distribution formula, their interest in evaluating the degree to which the funds are effectively addressing the issues they were designed to address might wane. Assessing the value and use of the PBAS might force allocation decisions to be reopened and the challenging distribution process to be restarted.

Evaluating Performance-Based Accountability Systems Is Challenging

Another potentially important reason there are few studies of PBAS effectiveness is that such research is challenging: It is difficult to study the operation of complex, dynamic systems, such as PBASs. For example, to study the effects of a child-care PBAS on kindergarten readiness, children must be followed over time and receive care in a given program during the entire study period. If children change providers during the study or leave formal child care entirely, their outcomes

are no longer a fair test of the effects of a given provider on children's outcomes. Yet, high attrition rates are common in child-care settings as families move, parents lose or change jobs, and children age out of care. Extremely high attrition undermined the power of one of the few studies that attempted to assess a PBAS's effects on ultimate child outcomes (as it has undermined other research on child-care programs) (Zellman et al., 2008). There are also technical challenges, as discussed in the following paragraphs.

Lack of Baseline Data. Many PBASs are implemented without the foresight or political will to collect baseline data, incorporate research designs into implementation plans, or phase in the system over time or geography. In health care, most PBASs have not captured performance data in time periods prior to the adoption of the PBAS (i.e., baseline data) and are thus handicapped in their ability to discern whether the policy change has led to improvements relative to what was occurring in the pre-PBAS period. In addition, health-care PBASs tend to offer program incentives to all providers that contract with a PBAS program sponsor at the outset rather than phasing in the program over time or geographic locations. This makes it impossible to assess program effects compared with usual practice. As a result, evaluation designs that call for these sorts of data might have to be rejected in favor of more costly or less powerful approaches later on.

It is sometimes possible to use historical data to estimate the impact of a PBAS or other intervention by comparing the pattern of effects over the period prior to implementation with the pattern after implementation. If the PBAS was the only thing that changed, it is reasonable to infer that any significant differences in measured outputs were caused by the new system. However, such longitudinal (or time-series) designs are difficult to achieve in practice; it is difficult to guarantee that the PBAS will be the only thing that changes during the period under study. Social systems are just too complex and volatile for this assumption to be valid in most cases. For example, the Medicare program implemented its hospital P4R program at the same time as it was launching its hospital P4P demonstration program. At the end of the three-year intervention, it was unclear how much of the observed improvement was due to the effects of publicly reporting per-

formance scores and how much was due to providing financial incentives for engaging in targeted behaviors. To add complexity, the P4P intervention occurred at the same time that the Joint Commission, the hospital-accrediting body, was focusing on the same performance measures within its accreditation review.

Similar challenges occurred in the education sector. During the time that the California Department of Education was implementing its Class Size Reduction initiative, it also implemented the Reading Initiative, Teaching Reading program, and the Mathematics Program Advisory, making it difficult if not impossible to attribute changes in student outcomes to any one program (Bohrnstedt and Stecher, 1999). In addition, the program was voluntary, but it was available statewide, and almost all districts agreed to participate. This rapid uptake eliminated the possibility of making treatment-versus-nontreatment cross-sectional comparisons.

Changes in Output Measures. Another technical problem in comparing program outcomes over time is that, in general, the same output measures are rarely used, unchanged, for years prior to and after the implementation of the PBAS. In health care in particular, new knowledge regarding best practices in the diagnosis and treatment of many conditions leads to changes in performance measures over time. For this reason, it can be difficult to track the identical performance measure prior to PBAS implementation and afterward.

Lagged Implementation for Pilot Testing. One of the less costly approaches to evaluation of PBAS is to conduct pilot studies or implement the PBAS in an incremental way. Such was the case with A+B contracting; DOT supported numerous trials before allowing this form of contracting to be used more broadly for federally funded transportation projects. These approaches enable jurisdictions or activities not yet affected by the PBAS to serve as untreated comparison groups against which the implementation and outputs of the PBAS can be assessed.

However, lagged implementation might be politically difficult, particularly if the PBAS comes with substantial resources. Jurisdictions resist being held back from implementing a policy that promises to bestow a range of benefits, even if these benefits come with risks. It is for this reason that the Medicare program, once it has decided to

implement a policy change (e.g., after a small-scale pilot-test)—such as a PBAS—does not implement it in an incremental fashion, but rather makes the policy change universal to all qualified providers.

Differing Conditions During Pilot-Testing. Even when programs are subject to pilot-testing, the conditions under which the pilots are implemented might differ in potentially important ways from the conditions likely to be in place when the PBAS is implemented at scale. A key example might be found in the health-care sector. The Medicare program developed its hospital P4P program and piloted it in 270 hospitals (out of 3,000 Medicare-participating hospitals nationally). The pilot program structured the incentive payment as "new money" that was not withheld from existing payments. This design feature created no risk for participating hospitals and allowed them to generate additional payments under the Medicare program if they performed well. In contrast, plans for rolling out P4P nationally at scale calls for a budget-neutral approach to financing the incentives, implying some form of withholding. The withholding approach might lead to different behavioral responses from what was observed in the pilot demonstration because this approach imposes risk and reduces benefits to hospitals.

In some cases, the very nature of a pilot makes it impossible to duplicate the conditions of a full rollout. For example, a pilot study of a child-care PBAS in a few geographically isolated counties in Ohio did not include the costly public relations campaign that was planned for statewide rollout because it was not feasible to launch the campaign in just a few small counties. But public awareness is a key component of child-care PBASs' logic model.

Sometimes, pilot programs draw participants in a different way than PBASs at scale, which also limits the conclusions that can be drawn from a pilot. For example, when PBAS pilots are implemented on a voluntary basis, the comparison between participants and nonparticipants can be biased by unknown factors that cause some units to volunteer to participate while others choose not to. Volunteering is often related to effectiveness, so comparing volunteers with nonvolunteers does not give a fair test of the impact of the PBAS if it were to be implemented widely. This selection effect is a serious threat to

the accuracy of implementation research. For example, the Medicare-sponsored P4P demonstration project, the Premier Hospital Quality Incentive Demonstration, allowed interested hospitals to volunteer to participate in the program. But whenever providers choose whether to participate, one must assume that those who opt in are different, in potentially important ways, from those who opt out. This selection bias in the set of hospitals that agreed to participate meant that the Medicare program has had to, on a post-hoc basis, attempt to match participating hospitals to nonparticipating hospitals (which can never be assumed to be totally effective in ruling out differences) to determine whether P4P had an effect.

Uncertain Funding Makes Pilot-Testing and Program Staging Less Appealing

There are other reasons as well that have been offered for not conducting pilot tests or staging the implementation of a PBAS—two relatively easy ways to build evaluation into PBAS implementation. A key reason noted more than once in the child-care sector among respondents in five states that pioneered QRISs is the fear that the funds to make the PBAS possible might go away during the course of a pilot project. For example, a supportive governor who has expressed strong support for a child-care QRIS might not be reelected; if he has indicated that he will work to get funds for a QRIS, it would be seen as folly to ask for a small fraction of the ultimate funds needed to go statewide to run a small pilot. What if the governor is not reelected? What if the pilot project is not completely successful? The general view in these states was that program advocates should get as much money as possible now and implement quickly and statewide, while the opportunity is there (Zellman and Perlman, 2008).

Cost Is Also a Key Barrier to Evaluation

Particularly in sectors in which data are not routinely collected, adding the costs of evaluation to the cost of PBAS implementation might be politically infeasible. While we argue that evaluation should be built into a PBAS from the beginning, this rarely happens, and the cost of evaluation is often viewed as a pricey add-on to a policy that itself

might be meeting opposition because of its high operational costs. Furthermore, small programs might not be able afford the cost of evaluation; as a result, they might not be evaluated unless some level of government steps in to fund the activity.

Often lost in these discussions of cost is the reality that continuing to implement a policy that is not working is the most costly approach of all: Not only are operating funds being wasted to carry out the PBAS or another policy, but potentially more-effective policies and approaches are precluded from being implemented.

Some Sectors Are More Focused on Service Provision Than on Experimentation

In some sectors, there is less demand for rigorous evaluation because the culture is more engaged with providing services than with scientific experimentation. For example, in the child-care sector, the implementation of PBASs has not provoked widespread calls for evaluation. The underlying logic model of the QRIS seemed to make sense, and providers were accepting of it on face value. They did not demand scientific proof; indeed, most advocates embraced the concept as a way to infuse both resources and support for improvement into a sector that had long had little of either. When the major challenge in the sector is attaining a basic level of services, providers might be less worried about *maximizing* the efficiency of the limited services that might exist. This is not to suggest that quality is unimportant or that key advocates have not pushed evaluations; rather, the first priority is to serve a greater proportion of those who need service, and increasing efficiency receives a lower priority.

In general, a preference for spreading intended benefits as widely as possible can undercut efforts to evaluate PBAS effectiveness (as well as program impact more generally). Judging the effectiveness of a PBAS (or any program activity) requires an estimate of the counterfactual condition: What would have occurred in the absence of the change? The best way to estimate this control condition is to have some agencies operate without the PBAS while others are subject to the system. If these two groups are chosen to be comparable with respect to all factors related to the operation of the PBAS, then the program effect can

be estimated by comparing the effects under the PBAS with the effects absent the PBAS. Unfortunately, when programs are implemented universally, this comparison cannot be made, and the effectiveness of the PBAS is difficult to determine.

This was the case in the California Integrated Healthcare Association's P4P program, which was implemented universally in all 225 medical groups with which the participating plans contracted. In the United Kingdom, the government modified its contract with primary-care physicians to provide a P4P program, then universally applied this contract to all general practitioners. This was also observed with the CAFE standards in transportation, which were applied simultaneously to all large auto manufacturers. The same problem exists with NCLB, whose reporting requirements apply to all schools and whose interventions must be applied to all Title I schools. To be fair, a case can be made for moving to universal implementation as quickly as possible so services are not denied to any deserving clients, but this argument breaks down when programs or systems are experimental and their efficacy still needs to be established.

All of these factors that operate against evaluating PBASs (and other policies) are relatively easily understood, but acquiescing to them would be a mistake. Despite often daunting logistical and political obstacles, evaluation of PBASs is important if policymakers and system designers are to learn how to improve the systems and, perhaps, come up with even better ways to meet their public policy goals.

Making Accountability and Evaluation More Appealing

Given the range and number of barriers to the design and implementation of evaluation, and the value of evaluation to maximizing PBAS potential, it is important to address evaluation barriers at the beginning of the PBAS design process. This section presents a number of approaches that might make the notion of evaluation more appealing and increase the likelihood that accountability systems will, in fact, be held accountable. Much of what follows is informed speculation; evalu-

ation was infrequent in the cases we examined, and there was little direct evidence about ways to promote it.

Reframe Performance-Based Accountability as One of Several Policy Options

As noted earlier, one route to winning acceptance for a new policy or approach is to make it seem like the clear and dominant choice. However, this is rarely if ever the case: A new idea, such as a PBAS, is not the only approach to improving quality, and any given PBAS is just one of many permutations of the PBAS idea. It might be effective to introduce a PBAS as one of several policy options, with evaluation as a necessary element to improve future options.

Embed the Evaluation Design into the System's Implementation

To the extent that the evaluation design can be subsumed in the rollout of the PBAS, the odds are greater that the evaluation can be carried out as planned. An evaluation that is embedded in the implementation of the PBAS might be more resistant to attack if cost or other factors become the focus of debate.

The ability to embed an evaluation design into the PBAS might depend on features of the sector in which it occurs. For example, in such sectors as transportation and health care, in which measures are routinely collected, it might be possible to develop an evaluation design that relies heavily on already-collected data. This reduces the cost and vulnerability of the evaluation. In contrast, in a sector like child care, in which there was little previous routine data collection, embedding evaluation in implementation is much more challenging. In this case, it might be more effective to try to capitalize on data that need to be collected to run the PBAS, e.g., data on staff education and training and data on provider ratings. Alternatively, evaluators might work with PBAS designers to embed into the system design and operation the collection of data that will later be needed for evaluation.

Propose Evaluation Designs That Are Developmentally Appropriate

In most of the cases we studied, the PBASs were complex, and full implementation took considerable time. At the extreme, all of the ele-

ments required by NCLB were not fully implemented in some states for five or six years. Until the system is operational, an assessment of PBAS outcomes might be premature, and an evaluation based on outcomes is less likely to demonstrate success (e.g., Zellman et al., 2008). Instead, it might be advisable to develop a multistage evaluation design that aligns with the stages of PBAS implementation. This developmental notion is not unique to PBAS evaluation. Indeed, the U.S. Food and Drug Administration (FDA) has a well-established protocol for drug testing designed to minimize risk by specifying a progression of tests (e.g., animals before human subjects). This developmental approach might be considered one part of a general strategy to encourage policymakers and stakeholders to understand evaluation in a more nuanced way. (See Chapter Seven for examples of developmentally appropriate evaluation in child care and health.)

Create a Realistic Evaluation Timeframe

Evaluations should be conducted according to a schedule that is linked to the PBAS's developmental phase. Early on, evaluations can be focused on improving system functioning. Once assessments demonstrate that the system and its components are working, evaluation can broaden, potentially assessing the system as a whole. A realistic timeframe is likely to cover at least five years; the timeframe can be longer if the PBAS becomes a successful learning organization that makes changes over time, because those also need to be assessed. A well-designed evaluation plan will produce many findings along the way and contribute to a better-functioning PBAS. The National Longitudinal Study of NCLB reported interim findings in four separate reports in 2007 based on an initial round of data collected in 2004–2005 (e.g., Le Floch et al., 2007); final reports were issued in 2009 and 2010 based on a second round of data collected in 2007–2008 (e.g., J. Taylor et al., forthcoming).

Assemble Sufficient Evaluation Expertise

Lack of evaluation capacity (or the perception of limited capacity) decreases the likelihood that a PBAS will be evaluated. Similarly, lack of evaluation sophistication can result in evaluations that fail to ade-

quately address key questions. One way that policymakers who have a stake in good evaluations might be able to encourage local implementers to implement evaluations is to provide a ready source of design expertise. For example, QRISs in child care are decentralized; in general, states undertake them and evaluate them (although, in some states without QRISs, localities have stepped up to design and implement QRISs). The agencies that oversee their operations might not be sophisticated with respect to evaluation. A centralized effort, perhaps funded by the federal government, could provide evaluation models and support for evaluation. In fact, the Office of Planning, Research and Evaluation (OPRE) in HHS's Administration for Children and Families has taken on some of this function through its recent support for a consortium designed to bring together researchers to discuss QRIS research in progress and develop models for evaluation of QRISs.

A centralized initiative to assemble evaluation expertise or conduct cross-site evaluations might also lead to more-sophisticated evaluation designs that capitalize on natural variation across sites and states. In the child-care sector, for example, each state is conducting its own evaluation, typically relying on a local university department to mount the effort. These efforts are largely conducted in isolation, and their sophistication is limited by small budgets and lack of variation on key dimensions given statewide implementation. A centralized evaluation initiative might encourage states to pool data on key PBAS elements, which could increase variation and increase the statistical power of these efforts. Of course, there are potential downsides to centralization of evaluation if the models imposed on individual sites are not sensitive to the local conditions that might influence PBAS operation.

In education, the enactment of NCLB represented a major change in policy for the federal government, and Congress mandated that a national evaluation be conducted under the direction of the Department of Education. Resources were set aside to conduct the National Longitudinal Study of NCLB to assess the implementation of the law, leading to a more comprehensive and detailed study than would have been possible had this task been left to the discretion of the individual states. Similarly, the federal government awarded a number of Teacher Incentive Fund (TIF) grants for districts to experiment with P4P com-

pensation reforms. Each project was required to have its own evaluation to assess project outcomes, as well as implementation, but the approaches to evaluation have varied widely across sites, making it difficult to synthesize the findings. The federal government also funded a national evaluation intended to produce programwide findings regarding implementation and outcomes.

Interest an Outside Funder in Supporting an Evaluation

In some sectors, there might be independent organizations interested in studying the changes brought about by PBASs to determine whether this approach to public policy has merit. For example, in health care, several foundations provided funding support under the Rewarding Results initiative to ensure that the P4P reforms were evaluated. The funding agency required the program sponsors that received foundation funding for program design and implementation to have an independent evaluator assess the program's impact. More generally, our analyses suggest that government agencies or large philanthropic foundations could advance our knowledge of PBAS effectiveness by assembling relevant databases and by developing evaluation templates that could help local jurisdictions and service providers evaluate their PBAS efforts.

Designing Evaluations to Maximize Their Value

Having argued for the importance of evaluation and offered suggestions for generating support and resources for it, we suggest four elements of evaluation design that would make efforts more successful: creating a separate mechanism to assess the validity of reported gains, including an appropriate comparison group, embedding the evaluation design into the PBAS's implementation, and using a logic model.

Consider a Separate Measure to Assess the Validity of Reported Gains

PBASs have been successful in many instances in motivating behavioral change among targeted providers, leading to changes in measured out-

puts. However, it is not always clear whether improved outputs reflect real movement toward the goal rather than gamesmanship or teaching to the test motivated by the PBAS's incentives. Having a separate measure that is not subject to incentives offers one way to assess the validity of reported gains. Such audit mechanisms are often important when evaluating PBASs because measured outputs do not always align well with long-term goals and because behaviors that promote the former might not encourage the latter.

The education sector provides a good example of the problem and the use of an alternative measure. While improved student test scores are the desired effect of education PBASs, research (and common sense) suggest that better student performance represents some combination of improved student learning (the desired outcome) and other things, including better test-taking skills and selective inclusion of test-takers. The desire to look good is far less important when students take tests that are not subject to PBAS scrutiny and incentives. Comparison of performance on incentivized and nonincentivized exams enables evaluators to assess the degree to which improved student performance is real and not an artifact of the PBAS (Podgursky and Springer, 2008). Educators are reluctant to devote more class time to testing, so it is uncommon to have students take both incentivized and nonincentivized tests. In one such study, Winters et al. (2007) found statistically significant gains in math scores on a nonincentivized test among students whose teachers were participating in a P4P initiative.

Include an Appropriate Comparison Group

In many cases, it is possible to identify an appropriate nonparticipating group against which PBAS effects can be compared. Measuring changes in the performance of this group clarifies how much of the change in the PBAS group should be attributed to the PBAS and how much might be due to outside factors.

Comparison groups might be created in a variety of ways, including random assignment of providers or clients to the PBAS or to usual procedures (the gold standard discussed in more detail in Appendix B). Phased-in implementation of the PBAS among providers is a way to permit comparisons without excluding anyone from eventual partici-

pation (see Appendix B for further discussion). Selection of comparison groups most likely will be constrained by politics, time, and cost. It is important to identify the most likely alternative explanations for apparent effects of a PBAS and, if possible, create comparison groups that allow those alternative explanations to be ruled out. For example, if it is widely believed that hospitals engage in quality-improvement initiatives in response to both public reporting of their quality and because they want to gain the rewards and avoid the penalties associated with an operating PBAS, a good evaluation of the PBAS argues for a public-reporting-only comparison group that does not receive the other incentives. This group will allow any behavioral changes motivated by public reporting alone to be factored out. A study by Lindenauer et al. (2007) did just this and found that, indeed, both the public-reporting-only and public-reporting-plus-PBAS groups of hospitals improved; controlling for baseline performance and other characteristics between the groups, the study found that the PBAS group improved more (about 2.7 percentage points higher) on most, but not all measures.

Embed Evaluation into System Implementation

As discussed earlier, embedding evaluation into the implementation process protects the evaluation function and provides the PBAS with far more timely feedback. This timely feedback, if used to change the system, might increase the odds that the PBAS will succeed later on. For example, a health PBAS assessed the degree to which doctors were aware of the PBAS and considered its incentives to be adequate to affect their behavior (Teleki et al., 2006). Feedback revealed low levels of awareness and near consensus concerning the inadequacy of system incentives. This gave the PBAS invaluable information quickly on which to base program modifications. In this case, the formative evaluation helped with the interpretation of the summative evaluation of the program's impact; without it, the evaluation would have erroneously concluded that the PBAS failed to produce outcomes because of a flawed concept rather than inadequate incentives and physician notification efforts.

Consider Using a Logic Model

Another worthwhile approach to evaluating a PBAS is to create a logic model, a systematic and visual diagram of the relationships among a program's resources, planned activities, outputs, outcomes, and impacts. This diagram articulates key processes and relationships considered necessary to achieve ultimate goals (Wholey, 1979; Rossi et al., 2004; Chen, 2005). Such evaluations can be far simpler and less costly to conduct if the program has developed a logic model. With a clearer understand of the expected links among program inputs, processes, and outputs, evaluations are more likely to produce worthwhile data on a PBAS's implementation and short-term prospects and outputs. PBASs in health care generally do not develop formal logic models. But there is general consensus about changes that must occur before a PBAS can be considered successful. For example, many health-care PBASs aim to change physician behaviors, which are anticipated to lead to gains in the desired outputs and outcomes. If an evaluation assesses physician behavior and determines that it has changed in response to the PBAS, a key condition for success has been met. Similarly, health-care PBASs often require the capacity to track physician behaviors. Evidence that a PBAS has resulted in improvements in data collection and monitoring (e.g., widespread installation and use of electronic data) could represent an interim success as well.

Chapter Summary

Despite the widespread proliferation of PBASs, there has been little rigorous research in the cases we examined about their functioning, outputs, or outcomes. The reasons for the lack of empirical investigation of these systems are varied and, to some degree, understandable, but the failure to examine these systems empirically remains an ironic reality: Systems designed to hold service providers accountable have largely avoided being subjected to accountability themselves. The dearth of evaluation has also held back progress in refining PBASs.

The reasons to evaluate are many; careful planning and strategizing could produce more evaluations better aligned to policymaker needs

and PBAS requirements. The reluctance to evaluate is understandable given what we know about policymaking and policy implementation. Yet, there are approaches that might make evaluation more attractive to stakeholders with an interest in PBASs. A number of evaluation designs are available that could be matched to the needs, context, and available resources in a given sector. For example, matching evaluation design to stage of implementation is most likely to produce usable, cost-effective evaluation outcomes and provide PBASs and competing quality-improvement approaches with a fair test of their merit.

CHAPTER NINE

Conclusions

We are now witnessing, across multiple service sectors, the development of PBASs, in which those responsible for overseeing a service-delivery activity establish a set of measures for gauging the quality of the service and a set of incentives—rewards or sanctions or a combination thereof—applied to service providers on the basis of measured performance, as a means of stimulating improvement.

The management literature is replete with normative, and often theoretical, advice on the institutional structuring of PBASs and the design of measures and incentives. Additionally, although to a lesser extent, the literature includes critical examinations of specific PBAS cases in different sectors. What has been missing to date is an effort to review multiple PBASs across different sectors and develop empirical and synthetic observations of how these governance structures work in practice, with special attention to the question of how contextual factors contribute to or hinder success.

Our goal in this study has been to conduct such a review, drawing on specific cases from child care, education, health care, PHEP, and transportation. In examining PBASs from these sectors, we have considered a broad range of potentially illuminating questions. What circumstances contribute to the introduction of a PBAS? How are goals for the PBAS established, given the often competing perspectives of different stakeholders? How is the PBAS integrated within the existing governance structure? How are measures selected and implemented? What incentive structures are developed, and to whom do they apply? What implementation strategies are pursued, and is the

system designed to evolve over time? Is the PBAS effective in achieving its goals? Are the goals achieved in a cost-effective manner? Would other governance or regulatory approaches prove more cost-effective? Do negative unintended consequences arise, and how might they be avoided? Finally, and as a recurring thread running through many of these questions, how do contextual factors influence the adoption, design, implementation, and functioning of a PBAS? While our review of available evidence has provided insight into many of these questions, certain issues—such as the cost-effectiveness of a PBAS approach relative to other policy approaches—would benefit from further core research.

The preceding chapters in this monograph have delved into a variety of design and implementation issues in greater detail. In this concluding chapter, we do the following:

- Describe high-level findings on the structure and functioning of PBASs.
- Distill potentially helpful recommendations for decisionmakers considering the design and implementation of a PBAS.
- Highlight areas of remaining uncertainty that would benefit from further research.

Observations on the Structure and Functioning of a Performance-Based Accountability System

As evidenced by the blossoming of PBASs in the five sectors examined in this monograph and in many other sectors as well,[1] the idea of greater accountability in publicly oriented services has considerable current traction in policy circles. Based on our analysis of available evidence, it is not possible to conclude that PBASs have fulfilled their

[1] Performance-based service acquisition, for example, has gained traction at all levels of government in the United States for most of the types of services that government agencies buy. Award and incentive fees reward good performance in the provision of services ranging from facility services to weapon-system maintenance to support deployed forces.

perceived potential in all contexts. We can, however, offer the following observations.

Performance-Based Accountability Systems Can Be Effective in the Right Circumstances

Perhaps the most unambiguous example of a successful PBAS is the A+B contracting example from transportation (CAFE standards, also from the transportation sector, are also generally viewed as a success). A+B contracting has led to dramatic reductions in the time required to complete road repair and construction projects. And, provided that the magnitude of the financial incentives is commensurate with the value to society of reducing the duration of construction-related traffic congestion, it can be claimed that A+B contracting achieves its ends in a cost-effective manner.

Yet, A+B contracting presents, in many ways, a "best case" set of circumstances for designing and implementing a successful PBAS. Consider the following:

- The goal of reducing construction time and, in turn, traffic congestion, is widely shared among decisionmakers and the public.
- The performance measure—days to complete—is relatively unambiguous and generally easy to observe. Disagreements might arise, of course, as to when a project is actually "complete" or whether completion can or might not be "conditional" upon further work that can be done after the roadway opens to traffic. Within practical limits, however, these performance metrics are relatively clear and actionable in relation to those that apply in many other case studies.
- Those held accountable under A+B contracting—the construction firms—have near-complete control over the relevant inputs and processes involved in road construction and thus can reasonably be held to account for the outputs (certain factors that might delay completion, such as adverse weather, fall beyond the control of the construction firm, but contracts are often structured with clauses to absolve the construction firm for problems that arise due to such circumstances).

- A+B contracting operates within a broader, ongoing regulatory framework that ensures that construction firms do not increase construction speed in ways that have adverse consequences. For example, highway departments closely monitor work as it is executed to ensure that construction firms achieve mandated engineering standards.
- Contracting firms are private-sector entities, thus relieving concerns over whether there is a level playing field among different competitors. Further, as for-profit entities, contracting firms are inherently motivated by financial incentives.[2] There is thus no need to worry that the application of financial incentives will have some adverse effect on any intrinsic motivation of service providers.
- Though requiring nontrivial transportation and economic modeling, estimates of the social value of mitigating traffic congestion provide a rational basis for setting an upper bound on the magnitude of the financial incentives.
- The relationship between a public transportation agency and a contracting firm is governed exclusively by the contract. Thus, the effectiveness of the A+B program is not undermined by other, potentially competing or contradictory, programmatic interests of requirements.

Most of the other PBASs that we examined in this study exhibit greater complexity in one or more of these dimensions, making it harder to determine the best design or implementation strategies for improving the PBAS and judge whether the PBAS has been successful in its aims. Consider these examples:

[2] While most highway contractors are for-profit entities, there are also examples of performance-based contracts with private, not-for-profit organizations in other sectors, such as employment manpower, training, and placement programs. In such cases, financial rewards might still be important, but other factors might come into play as well. See Liner et al. (2001) for examples and discussion.

- With education, child care, and transit service, there are multiple and often conflicting goals of interest and relatively little consensus regarding their priorities.
- For both child care and PHEP, the measurement of performance poses significant and largely unresolved difficulties. With the CAA, the performance metrics are based on modeled emission forecasts for future years, raising important concerns regarding the accuracy of the underlying models and assumptions.
- In child care, education, and health care, those held accountable have only partial influence over the outputs of the service-provision process. Some students are less prepared or motivated to learn than others, or face more-difficult home environments; likewise, in health care, some patients are in poorer health to begin with, or are less likely or able to comply with a physician's instructions. This condition of "unequal playing fields" raises significant concerns regarding the fairness of a PBAS and, in particular, introduces a potential conflict between the interests of service providers and those of end users. If end users are given higher priority, the PBAS might be designed to focus on output measures, regardless of the degree to which service providers can influence the outputs. Such a decision, however, is almost certain to increase service-provider resistance to the PBAS. To achieve greater fairness from the service provider's perspective, the PBAS might instead adopt performance measures based on inputs, on structure, or on processes rather than on outputs. Or, as in health care, output measures might be adjusted for differences in the mix of the population served.
- Perhaps most notable in child care, education, and health care, but in other examples as well, a strong case can be made that service providers are already intrinsically motivated to work to the best of their ability. There is some concern that an excessive focus on competitive financial rewards in such circumstances could undermine rather than bolster this motivation.
- In many of the cases we examined, the PBAS was just one of multiple—in some cases, competing—programs intended to influence or govern provider behavior. When these programs are

not aligned or have conflicting goals and measures, the strength of the signal provided by the PBAS's incentive structure is weakened.

The comparison between A+B contracting and the other PBASs serves to reinforce the fact that, while PBASs can be instantly effective in the right circumstances, it is rare for all of those circumstances to be present. In most of the service sectors we studied, PBAS designers had to contend with one or more of the challenges just described.[3]

Performance-Based Accountability Systems Motivate Some Changes in Service-Provider Behavior

While few of the PBAS examples we examined could be characterized as unqualified successes in terms of achieving their purported goals, most of the PBASs did stimulate at least some changes in service-provider behavior and in desired outputs. In health-care P4P programs, doctors devote additional effort to improve performance against incentivized quality-of-care measures, e.g., by investing in information systems, increasing quality-improvement resources, and focusing staff on ensuring that patients come in for recommended care. With NCLB, teachers allocate more instructional time to mathematics and reading. In child-care QRISs, providers take concerted steps, within the constraints of available resources, to improve their ratings. In response to CAFE standards, auto manufacturers modify the composition and pricing of their fleets to achieve specified mileage targets.

A System's Structural Details Strongly Influence Providers' Responses

PBASs are designed to support broad social goals—for instance, producing a productive and engaged citizenry in education, reducing morbidity and mortality in health care, or reducing future aggregate fuel consumption with CAFE standards. It is typically difficult, if not impossible, to measure progress against these broader goals in a timely manner; some goals are sufficiently ill defined as to defy measurement,

[3] These factors also hinder the success of more-conventional management systems, such as traditional government oversight.

while others might not unfold until many years have passed. As a result, a PBAS must instead focus on near-term performance measures that serve, at best, as proxies for the broader social goals.

It is reasonable to expect that service providers will bend their efforts toward improving performance on these specific measures—possibly at the expense of other activities that might also serve broader sectoral goals. The strength of the provider response, in turn, will depend on the magnitude of the incentives in comparison to the cost of changing behavior in the desired manner. One can thus assert that provider response is quite sensitive to the structural details of a PBAS: *what* it measured, *how* incentives are applied, and the *size* of those incentives.

In some cases, measures and incentives align to produce the desired outcomes. Here again, A+B contracting serves as a good example; the measure of days to complete a construction project correlates closely with the broader goal of reducing construction-induced congestion delays, and the size of the rewards has been sufficient to motivate construction firms to complete projects much more quickly. Yet, success is far from the norm. Some PBASs might fail to achieve their intended goals, while others might create unintended and undesired consequences. Almost invariably, such failures stem from the detailed structure of the performance measures and incentives that define, with great precision, what the PBAS is asking of providers. Here, the common aphorism is quite relevant: Be careful what you ask for, as you just might get it.

There are at least three ways in which the performance measures and incentives within a PBAS might fail to support the intended outcomes:

- First, the PBAS might not stimulate a significant provider response simply because the incentives are not large enough. In many health-care P4P programs in the United States, for instance, the potential financial rewards represent a very small percentage of overall physician pay and thus might not garner much attention.
- Second, a PBAS might lead to improved performance for a particular measure without corresponding progress for the underly-

ing goal. With NCLB, as an example, teachers might teach to the test—that is, pursue instructional strategies that result in higher standardized math and reading scores—without making comparable improvements in broader student capabilities in math and reading.

- Third, the structure of measures and incentives might give rise to unanticipated and undesired consequences. The cases we examined provide many examples of this. CAFE standards, for instance, led auto makers to produce smaller and lighter cars, and many researchers have argued that this resulted in a higher number of crash-related injuries and fatalities. In health care, there are concerns about ignoring nonincentivized areas of care, as well as dropping sicker patients to score well. With NCLB, the emphasis on math and reading has led many teachers to spend less time on other subjects, thus narrowing the education of students.

The preceding observations suggest that the designers of a PBAS should pay close attention to the structural details and envision, to the extent possible, the type of provider response that the measures and incentives will stimulate. Further, if a PBAS does not initially meet its aims, it might mean that some of the structural details require further refinement.

Initial Success Is Rare, and the Need for Modification Should Be Anticipated

PBASs are dynamic systems that operate in complex real-world environments. As such, it is often necessary to make fine-tuning adjustments—altering, for instance, the measures or incentives—to facilitate improved functioning of the PBAS over time. This highlights the importance of developing a plan for monitoring and evaluating the performance of a PBAS, both to detect problems and to identify strategies for bringing the operation of the PBAS into closer alignment with its goals.

A System's Success Should Be Assessed in Relation to Previous Conditions Within the Sector

In the process of building support for a PBAS, decisionmakers often make optimistic assertions about the benefits that this approach will bring. If, after a few years of operation, these benefits have failed to materialize to the extent promised, there might be a temptation to view the PBAS as a failure. Yet, in most cases, the main reason for introducing a PBAS is that the prior governance structure was failing—often to a dramatic degree—to support sectoral goals. It is therefore useful, when evaluating a PBAS, to compare its results with prior performance in the sector. If there has been some improvement, the PBAS can be judged as at least partially successful, even if it has yet to fulfill its original expectations.

The findings of this study are consistent with the literature describing other innovations in public policy and other evaluations of policy initiatives over more than half a century. The consistency of these findings with those of earlier studies in the public policy and evaluation literature is possibly due to the fact that the fundamental structures of governance evolve very gradually. When Arnold Meltsner, for example, described the public decisionmaking process in his classic work *Policy Analysts in the Bureaucracy* (1986), PBASs were not yet known. But many of his observations about the behavior of analysts and the institutions in which they work are similar to conclusions reached in this study. Similarly, when Pressman and Wildavsky wrote their treatise, *Implementation: How Great Expectations in Washington Are Dashed in Oakland* (the first of three editions was published in 1973), they described the failure of economic development programs in Oakland, California, in comparison with the goals for those programs as conceived in Washington. Many of their insights remain surprisingly consistent with our own conclusions regarding implementation of PBASs in education and health care. Finally, most of our descriptions of the ways in which more-meaningful evaluations could be conducted are completely consistent with the widely read treatise *Handbook of Practical Program Evaluation*, edited by Wholey, Hatry, and Newcomer (2004).

Practical Insights on System Design and Implementation

Designing and implementing a PBAS, as suggested by the preceding, is a complex undertaking, and many of the decisions that will need to be made are heavily dependent on sector-specific contextual circumstances. Still, our analysis suggests a series of recommendations that might assist decisionmakers in the process of considering, designing, or implementing a PBAS. We group these recommendations as follows:

- whether to pursue a PBAS
- high-level design of a PBAS
- designing performance measures
- designing incentives
- implementing a PBAS
- evaluating and improving a PBAS over time.

Whether to Pursue a Performance-Based Accountability System

Despite the current popularity of performance-based accountability within policy circles, it is not evident that a PBAS is always the best approach. Prior to embarking on this path, it is therefore helpful to develop an understanding of why a PBAS might or might not be expected to succeed.

Consider the Factors That Might Hinder or Support the Success of a PBAS to See Whether Conditions Support Its Use. There are no hard and fast rules about the circumstances that would ensure or preclude the success of a PBAS. From the cases that we have examined, however, it appears that the factors listed in this section will tend to support a successful PBAS implementation. If a large share of these factors does not hold for the application under consideration, decisionmakers might wish to consider alternative policy options. Or they should think about ways to influence the context to create more-positive conditions for a PBAS. The supportive factors for a PBAS are as follows:

- broad agreement on the nature of the problem
- broad agreement on the goals that the PBAS should address

- knowledge that specific changes in inputs, structures, processes, or outputs will lead to improved outcomes
- ability of service providers, through changes in behavior, to exert significant influence on outputs and outcomes
- ability of the implementing organization to modify the incentive structure for service providers
- absence of competing programs that send conflicting signals to service providers (or, alternatively, the ability to better align or coordinate the activities of the PBAS with other programs)
- political context in which it is acceptable if the PBAS does not immediately achieve its goals but rather is gradually improved over time
- sufficient resources on the part of the implementing agency to create the PBAS and sufficient resources on the part of service providers to respond to the incentives.

High-Level Design Principles

If a decision is made to develop a PBAS, the following high-level design recommendations become relevant.

Account for Constraints and Leverage Opportunities Presented by the Context in Which the PBAS Will Be Implemented. Our analysis indicates that it is very important to consider context when designing the structure of a PBAS and, in particular, when setting up the measures and incentives. Subsequent recommendations in this chapter provide more-specific guidance, but, at the outset of the design process, it is important to consider the following questions:

- In what ways can the implementing organization alter the incentive structure that service providers face?
- What other mechanisms—for example, safety requirements, licensing or accreditation requirements, restrictions on local use of resources, and the existence of other incentives—will continue to frame service providers' behavior, and to what extent do these mechanisms support or hinder the PBAS?
- What mechanisms are already in place that can be used or modified to support the PBAS—for example, performance informa-

tion that is already being collected, mechanisms to insert performance information into existing personnel management and compensation systems, and budgetary systems?
- How much is known about the service activity to be subjected to a PBAS, and what implications does this have for the types of performance elements that might be monitored—for example, process measures versus output measures?

Consider Applying Performance Measures and Incentives at Different Functional Levels Within the Service Activity. This is especially important for hierarchically organized service-delivery activities. In education, for example, it might be helpful to set up different performance measures and incentives for school districts, school principals, and teachers. In health care, separate measures and incentives might be set up for hospitals or clinics and doctors. Designing a PBAS to span multiple levels of a service-delivery activity can be viewed as appropriate when different parties contribute to the overall outputs or outcomes of interest and those parties are able to influence different aspects of the service-delivery process. Provided that the performance measures and incentives are structured in a complementary fashion, the results can be additive and mutually reinforcing.

Design the System So That It Can Be Monitored over Time and Improved as Needed. As already noted, the design and operation of a PBAS is a complex undertaking, making it difficult to get everything right on the first attempt. To obtain the best results over the long term, therefore, it is important to develop a plan for monitoring the PBAS, identifying shortcomings that might be limiting the effectiveness of the PBAS or leading to unintended consequences, and modifying the program as needed. More-specific details on how to structure the PBAS to improve over time are provided in subsequent recommendations.

Designing Performance Measures

The selection of performance measures is of vital importance, as the measures dictate what the service providers should focus on and what they might choose to ignore or neglect. At the same time, selecting performance measures can be complicated by a variety of factors. What

performance data are already available? How much would it cost to collect new performance data? What is known about how different inputs, structures, processes, or outputs relate to the ultimate outcomes of interest? How much control can service providers exert over various measures of interest, and to what extent do external factors come into play? How many different measures can a service provider be expected to focus on? Are the measures of interest resistant to manipulation by the service provider? Are the measures sufficiently transparent? With all that in mind, we offer the following recommendations as ideals for which to strive.

Focus on Performance Measures That Matter. Performance measures will determine how service providers focus their efforts. To the extent possible, therefore, it makes sense to include those measures believed to have the greatest effect on the broader goals of interest.

Create Measures That Treat Service Providers Fairly. In certain settings, service providers' ability to influence desired outputs might be limited. In education, for instance, teachers can offer more-effective instruction, but the quality of a student's home life will also affect performance. Likewise, in medicine, a doctor can follow appropriate care processes, but a patient's failure to heed doctor instructions might still prevent a successful outcome.

In such cases, there are three options for creating performance measures that treat service providers fairly:

- Create "risk-adjusted" output measures that account for relevant social, physical, or demographic characteristics of the population served. As an example, schools with a higher percentage of non-native English speakers might have lower performance targets for reading-test performance.
- Establish measures based on inputs, structure, or processes rather than on outputs. So, for instance, doctors might be judged on the basis of whether they have implemented an electronic data system (i.e., patient registry) to better manage patients with chronic conditions.
- Measure relative improvement rather than absolute performance. With this approach, a teacher might be judged on the improve-

ment (or lack thereof) of student test scores from one year to the next rather than on just the current year's scores.

Avoid Performance Measures That Focus on a Single Absolute Threshold Score. The threshold approach can be intuitively appealing, in the sense that the specified score represents a quality bar that all service providers should strive to achieve. In practice, however, measures that focus on a single threshold can prove quite problematic. Low achievers with no realistic prospects for achieving the absolute threshold score will have no incentive to seek even modest improvements, while high achievers will have no incentive to strive for further improvement.

There are alternatives to the single threshold value:

- Develop multithreshold (e.g., low, medium, high, very high) or continuous (e.g., 0 to 100 percent) scores, and provide different incentives at different points along the spectrum.
- Measure year-over-year improvement, offering rewards for meaningful gains, rather than focusing on single-period scores.

Either of these approaches can help ensure that all providers will be motivated to seek continued improvement.

Designing Incentives

If performance measures dictate what service providers will focus on, it is the type and magnitude of the incentives that governs the level of effort they will expend. When setting up incentives, there are a variety of potential constraints to consider—for instance, the ability to alter the incentive structure for service providers and the availability of resources for creating incentives. With those in mind, we offer the following recommendations for structuring the system of incentives within a PBAS.

Create an Incentive Structure Compatible with the Culture of the Service Activity. The appropriate types of incentives to offer or impose can vary considerably from one sector to another, as well as for different parties within a given sector. Common options include

cash, promotions, status, recognition, increased autonomy, and access to training or other investment resources. The goal is to adopt forms of incentives that will motivate desired behavioral changes on the part of service providers within a given sector while not undermining intrinsic service motivation. Key issues that might influence the appropriate form of incentives include whether the service providers are public, private for-profit, or private nonprofit entities and whether the incentives will be applied to individuals or organizations.

Make the Rewards or Penalties Big Enough to Matter but Not Exceeding the Value of Improved Performance. The size of the incentive should outweigh the effort required by the service provider to improve on the performance measure; otherwise, service providers will simply not make the effort. Ideally, from a cost-effectiveness perspective, the size of the incentive should just barely exceed the cost of making changes in the service-delivery activity. In other words, the goal is to create a set of incentives that induce desired behavioral changes in the most cost-effective manner.

However, if the size of the incentives exceeds the value obtained from improved provider behavior, by definition, the PBAS will not be a cost-effective approach. A good example of this principle is provided by A+B contracting, in which incentives for early completion are tied to monetized estimates of the social value of reduced traffic congestion.

It should be noted, of course, that efforts to estimate the value of service improvements in monetized terms can be fraught with methodological challenges. Even the relatively simple example of A+B contracting involves potentially problematic assumptions regarding the value that travelers place on time savings. Even so, to the extent possible, and recognizing that some subjective judgment might be required, it is important to consider the benefits of improved performance in relation to costs of motivating the improvement.

Implementing a Performance-Based Accountability System

It is possible to create an effective design for a PBAS and then fail to implement the design successfully. Our reviews of PBAS cases in different sectors suggest the following recommendations for implementing a PBAS.

Implement the Program in Stages. Because most PBASs are quite complex, it is often helpful to develop and introduce different components in sequence, modifying as needed in response to any issues that arise. Strategies for a staged implementation process include the following:

- Focus initial efforts and funding on the development of capacity to measure and report performance; introduce performance incentives only after the measurement capacity is in place.
- Start small and gradually expand both the measures and incentives over time.
- Collect data to track the progress of implementation and effects, both positive and negative, and update these data over time.
- Incorporate a pilot-testing phase to test measures and other design features.

Integrate the System with Existing Performance Databases, Accounting Systems, and Personnel Systems. A PBAS is not created in a void; rather, it must be incorporated within existing structures and systems. The main point of this recommendation is to carefully think through all of the ways in which the PBAS will need to interact with preexisting infrastructure—for example, performance databases, accounting systems, and personnel systems—and ensure that this can occur in a seamless manner. In some cases, this might suggest changes in the design of the PBAS; in other cases, it might highlight ways in which the existing infrastructure needs to be modified at the same time the PBAS is being created. It is also important to evaluate whether what is being asked of service providers can in fact be accomplished, given constraints imposed by the existing systems.

Engage Providers and, to the Extent Possible, Secure Their Support. To garner providers' support, it is helpful to develop measures that are credible (i.e., tied to outcomes about which they care), fair (i.e., that account for external circumstances beyond providers' control), and actionable (i.e., that the service provider can positively influence through appropriate actions). A good model for this, from the health-care and PHEP sectors, is to involve providers in the process

of developing the measures and incentives. Though some of the input they provide might be specific to their own context, service providers can often provide extremely valuable input on what is important and what is realistic or feasible.

CAFE standards provide a counterexample to illustrate the problems that might arise when provider support is not secured. When CAFE standards were first introduced in the 1970s, auto manufacturers were understandably resistant to the legislation, and they have remained so in the intervening years. Drawing on their considerable lobbying clout, the large manufacturers fought tenaciously and effectively to block increases in the required mileage standards throughout the late 1980s, 1990s, and early 2000s. As a result, the net effect of CAFE standards in terms of reducing fuel consumption has been much less than what would have been possible had policymakers gained providers' support before implementation.

To some degree, it can be expected that service providers, if afforded the opportunity to influence measures and incentives, might seek to weaken the targets or standards to their benefit. In such cases, those responsible for implementing and overseeing the PBAS will need to judge whether lowering performance expectations would ultimately undermine the success of the PBAS. One possible strategy for overcoming this dilemma, pursued in both health care and PHEP, is to begin with less stringent performance targets with the expectation that the measures and incentive structures will become progressively more demanding over time.

Ensure That Providers and Other Stakeholders Understand Measures and Incentives. Communication is key. Particularly in cases in which there are numerous providers with varying internal support systems to enable engagement—as, for example, with health-care P4P systems and child-care QRISs—it can be helpful to employ multiple communication channels (e.g., email, website, conference presentations) as appropriate. It is also beneficial to keep other stakeholders (e.g., provider professional organizations, referral agencies, business community, consumers) apprised of progress to date and remaining areas for improvement.

Plan for the Likelihood That Certain Measures Will Top Out. As service providers improve their performance in response to the incentive structure, a growing percentage might achieve the highest possible scores for certain measures. PBAS designers should plan for this eventuality:

- One option is to replace a topped-out measure with an even more challenging performance goal so as to induce continued improvement.
- Another alternative is to require that service providers maintain a high level of performance for topped-out measures in order to qualify for incentives based on other measures of interest.

Provide Resources to Support Provider Improvement. In some cases, service providers might be genuinely motivated to improve but lack either the knowledge or the necessary resources to do so. It therefore can be valuable to devote program resources to support efforts at improvement. This might take the form of educating providers with strategies for becoming more effective, or it might involve infrastructure investments. In health care, for example, the federal government has allocated $35 billion to implement information technology in physicians' offices to support improved measurement and practice.

Evaluating and Improving a System Over Time

Ironically, given the spirit of accountability embodied in the PBAS approach, most of the examples that we reviewed in this analysis have not themselves been subjected to rigorous evaluation. In our view, rectifying this lack is of vital importance. Most PBASs are sufficiently complex that any initial design is likely to contain at least some flaws or limitations. It is only through careful monitoring and evaluation that decisionmakers can detect problems and take steps to improve system functioning over time. The following recommendations are intended to foster a more effective monitoring and evaluation process.

Consider Using a Third Party to Evaluate the System. Not all organizations that implement a PBAS possess the necessary methodological expertise to conduct a sound programmatic evaluation.

Additionally, many implementing organizations, for understandable reasons, will tend to be biased in favor of positive results. For these reasons, it is beneficial to rely on an independent and qualified third party to conduct an evaluation of the PBAS.

Structure the Evaluation of a System Based on Its Stage of Development. When a system is first developed, it might be most helpful to evaluate implementation activities. For example, have the appropriate mechanisms for capturing and reporting performance measures been developed? As the system matures, the focus of the evaluation should shift to evaluating the effects, in terms of observed provider behavior and service outputs, of the performance measures and incentive structure. Note that, if a system is likely to continue to change over time, it is appropriate to focus more on cross-sectional comparisons (comparing the behavior and outputs of different sets of providers at the same time) than longitudinal comparisons (comparing the behavior and outputs of the same set of providers at different points in time).

An evaluability assessment is a good way to start the process (Wholey, Hatry, and Newcomer, 2004). It can help to match different forms of evaluation to the specific features of a particular PBAS. The evaluation should take place only after performance measures and incentives have been in place long enough to influence behavior. With CAFE standards, for example, the required timeframe for designing and manufacturing a more efficient vehicle fleet is at least several years. Thus, although the program was enacted in 1975, manufacturers were not held accountable until the model years 1978 (for passenger cars) and 1979 (for trucks).

Examine the System's Effects on Both Procedures and Outputs. The evaluation should consider the PBAS's effects on both procedures and outputs. A logic model might be used to illustrate the ways in which the PBAS is intended to influence provider behavior. The logical connections among the system's elements, as described by the model, become testable hypotheses that can be used to help structure the evaluation. Inferring from the logic model, one might ask, for example, Does the magnitude of the incentives motivate behavioral change on the part of service providers? Do the measures and incentives in combi-

nation induce the intended behavioral response? Do service providers' actions translate to improved outputs as envisioned?

Use the Strongest Possible Research Design Given the Context in Which the System Exists. Options, sorted in order of decreasing rigor, include RCT, regression discontinuity design, regression analysis with instrumental variables, propensity-score matching, nonequivalent-group design, lagged implementation design, and case studies. Appendix B discusses these alternative research designs and the circumstances in which they might be applied.

If certain design aspects are flexible, it might be possible to implement variations in the PBAS coupled with common evaluation frameworks to provide rigorous comparison and help choose the most effective options. Such variations could include different performance measures, different types of incentives, or different incentive levels (e.g., significant versus modest financial rewards).

Implement Additional, Nonincentivized Measures to Verify Improvement and Test for Unintended Consequences. As noted earlier, a PBAS might induce service-provider responses that lead to improved performance scores without corresponding improvement in the underlying objectives. Consider education, for example: A teacher might choose to invest additional instructional effort on test-taking strategies or on specific questions that he or she believes are likely to appear on the test (as opposed to additional instruction on the subject matter itself). This could lead to improvement on standardized test scores that overstates actual student gains in mastery of the broader subject matter. To detect when this might be occurring, it can be helpful to include nonincentivized measures intended to test similar concepts. Students might, for instance, be administered additional math and reading exams in alternative test formats to check whether there has been a comparable level of improvement. If scores on the nonincentivized measures fall short of scores on the incentivized measures, there is reason to question whether the incentivized measures reflect actual improvement gains or simply the results of teaching to the test.

It is also the case that, as service providers focus more effort on incentivized measures within a PBAS, other aspects of service might suffer. Again using education as an example, additional instructional

effort devoted to math and reading might limit the attention devoted to other subjects, such as arts, science, and social studies. If evidence suggests that the current structure of a PBAS's measures and incentives is leading to unintended consequences that are viewed as unacceptable, then the structure of the PBAS can be modified accordingly.

Link the System Evaluation to a Review and Redesign Process. The true benefits of evaluation come not from simply understanding what is working and what is not but rather from applying that understanding to improve the functioning of the PBAS. Evaluation should thus be embedded within a broader framework for monitoring and continuing to refine the PBAS over time.

Areas for Further Research

Because so few of the PBASs that we examined have been subjected to rigorous testing and evaluation, there are a number of fundamental questions about PBAS design, implementation, and performance that our study is unable to answer. Further research to gain greater insight on some of these questions would aid understanding of the circumstances under which a PBAS represents a good policy option and the design and implementation approaches that can make a PBAS as effective as possible. In fact, a good starting point might be to conduct a nationwide survey of existing PBASs in diverse policy areas. Key questions for further research—within individual sectors, as well as across sectors—include the following:

- Do PBASs generally improve key outputs and outcomes?
- What are the links between the policy outcomes that interest PBAS creators and the processes or outputs that they must monitor to induce meaningful and productive change?
- What PBAS elements are the most important in terms of effectiveness?
- Are there certain types of unintended consequences (e.g., decreased performance for nonincentivized aspects of service) that routinely

occur with PBASs, and are there effective strategies for preventing such consequences?
- How much does it cost to create and run a PBAS—for the implementing agency, for the service providers, and for other stakeholders?
- Are PBASs generally cost-effective?
- How do PBASs compare to other policy approaches in terms of cost-effectiveness?
- How can we encourage a greater level of programmatic evaluation within PBASs?

Final Thoughts

The concept of measuring and rewarding performance as a means of improving service delivery appears to be gaining considerable traction in policy circles. The five sectors that we examined in this study—child care, education, health care, PHEP, and transportation—yielded numerous examples of operational PBASs. Yet, implementation has, to date, far outpaced efforts to evaluate the merits of this approach in real-world settings. It is thus difficult to answer such seemingly basic questions as whether PBASs are generally effective in achieving their aims and whether they do so in a cost-effective manner.

Even so, the analysis presented in this monograph suggests that PBASs represent a promising policy option for improving the quality of service-delivery activities in many contexts. In several of the cases that we examined, most notably A+B contracting, the results of the PBAS approach appear unambiguously positive. In other examples, in which the results to date have been more modest, there are reasons to suspect that the effectiveness of the approach could be considerably strengthened by simple refinements to the performance measures or incentive structures (e.g., increasing the magnitude of the incentives). In short, though we still have much to learn about the design, cost, and effectiveness of PBASs, the research and analysis reviewed in this monograph support continued experimentation with and adoption of PBASs in appropriate circumstances.

This recommendation comes with two qualifications. The first is that the appropriate design for a PBAS and, ultimately, its prospects for success are highly dependent on the context in which it will operate. Is the responsibility for service delivery concentrated within a single organization or dispersed across multiple levels of government? Are the services publicly or privately provided? What degree of influence do providers have on the outputs of their service? Does the responsibility for improving service logically rest with individuals or organizations? Are the mechanisms through which changes in service-provider behavior can lead to improved outputs or outcomes well understood? Will existing regulatory or governance structures undermine or support the intent of a PBAS? As discussed throughout this monograph, decisionmakers considering, designing, or implementing a PBAS should devote careful attention to such questions and make their choices accordingly.

Second, ongoing system evaluation and monitoring should be viewed, to a far greater extent than in prior efforts, as an integral component of the PBAS. That is, PBASs should, by design, be structured as learning systems. Because PBASs are typically complex, getting all of the details right for the initial implementation is rare. Evaluation and monitoring provide the necessary information to refine and improve the functioning of the system over time. Additionally, morethorough evaluation and monitoring of PBASs will lead, gradually, to a richer evidence base that should help future decisionmakers understand (1) the circumstances under which a PBAS would be an effective and cost-effective policy instrument, and (2) the most appropriate design features to employ when developing a PBAS for a given set of circumstances.

APPENDIX A

The Five Sectors

This appendix provides a brief description of each of the sectors and the relevant PBASs covered in this monograph. The descriptions are based on our knowledge and research expertise.

Child Care

Child care is funded and delivered by public agencies at all levels of government and a variety of private organizations as part of what has been described as a "non-system of micro-enterprises" (Kagan, 2008). Formal child-care programs operate in a range of settings, including free-standing centers, public school campuses, churches, community centers, and family homes. Program models include full-day care for ages 0–5, pre-K programs for four-year-olds (and, in some states, three-year-olds), and part-day preschool programs. Programs might be funded entirely by parent fees or federal monies, receive subsidies for children whose families qualify, receive in-kind subsidies from the churches or other organizations that sponsor and house them, or rely mainly on parent volunteers as part of child-care cooperatives. Child care is an imperfect market in several respects: (1) Programs are generally underfunded because most parents cannot afford to pay the full cost of care and public subsidies are set at less than full cost; (2) the supply of affordable care in most areas is limited; and (3) most parents are grateful to find an affordable place for their child that accommodates their work hours.

Although studies have consistently found that average child-care quality is mediocre, the sector has not focused much attention on program quality. Until quite recently, quality standards were largely defined by state licensing requirements, which represented a fairly low bar. Licensing is focused primarily on the adequacy and safety of a program's physical environment, including fencing, square footage, and protection of children's well-being (i.e., are electrical plugs covered? Are cleaning supplies locked up?).

The growing policy attention on K–12 accountability has raised questions about child-care outcomes, particularly school readiness. These questions have led the sector to focus on quality and devise ways to improve it. QRISs represent the most popular current approach to doing so. QRISs produce a single, easy-to-understand rating for each provider, much like restaurant ratings in some cities; QRISs differ from other PBASs in that participation is voluntary. QRISs define quality standards and measure and rate providers, thus making program quality transparent, and provide incentives and supports for quality improvement. Ideally, these systems promote awareness of quality and encourage programs to engage in a process of continuous quality improvement. While QRISs ultimately are expected to promote improved child outcomes, such as increased school readiness, QRISs focus primarily on assessing and improving program inputs and processes. States have found QRISs an attractive approach to improving child-care quality; 19 states now operate QRISs, and several others are developing them.

The rating process and the ratings that result are the major QRIS monitoring activities. Rating systems essentially define child-care quality by identifying which program components will be assessed. States generally measure child-staff ratios, group size, staff education and training, and some indicator of the classroom environment. States differ in whether to include and how to weight parent involvement, national accreditation, and management processes. A number of issues surround QRIS ratings. A key one is integrity: Most measures were designed for low-stakes use, but QRISs are high-stakes systems; summary ratings might affect both program funding and enrollments. Ratings are also quite costly; they typically require hours-long classroom

observations. How the different component measures are combined into a single program rating has received no empirical attention.

Early on, one of the key incentives was the prestige associated with a high rating. But significant funds are required to support key quality improvements, such as reduced child-staff ratios and improved staff education and training. As a result, most QRISs now provide financial incentives to support improvements and motivate providers to participate in rating systems. Incentives are generally linked to the summary rating: Higher-rated programs receive more funds to support their higher-quality programs. In higher-stakes QRISs, rating might also affect the level of funding provided for subsidy-eligible children. Many states also provide staff-level incentives, including scholarships or other professional development programs; eligibility generally requires a program rating that denotes at least reasonable quality.

Incentives might also occur in the form of hands-on quality-improvement support. Often, this support begins with detailed feedback on rating results and a specific quality-improvement plan. In many systems, coaches provide specific technical assistance. This package of supports can be very motivating for providers, who often do not know how best to spend the limited quality-improvement funds they receive through their participation in the QRIS process or how to initiate quality-improvement efforts.

Research on QRISs has been limited in both focus and depth. Most efforts focus on testing the validity of these systems and ask basic questions appropriate to this task: Do summary ratings relate to other measures of quality? Are the quality-improvement efforts resulting in improvements in participating-provider quality? A large share of the evaluation studies has focused on examining correlations between environmental ratings and overall program ratings. Typically, moderate correlations are found. Several studies have examined whether average ratings improve over time; generally, they do. However, we do not know how well QRISs measure what they purport to measure or whether children benefit from the improved care they receive as their providers receive quality-improvement support. Many of the measures used to assess the components were developed in low-stakes settings, such as research studies or self-assessments, in which there were few, if

any, consequences attached to a particular score. These measures might not be appropriate in high-stakes settings, in which summary ratings could substantially affect a program's bottom line.

Education

Public education in the United States is primarily the responsibility of state and local governments. Traditionally, states reserve for themselves the functions of school accreditation, teacher certification, curriculum adoption, and financial auditing, and states delegate to local districts the responsibility for operations, instructional materials, and staff supervision (although there is considerable variation in this pattern). Districts, in turn, delegate many operational decisions to individual schools. The educational governance system has been described as "loosely coupled" because responsibility is distributed across levels without rigid monitoring and accountability (Weick, 1976). In most districts, teacher and principal salaries are determined by a negotiated schedule that rewards postsecondary education and job experience but not individual performance.

Until recently, the federal role in public K–12 education has been limited to regulations and supplemental funding designed to promote equity for economically disadvantaged students and students with disabilities. The federal government contributes about 10 percent of the total cost of K–12 education.

In the 1980s and 1990s, some states began to adopt standards-based reforms (SBRs), which were designed to shift the focus of governance from inputs (finance, accreditation, certification) to outputs (student achievement) and to align the elements of the educational system to foster higher achievement. In 2001, this idea was incorporated into federal legislation (NCLB), which is essentially a PBAS that uses schools and districts as the units of accountability. NCLB requires that all states create accountability systems that include state standards in reading, mathematics, and science and annual testing of students in selected grades in these subjects.

A second form of PBAS (P4P) is being adopted in some districts and states. P4P systems usually pay bonuses directly to teachers or principals for meeting specific performance criteria, usually in terms of student achievement but, in some cases, including other outputs relevant to students (such as graduation) or educators' practices.

Under NCLB, each state must identify students who are proficient based on reading and mathematics tests. The percentage of students who are proficient in reading and mathematics in each school is compared to a target value, which must increase to 100 percent by 2014. The calculation must be made for the school as a whole and for each significant subgroup of students, including the major racial and ethnic groups, students of low socioeconomic status, English-language learners, and special education students.

Most P4P programs also use the state tests as their primary measure and compute some form of value-added metric to try to determine how much growth is associated with a particular teacher or principal each year.

NCLB includes a graduated set of interventions that are intended to motivate better performance and effect specific changes while also providing schools with needed assistance. P4P systems tend to focus more on rewards than on sanctions, offering cash bonuses (or sometimes salary increases) to teachers and principals whose students meet the performance targets. Some of these systems are competitions, with the highest-performing teachers receiving bonuses, whereas others set fixed growth targets and pay bonuses to any teacher or principal reaching the target.

A large body of research indicates that high-stakes testing has a strong effect on teaching practice (see, e.g., Stecher, 2002). Teachers tend to align their lessons with both standards and assessments, which often leads to a reduction in time spent on topics and subjects that are not tested. This focus on tested material can lead to *score inflation*, which refers to gains on a test that do not generalize to other measures of the same topic or subject. Some research suggests that, when performance is measured according to a single threshold, such as proficiency, teachers tend to focus on students near that threshold (often called bubble kids). Furthermore, there is some evidence that high stakes lead

to excessive test preparation (i.e., practice with specific formats, such as multiple choice) and even cheating.

The literature on P4P is also mixed; some programs have been associated with improvements in achievement, but it is not always possible to distinguish real gains from score inflation. In addition, there is some evidence that teachers and administrators have trouble understanding the information on performance gains reported as value-added measures and that, in some systems, teachers believe that the P4P program has led to negative effects in their schools.

The impact of NCLB and other SBR policies on achievement is uncertain; while scores have increased in many places, it is difficult to know whether these are real gains or stem from score inflation. NCLB appears to have had beneficial effects of focusing attention on student outcomes and highlighting the performance of traditionally low-performing subgroups of students. However, there are no studies of the overall costs and benefits of NCLB or P4P in education. Although only a small percentage of educational spending goes for NCLB accountability provisions, many states have reacted negatively to the accountability requirements and characterized them as unfunded mandates. Some states lack capacity to create and manage high-quality, test-based accountability programs and address the needs of identified schools.

Health Care

Unlike most of the other sectors represented here, health care in the United States is provided primarily by the private sector. Only a quarter of the population is covered by public health insurance (i.e., the elderly and low-income, by Medicare and Medicaid, respectively). In contrast, 60 percent are covered by private health insurance plans. (Roughly 15 percent have no health insurance coverage and typically rely on emergency services.) Most private plans are employer-sponsored, managed care plans. That is, care is based around a network of preferred providers offering lower-cost and more-comprehensive benefits than out-of-network providers. (In the case of HMOs, out-of-network care is restricted altogether.)

Performance measurement in health care was initially used internally for quality improvement, but, with the 1990s expansion of managed care, it grew into a mechanism to hold plans accountable. NCQA provides accreditation of health plans and has developed HEDIS to provide plan-quality information to employers. In addition, health plans, employers, consumer advocacy groups, and various government agencies publish quality report cards to assist individuals in choosing their health-care providers (at the level of health plans, medical groups and hospitals or individual physicians). Regardless of these efforts, a 2003 RAND study (McGlynn et al., 2003) found that adults in the United States receive only 50 percent of recommended care on average. Currently, most health-care providers are reimbursed regardless of how well they provide care or how efficiently they use resources.

Despite this emphasis on the private sector, CMS remains the dominant purchaser in the market for health-care services because use is much greater in the senior population. (Even though seniors represent only a fraction of the total market, they represent a far greater share of overall health spending.) As a result, any reform to physician or hospital reimbursement under Medicare will have repercussions for the rest of the health sector. While performance-based accountability was initiated in the private sector, P4R legislation in 2003 and 2006[1] has set the stage for performance-based accountability to move into the public sector through Medicare.

At last count, there were more than 40 hospitals and more than 100 physician and medical-group performance-based accountability (P4P) programs in place in the private sector in the United States. Additionally, CMS is staging a number of P4P demonstrations targeted at hospitals, physician group practices, end-stage renal-disease facilities, nursing homes, and home health workers. Outside the United States, in 2004, the UK's National Health Service rolled out a large-scale P4P program for general practitioners. No single approach to P4P is being used, and there is a wide variation in program designs.

---

[1] P4R legislation was part of the Medicare Prescription Drug Improvement and Modernization Act of 2003 (Pub. L. 108-173), which established the Reporting Hospital Quality Data for Annual Payment Update program and the Physician Quality Reporting Initiative.

The number and set of measures used in the PBAS programs vary widely from only a few (typically no more than five to 10 in the small-scale U.S. programs) to many (e.g., 146 in the case of the UK). The measures are centered primarily on quality, although, recently, more programs have incorporated measures of efficiency. Many of the quality measures are process measures—that is, they evaluate actions taken by providers, and they tend to measure the proportion of patients in a certain risk group who received some specific type of evidence-based care (e.g., proportion of women ages 52–69 who got a mammogram in the past two years). The measures can be computed from such sources as administrative data, electronic health records, and medical charts. Cost depends on data infrastructure, and auditing is often in place to avoid gaming and cherry-picking patients. The measures might or might not be made available to patients, as well, in which case the information must be useful and understandable to patients.

Similarly, the reward structures of PBAS programs in health care vary widely, from the target of the incentives (physician, medical groups, or hospitals) to the amount of money at risk (varying from $500 to $5,000 per doctor for most programs in the United States to almost $40,000 in the UK program). Such programs are usually structured around meeting specific performance thresholds, but these thresholds can be absolute or relative to other providers. Some programs also pay for improvement. Paying for improvement motivates low performers to improve rather than simply rewarding high performers. The source of the incentive money can be existing funds (so the program is budget-neutral, in which case poor performers might be penalized) or new money (paying out bonuses in addition to existing reimbursement to the successful performers). Since variation in the measures often has some random component (or, in some cases, depends in part on patient behavior), issues of fairness often arise, and some physicians are unwilling to participate in such programs because of the risk. The frequency of the incentive payments (e.g., quarterly, annually) also might have an impact on the effectiveness of PBAS programs, as more-frequent payments might be more salient to providers and thus motivate more-persistent gains in performance.

Despite the popularity of PBAS programs in health care, only a handful of studies have rigorously evaluated their impact. Since many programs in the early stages of PBASs were small in scale, many studies have found only marginal impacts if they found any impact at all. Due to data limitations, these studies have tended to focus on changes in rewarded performance measures only, as opposed to unrewarded measures, which might decrease if providers respond to PBAS incentives by multitasking or teaching to the test (as they do in education). So far, no study has shown PBASs to result in a notable disruption in care. While it is known that design features matter, existing studies do not provide information on the impact of various design features (e.g., number of measures, payment structure, target of the incentives) in any intervention's success or failure.

Public Health Emergency Preparedness

PHEP involves efforts to prevent, protect against, quickly respond to, and recover from large-scale health emergencies, including bioterrorism, naturally occurring disease outbreaks, and natural disasters. Primary legal responsibility for PHEP (as with other aspects of public health) lies with state health departments, which delegate varying degrees of responsibility to local health departments. While PHEP efforts are typically coordinated and led by governmental public health agencies, PHEP also requires active involvement by health-care systems, emergency management, law enforcement, communities, and individuals.

Until recently, the federal role in PHEP was limited largely to providing assistance and coordination during large-scale incidents that stretched state and local capabilities. During the late 1990s, increasing concern about the threat of weapons of mass destruction led to a small federal effort to build state and local ability to prepare for large-scale public health emergencies. That effort grew considerably after September 11, 2001, and the anthrax attacks of October 2001. A survey by the National Association of County and City Health Officials (NACCHO) (2007) estimates that 41 percent of local health departments receive all

of their PHEP funding from federal sources, while another 40 percent get more than three-fourths of their funding from federal sources.

The two most important federal PHEP programs focus on hospital preparedness (the HHS Hospital Preparedness Program) and all-hazards public health preparedness (CDC PHEP cooperative agreement). Although the 2002 National Strategy for Homeland Security (Office of Homeland Security, 2002) required that these and other programs create performance measures to evaluate progress, until recently, there were no clearly defined consequences associated with them. PAHPA clarified those consequences by requiring that HHS, as of 2009, withhold federal funding for failure to meet performance benchmarks. The remainder of this summary focuses on the PHEP cooperative agreement.

The PHEP cooperative agreement requires grantees (including the 50 states, four separately funded large cities, and eight territories) to report data on performance metrics for two program areas: (1) mass medical countermeasure delivery and (2) all other aspects of PHEP. Early metrics focused on infrastructure (e.g., plans, personnel, and equipment), but, more recently, the cooperative agreement has utilized an increasing number of drill-based metrics for assessing operational capabilities (i.e., whether grantees can use infrastructure to complete operational tasks). For instance, a performance metric for the 2009 grant year assesses the amount of time required to notify key incident management staff of the need to report for duty.

Currently, countermeasure delivery is assessed through an extensive written assessment plus five drill-based metrics (grantees must report on three). There are 14 metrics for the remainder of PHEP, focusing on both infrastructure and operational capabilities. Some, but not all, of these metrics are not associated with clear consequences. With the exception of the written assessment on countermeasure delivery (which is administered during site visits by CDC staff), data collection relies on grantee self-reports. While federal data-reporting requirements have applied only to the (mostly state-level) grantees, many states have chosen to require their local grantees to provide data on state-level performance measures before releasing the funds to the local level.

Under the PAHPA legislation, in 2010, CDC must begin withholding funds for failure to meet performance benchmarks. Initially, these benchmarks are linked to a subset of measures that focus on infrastructure and completion of (but not performance on) operational assessments, but, in future years, it is expected that funding will be tied to *levels* of operational performance. States or other grantees failing to meet a benchmark for one year lose 10 percent of their funding. The penalty increases to 15 percent for failure during two consecutive years, 20 percent for three consecutive years, and 25 percent for four consecutive years. HHS might reduce or waive penalties for mitigating conditions, and funds withheld are allocated to hospital preparedness activities within the same jurisdiction. (It should be noted that total budget for the cooperative agreement decreased more than 25 percent in real terms between 2007 and 2009, from $767 million to $609 million).

There are also numerous anecdotes about poor performance ratings on mass medical countermeasure delivery being used by state and local policymakers to justify replacement of key PHEP personnel, thus adding another potential set of consequences associated with the measures.

It is too early to assess the impact of performance-based accountability on PHEP. Nonetheless, there are numerous and widespread anecdotes suggesting that the relatively strong emphasis on performance measurement for mass medical countermeasure delivery has led state and local health departments to invest in those capabilities at the expense of other capabilities. Moreover, the NACCHO survey noted earlier suggests that the threat of funding cuts (on top of those sustained during recent years) are leading local health departments to scale back on preparedness activities.

Transportation

The surface transportation (highway and transit) sector consists of public agencies at all levels of government along with a wide range of private actors. Federal, state, and local governments, in varying degrees, are largely responsible for activities, such as setting policies, raising and

distributing revenue, planning and developing projects, and maintaining and operating existing infrastructure. The private sector manufactures personal, commercial, and transit vehicles, and private firms might also contract for such activities as building or maintaining highways or operating transit services.

Performance measurement has a long tradition in the transportation sector, and literally thousands of performance metrics have been developed or discussed. Most commonly, however, performance metrics are used to inform policy and planning decisions. Examples in which performance measures are used to enforce greater accountability are the exception rather than the rule.

PBASs are of growing interest in the field of transportation planning and policy. Federal legislation is expected to be enacted in 2010 that will renew the national transportation-funding program for the following six years. Many politicians, interest groups, and scholars are advocating that familiar formulas by which federal funds are distributed to states for particular programs be replaced by funding arrangements that are "performance-based" (National Transportation Policy Project, 2009). Thus far, many state and federal programs purport to measure and report on the performance of the transportation system, but relatively few include any sort of accountability requirements. Current debates suggest that funding should, in the future, be tied more directly to measures of the attainment of major programmatic objectives, such as improved mobility, increased accessibility, and congestion reduction, yet it is not clear that consensus can be reached on approaches by which to measure the attainment of these objectives. For this study, we were unable to find transportation programs that incorporated PBASs, but several specific PBASs were identified within or related to transportation. These were, in general, narrower in scope than some of the pending proposals. For example, in this study, we included examination of road construction contracts that provide bonuses for early completion of a road project and penalties for late project delivery. We also looked at penalties imposed on regions under the CAA amendments when their regional transportation plans failed to result in specified targets for the reduction of air pollutant emissions. Also included was the CAFE program of the federal govern-

ment, which financially penalizes automobile manufacturers that fail to achieve improvements in fuel economy in pursuit of environmental goals. A fourth transportation-related PBAS is an attempt to reward financially public transit systems that increase their daily patronage in relation to other transit systems. While these four examples cannot, even when taken together, suggest how a more integrated PBAS might work in the field of transportation, they provide many lessons that should influence the design of such a system over the coming few years.

A+B Highway Construction Contracting Overview

State and local governments often contract with private firms for road construction activities. Traditionally, contracts have been awarded to the firm offering the lowest bid. Some states have adopted performance-based contracting, which is referred to by contractors as A+B contracting because the incentives are enumerated in a section of the contract (part B) that follows the basic contracting language. Under such an incentive-based contract, both the financial cost and the time to complete the project are included in the contract; part A specifies the financial cost, while part B provides rewards for early completion and penalties for late completion. This innovation was motivated by the fact that construction activities create or exacerbate traffic delays. Speeding up project delivery will therefore reduce public costs—in terms of wasted time and fuel—associated with road construction. A+B contracting represents a shift in emphasis from lowest agency cost to best overall public value. Because most highway construction is funded by states (sometimes using federal funds), the gradual shift to A+B contracting has required enabling federal and state policy frameworks.

The principal metric is the time required to complete construction. Contractors are also held to various design and engineering standards, but such standards are not unique to A+B contracting.

A+B contracting relies on financial rewards or sanctions to the construction firm to encourage faster delivery, though the specific form of the incentive can vary depending on how the contract is implemented. In some cases, both time and cost are specified in the bid, but the award price is reduced if the contractor is late in delivering the project. In other cases, only the cost is specified within the bid, but the

contractor receives additional bonuses if the project is delivered earlier than the target delivery date. In either case, the size of the bonus or penalty is a function of the number of days that the project is ahead of or behind schedule.

A+B contracting has proven quite successful in reducing the time to complete projects when compared with traditional lowest-bid contracting, and there are many examples in which total construction time has been reduced by more than 50 percent. Provided that the daily bonuses or penalties for early or late delivery are commensurate with the costs, in wasted time and fuel, of construction-related traffic congestion delays, A+B contracting appears to be an effective strategy for minimizing the net social cost of highway construction activities.

Clean Air Act Conformity Requirement Overview

Under the CAA Extension of 1970 and subsequent CAA Amendments of 1977 and 1990, EPA sets ambient air quality standards for several criteria pollutants, including carbon monoxide, nitrogen dioxide, ground-level ozone, sulfur dioxide, fine particulate matter, and lead. Metropolitan regions failing to meet one or more of the EPA criteria-pollutant standards are designated as nonattainment areas. States with nonattainment areas must develop a SIP demonstrating how they will achieve compliance with EPA standards by a certain date. Because automobiles and trucks—described as mobile sources—represent a major source of certain pollutants, SIPs often include strategies for reducing mobile-source emissions within each nonattainment area. To strengthen this link, the CAA Amendments of 1990 specified that federal transportation funds be withheld from nonattainment areas that adopt transportation investment plans likely to prevent air quality compliance within the required timeframe.

The initial determination of compliance with EPA's ambient air quality standards is based on sampling the average concentration of the regulated pollutants at different locations over different time intervals. Once an area is found to be in nonattainment, the emphasis shifts to an exercise in forecasting whether the strategies outlined in the SIP, including mobile-source emission-reduction measures, will be sufficient to achieve compliance by the specified target date. From

the transportation perspective, a key requirement is to ensure that the planned investments specified in regional transportation plans, such as highway investments, do not undermine the mobile-source emission-reduction targets specified in the SIP. In addition to linking federal transportation funding with air quality compliance efforts, the 1990 CAA Amendments also required that the transportation and emission models used by nonattainment areas be more accurate than in previous years, including such features as travel projections by time of day, congestion levels, vehicle speeds, and the interaction of land use and accessibility with travel demand. In effect, the more-accurate modeling requirements make it more difficult to demonstrate compliance, but the results can also be viewed with greater confidence.

The main incentive for complying with EPA's air quality regulations, or, in the case of nonattainment areas, making progress toward compliance, is access to federal transportation dollars. Because large urban regions might receive hundred of millions of dollars per year in federal funding, this is a powerful incentive. Since 1997, conformity lapses (failure to demonstrate progress toward compliance) have occurred in at least 63 areas across 29 states. Most of these areas have returned to conformity quickly and received deferred federal funding without major effects on their transportation program. Five areas, however, had to make significant changes to their transportation plans in order to resolve a conformity lapse. In the most extreme example, the Atlanta region had to strip out the majority of its planned highway expansion projects in order to qualify for about $700 million in federal transportation support.

The success of tying federal transportation funding to compliance with EPA's ambient air quality standards can be viewed as mixed. While air quality has generally improved in the past several decades, there are still many nonattainment areas across the United States. Based on EPA (2010) data, in 1992, when federal transportation funding was first linked to conformity, 199 metropolitan areas (out of 340 areas in total), representing a combined population of close to 100 million residents, were classified as nonattainment for one or more criteria pollutants. By 2003, the number had dropped to 100 areas with a total population of 34 million. In 2004 and 2005, after new

EPA standards related to ozone and fine particulate matter came into force, the number of nonattainment areas jumped to 201, representing around 190 million residents. By 2008, however, the number of nonattainment areas had declined to about 140, with a combined population of 178 million. In short, the past two decades have witnessed slow but steady improvement toward attainment of air quality goals; while the number of nonattainment areas increased in 2004 and 2005, this was due to the application of two additional demanding standards.

Several additional factors complicate the assessment. First, many of the recent improvements in air quality can be attributed to innovations in vehicle emission control technology rather than transportation infrastructure planning and investment. At the same time, mobile sources are not solely responsible for air quality problems; stationary sources (e.g., factories or refineries) also emit harmful pollutants, so failure to achieve compliance is not just a function of transportation planning. What we can say for certain is that the link to federal funding has been sufficient to induce some regions, such as Atlanta, to cancel infrastructure investments likely to further exacerbate air quality challenges.

Corporate Average Fuel Economy Standards Overview

Introduced with EPCA in response to the oil shocks of 1973–1974, CAFE standards require that automobile manufacturers achieve a minimum level of fuel economy for the fleet of vehicles sold each year in the United States. Manufacturers failing to meet the standards are subject to significant fines. Separate CAFE standards are applied to passenger cars and light trucks (e.g., minivans, SUVs, and pickup trucks), with the former being more stringent. Passenger-car standards were first enforced in 1978, and light-truck standards in 1979. During the 1980s and early 1990s, the standards were made more demanding with some regularity, but, in recent years, the standards have been allowed to stagnate. The passenger-car standard of 27.5 miles per gallon (mpg), for example, has remained constant since 1990, while the light-truck standard of 20.7 mpg has not been increased since 1996. With growing concern over such issues as climate change, energy security, and fuel price volatility, however, there has been a renewed interest in more-

stringent fuel economy standards. In response to increased public pressure, Congress passed the Energy Independence and Security Act of 2007 (Pub. L. 110-140), which requires that auto manufacturers once again begin to increase the average fuel economy of their fleets. The new regulations take effect in model year 2011 and will culminate in a fleetwide average of 35 mpg by 2020.

Under CAFE regulations, EPA is responsible for rating the fuel economy for each vehicle model that a manufacturer produces, using a standardized test procedure on new vehicles taken at random from assembly lines. Manufacturers are then judged according to the sales-weighted fuel economy of the fleet of vehicles they sell in the United States each year, based on a set of assumptions regarding typical driving patterns. Fleet fuel-economy measures are calculated separately for passenger cars and light trucks.

If the average fuel economy of a manufacturer's passenger-car or light-truck fleet fails to meet the corresponding CAFE standard, the manufacturer must pay a penalty of $5.50 for each 0.1-mpg shortfall multiplied by the number of vehicles produced for the U.S. market. Manufacturers earn CAFE credits in years when they exceed the standard, which can be applied to offset shortfalls in the preceding or following three years. The rationale for such credits is to ensure that manufacturers are penalized only for persistent failure to meet requirements, not for temporary noncompliance due to anomalous market conditions in any specific year. The threat of sanctions appears to be generally effective, as most manufacturers, including the major U.S. and Japanese firms, consistently meet CAFE standards. Some high-end producers, however, such as BMW, Daimler, Ferrari, Porsche, and Maserati, choose to pay the fines rather than trying to meet the CAFE requirements.

CAFE standards are generally recognized as having had a positive effect on reducing motor fuel consumption in the United States. That said, the CAFE approach to fuel economy remains controversial for several reasons:

- A common strategy for improving vehicle fuel economy is to reduce vehicle weight. Safety proponents have argued that lighter

cars increase the risk of crash fatalities, though other studies suggest that vehicle design and quality are stronger determinants of safety than vehicle weight.
- Economists have suggested that simply taxing fuel consumption, though perhaps more politically challenging, would be a far more-efficient approach for stimulating the production and consumption of more fuel-efficient vehicles.
- The decoupling of light-truck and passenger-vehicle mileage standards, when combined with significant growth in the market share for light trucks in the past several decades, has undermined the fleetwide improvement in fuel economy that might otherwise have been achieved.

Transit Subsidy Formula Allocation Overview

Congress established the National Transit Database (NTD) as the primary source for information and statistics on U.S. transit systems. The database was intended to collect and publish data that individual transit systems could use for service planning and that federal, state, and local governments could use for transit investment decisionmaking. More than 660 transit providers in urbanized areas currently report to the NTD through a standardized reporting system. Each year, NTD performance data are used to apportion more than $5 billion of FTA funds to transit agencies in urbanized areas, mostly for *capital* projects. State and regional transportation agencies, however, support more than 95 percent of transit *operating* costs, and the uses of NTD data for subsidy allocation vary widely from state to state.

An extensive body of research dating back to the late 1970s evaluates and establishes measures of transit service effectiveness, efficiency, and productivity. These studies examine the appropriateness and uses of various performance measures (such as cost efficiency, cost-effectiveness, service utilization, vehicle utilization, quality of service, labor productivity, and coverage) and performance indicators (e.g., cost per mile, cost per passenger trip, passenger trips per vehicle mile, miles per vehicle, average speed, passenger trips per employee, vehicle miles per capita). These studies also examine the use of nontraditional indicators that are not collected by the NTD. Most research finds general

consensus among transit agencies and experts that funding and investment decisions should incorporate a combination of performance indicators when comparing peer groups and that such indicators should be consistent with transit agencies' goals. However, there are strong disagreements about which measures to use and how to combine them into a composite index.

Most states allocate operating subsidies based on a formula process; some states include performance-based measures in these formulas, while the majority use only non–performance-based measures (such as level of local financial match, or service population). Very few states, if any, use wholly performance-based measures and procedures. Studies have identified three general trends in the use of performance measures for transit funding allocation: When performance measures are used, they are (1) combined with nonperformance measures in a composite index, (2) used to determine an incentive level of funding above a baseline set by non–performance-based measures, or (3) eventually eliminated completely from any formula allocation procedures.

For example, following a six-year implementation process, transit operators in Indiana now are categorized into four peer groups based on scale and scope of services and agency size. Funding is then allocated within each of the four peer groups based on a formula that provides equal weight to passengers per operating expense, miles per operating expense, and locally derived income per operating expense. The data used are calculated on a three-year rolling average to enhance funding stability and predictability. Other states that currently use performance measures in transit funding include Florida, Iowa, Ohio, New York, North Carolina, Pennsylvania, Missouri, and California. Texas created a formula based on demographic and performance data in 1989 but abandoned it in 1994; it now allocates based on financial need.

There has been little evaluation of whether the use of performance measures in transit funding allocation has resulted in service-delivery improvements. However, some studies have examined reasons that state and regional governments have followed the general trend of separating performance measurement from funding. Findings suggest that (1) formulas have not produced revenue or funding changes significant enough to affect service-delivery behaviors, especially when perfor-

mance measures are used to determine an incentive level, rather than a base level of funding; (2) philosophical debates arise about whether to penalize agencies most in need; (3) the political process of funding allocation tends to favor distributional equity over operational goals; (4) transit satisfies a broad-based set of goals, many of which cannot be captured in performance measures; (5) lag time between reporting and allocation decisions make the PBAS difficult to administer; and (6) the zero-sum nature of limited transit funding is not likely to garner support for a PBAS that rewards some operators at the expense of others.

APPENDIX B

Designs for Evaluation

As discussed in Chapter Seven, evaluation designs might focus on a broad range of policy questions, might vary in complexity, and might impose more or less burden on those evaluated. Key in designing an evaluation is to clarify, in advance of implementation and, ideally, in the design phase, which questions need to be addressed and in what timeframe. That way, the evaluation design will ensure that these key questions are answered in a timely manner so that the PBAS can benefit from the evaluation process. The designs discussed in this appendix are presented in order of their rigor, although it is important to note that the most rigorous design might not be the most appropriate one in a given situation.

Randomized Control Trial

The RCT is considered the gold standard of evaluation because affected entities (programs, patients, service providers, or users) are randomly assigned to the condition in which they find themselves. The strong appeal of randomization is that it essentially eliminates an important and often unmeasurable source of explanation for demonstrated effects: preexisting or coexisting differences in the groups being compared. Randomization instills confidence that the intervention and comparison groups were equivalent at the point of randomization and that the comparison group will demonstrate all of the nonprogram factors that might affect outcomes (Karoly, Kilburn, and Cannon, 2005). Moreover, RCT designs are highly efficient; statistically significant effects can be found with fewer participating entities than is

the case with other designs (see discussion of regression discontinuity designs in the next section). So, for example, if one wishes to determine whether an intervention designed to improve child-care provider quality is working, one might compare a group of providers who receive a package of coaching, assessment, and quality improvement that is part of a child-care PBAS with a group of providers who receive nothing beyond the usual licensing inspection visits and materials that remind staff to wash their hands frequently and so on. If the providers who got all the quality-improvement support improve their quality more than the providers who got the usual treatment, it might seem obvious that the quality-improvement interventions delivered through the PBAS were the cause. But another possible reason for the difference is that the providers who received quality improvement were different in some a priori way from the "usual treatment" providers. This would be even more likely if the quality-improvement intervention was voluntary, since then it would be even more likely that those who stepped forward asking for quality-improvement support were different from those who did not (either more open to change and improvement, or lower in quality, or some mix of both). However, if all providers who volunteered to participate in the PBAS were randomly assigned to receive the quality-improvement intervention or serve as comparisons, all of the ways in which they might differ—e.g., the socioeconomic status of the children who attend the centers, the skill of the director, staff motivation to improve—would be randomized out, allowing an unambiguous claim to be made about the effects of the quality-improvement intervention.

This design concept—random assignment to conditions—can be applied to multiple interventions, as well. For example, providers who volunteer might be assigned to quality-improvement support in the form of written rating-based feedback alone, a feedback-plus-coaching condition, or usual practice. Indeed, we strongly recommend that such multicondition designs be put in place because they unambiguously communicate that the policy community is not putting all its eggs into one basket. Indeed, in such a comparative study, there is less chance of creating winners and losers: If policy A is shown to be better than policy B, fine. If the reverse proves to be the case, that is fine, too. It is

unlikely that the answer will be that they are all equally inadequate, but, even if that happens, the fact that multiple approaches were tried confers on the policymakers responsible for the effort a mantle of openness, rigor, and responsibility.

However, it is not always possible or even wise to conduct an evaluation based on an RCT design. One reason not to do so is that the PBAS might not yet be ready for such a costly and rigid trial. A series of pilot studies at the onset of a PBAS might make more sense; once the most promising system elements are identified, an RCT design could be considered.

A key reason that RCT designs might not be used is that these designs require that resources be put into assessing an untreated comparison group, a potentially costly process that contributes to evaluation findings but not to other aspects of the program. Moreover, the vaunted controls associated with this design are very difficult to retain, particularly when the design is applied to government services rather than laboratory studies. Sometimes, particularly in health-care PBASs, it is not politically or ethically feasible to randomly assign patients to physicians or treatments, or a promising treatment might need to be provided to everyone who has a particular diagnosis. An RCT might also be unacceptable because the costs of tracking an untreated comparison group in a sector with limited funds and limited regular data collection impose unacceptably high costs, as well as opposition from those who are less supportive of evaluation. In these instances, other quasi-experimental design approaches might be used, such as the two designs described in more detail in the next two sections.

Regression-Discontinuity Designs

When it is simply not practical from a cost, design, or political perspective to pursue an RCT design, other evaluation designs might be employed that can provide answers to the same questions that RCT designs address. A design that is becoming more frequently used of late is the regression-discontinuity design, a pretest-posttest program–comparison-group strategy that evaluates the causal effects of interventions. These designs were first introduced in the evaluation literature by Thistlethwaite and Campbell (1960) but have received scant attention

until quite recently. What sets regression-discontinuity designs apart from other pre-post group designs is the way in which participating entities (patients, schools, counties) are assigned to either the treatment or the comparison condition: Assignment is essentially determined by the participant's score on a key measure lying on either side of a designated threshold (Imbens and Lemieux, 2007). In practice, this means that a cutoff score is established on a preprogram measure, such as blood-pressure level, score on an achievement test, or customer satisfaction level, and entities are assigned to the intervention or comparison group based on that score. The assumption underlying the design is that observations close to the cutoff are similar at baseline in observed and nonobserved ways. For example, if applicants to a state university are admitted on the basis of their rank in their high-school class, it might be reasonable to assume that those who just made the cutoff are not different from those who just missed it. However, in some cases, for example, if the cutoff is based on students' scores on an exam and some students retake the exam until they make the cutoff, then the assumption of equality cannot be made, and one is looking at fuzzy regression discontinuity. While an evaluation can still occur, steps must be taken to deal with the obvious selection effects. Having good institutional knowledge is obviously key in designing evaluations based on regression discontinuity (Imbens and Lemieux, 2007).

A major advantage of the design is that all of those who most need an intervention based on their preintervention score receive it; there is no need to create an untreated comparison group to determine whether an intervention worked, as is the case with RCT designs. Another is that the design relies on existing data and can be implemented in cases in which RCT cannot. The design might also be used to compare entities that receive two different interventions, which might provide some political advantage. Analyses are based on the assumption that, in the absence of the intervention, the pre-post relationship would be equivalent for the two groups. Intervention effects are demonstrated through discontinuity at the cutoff point.

The strength of regression-discontinuity designs depends on the degree to which the assumption of no spurious discontinuity in the pre-post relationship at the cutoff point holds and the degree to which the

pre-post relationship can be known and correctly modeled (Trochim, 2006b). A key disadvantage of these designs compared with RCT is their more-limited statistical power: More than twice as many entities must be involved to achieve comparable statistical power. Another is that their generalizability might be limited (Imbens and Lemieux, 2007).

Nonequivalent-Group Designs

Nonequivalent-group designs are structured like a pretest-posttest randomized experiment, but they lack the key feature and strength of those designs: random assignment. These designs typically rely on intact groups to serve as treatment and comparison groups, e.g., two classrooms at the same grade level in a single school. While these designs attempt to select entities that are as similar as possible to serve as intervention and comparison groups and frequently use statistical controls to adjust for known or suspected preselection differences, it is not possible to ensure that the groups are, in fact, comparable. As a result, any differences or lack of differences that are found cannot unambiguously be attributed to the intervention. In the worst case, this can lead to conclusions that the program did not make a difference when, in fact, it did or that it did make a difference when, in fact, it did not (Trochim, 2006a).

Despite their drawbacks, these designs are popular because they are easier and less costly to implement in many cases and because statistical controls are generally assumed to be adequate to deal with the most egregious threats to internal validity.

Lagged-Implementation Designs

These designs rely on time and geography to create comparison and intervention groups. By implementing an intervention in a particular county or part of a state, for example, while measuring key behaviors in both the intervention and nonintervention sites, the design effectively creates intervention and comparison groups that can be compared on key dimensions. In addition, comparison-group indicators can be compared before and after the intervention is implemented to control for the effects of outside events that otherwise might be attributed spuri-

ously to the intervention. The appeal of this design is considerable: It does not deny anyone the intervention but merely delays it. It allows jurisdictions or other groups to serve as their own pre-post controls as well. However, the design has several drawbacks that have reduced its popularity. First, it requires that some entities wait to receive the intervention. This might be politically unpalatable. It might also increase costs, since some intervention components, e.g., a public relations campaign to inform parents of a new child-care quality rating system, can be accomplished more inexpensively statewide. Of even greater concern, when lagged-implementation designs are promoted as a way to pilot an intervention, the natural inclination to refine the intervention in response to pilot-test experience and results interferes with the need, in the lagged-implementation design, to keep the intervention consistent across sites or entities. To the degree that the intervention changes, the design is weakened as an assessment of its effects.

Case Studies

Case studies are the least rigorous approach to evaluation but are often extremely valuable because their lack of rigor enables researchers to identify new ideas and pursue promising leads in the course of the work. In depth, qualitative assessments and observations, which are the hallmark of many case-study designs, provide insights into the production process in a way that more-structured tools might not. Conversations with key stakeholders about important PBAS components and the larger context in which the PBAS operates can help pinpoint important issues and problems that can help to improve the PBAS design and better align it with other, ongoing efforts to which PBAS target groups must attend.

The in-depth interviews that are the hallmark of case studies also might promote candor that might not emerge through other data-collection approaches. For example, in RAND's landmark studies of the implementation of education innovations in the 1970s, teachers told researchers that they often put new curricula in closets in the back of their classrooms and continued to teach as before, which led researchers to understand the complexity and importance of the imple-

mentation process in understanding how it is that innovations sometimes appear to fail (Berman and McLaughlin, 1978).

Choosing an Evaluation Approach

Clearly, there are many ways to evaluate a PBAS. In general, it is advisable to consider, in choosing an approach, issues of context, costs, ambition, and developmental level of the PBAS. Important as well is to determine what questions need to be answered to build a better PBAS and determine its value. It is often wise to conduct an evaluability assessment, a systematic process that helps identify whether program evaluation is justified, feasible, and likely to provide useful information. A good evaluability assessment shows not only whether a program can be meaningfully evaluated but also whether conducting the evaluation is likely to contribute to improved program performance and management (see Wholey, Hatry, and Newcomer, 2004, for discussion of this approach).

Newer PBASs might benefit most from targeted evaluation approaches, in which the focus is on just one or a handful of implementation measures and no attempt is made to assess the effectiveness of the intervention as a whole. For example, an assessment of a child-care quality rating system might begin by conducting a survey of parents who call an R&R in pursuit of child care to determine how many have heard of the new rating system. Or, the results of a physician training intervention might be assessed through a patient preference survey. These approaches might be a response to limited evaluation funds, a sense that certain aspects of a PBAS are in particular need of scrutiny, or a preference for first determining whether key system components, e.g., sufficient parental awareness, are in place before a launch of a fuller evaluation effort.

These designs can be extremely helpful in the short term in pinpointing areas in which more work needs to be done to effectively implement a PBAS. If, for example, a parent survey reveals that few parents know of the rating system, these results might spur efforts to better publicize the effort, since informed parents are a key part

of the logic model of QRIS effectiveness. Later on, such evaluations might also help to explain limited intervention effects. If, for example, a QRIS evaluation reveals little quality improvement among providers, findings from an early parent survey that showed limited awareness of the QRIS might help explain the lack of change in provider quality, especially if the survey results did not spur the development and implementation of more parent-awareness efforts. If parents are not aware of the rating process or the QRIS more generally, the underlying logic model suggests that providers will be less likely to try to improve the quality of the care they provide.

Perhaps the best approach to evaluation is an incremental one, in which designs are matched to stages of implementation. Such matching ultimately saves money (since more-expensive rigor is reserved until the point at which the PBAS is ready for it), reduces the likelihood that promising PBASs are discarded prematurely, and permits refinement of PBAS designs and implementation processes based on findings from short-term, targeted evaluation efforts. We know, for example, that interventions take time to be fully implemented. It makes little sense to launch a full-bore outcome evaluation when an intervention has not been fully implemented, those who are supposed to work together have not yet begun to do so, and necessary changes to standard operating procedures or infrastructure have not yet been put in place.

Bibliography

AASHTO—*See* American Association of State Highway and Transportation Officials.

American Association of State Highway and Transportation Officials, *Transportation: Invest in Our Future: Accelerating Project Delivery*, Washington, D.C., 2007. As of June 6, 2010:
http://www.transportation1.org/tif7report/

American Law Institute, *Restatement of the Law Third: Restatement of the Law, Agency*, St. Paul, Minn., 2006.

Amrein, Audrey L., and David C. Berliner, "High-Stakes Testing and Student Learning," *Education Policy Analysis Archives*, Vol. 10, No. 18, March 28, 2002. As of June 6, 2010:
http://epaa.asu.edu/ojs/article/view/297

Amundson, Gail, Leif I. Solberg, Maureen Reed, E. Mary Martini, and Richard Carlson, "Paying for Quality Improvement: Compliance with Tobacco Cessation Guidelines," *Joint Commission Journal on Quality and Patient Safety*, Vol. 29, No. 2, February 2003, pp. 59–65.

Armour, B. S., C. Friedman, M. M. Pitts, J. Wike, L. Alley, and J. Etchason, "The Influence of Year-End Bonuses on Colorectal Cancer Screening," *American Journal of Managed Care*, Vol. 10, No. 9, September 2004, pp. 617–624.

Artley, Will, *The Performance-Based Management Handbook*, Vol. 3: *Establishing Accountability for Performance*, Oak Ridge, Tenn.: Oak Ridge Associated Universities, Performance-Based Management Special Interest Group, September 2001. As of June 7, 2010:
http://www.orau.gov/pbm/pbmhandbook/Volume%203.pdf

Asch, Steven M., Elizabeth A. McGlynn, Mary M. Hogan, Rodney A. Hayward, Paul Shekelle, Lisa Rubenstein, Joan Keesey, John Adams, and Eve A. Kerr, "Comparison of Quality of Care for Patients in the Veterans Health Administration and Patients in a National Sample," *Annals of Internal Medicine*, Vol. 141, No. 12, December 2004, pp. 938–945.

Atkinson, A. B., *Atkinson Review: Final Report—Measurement of Government Output and Productivity for the National Accounts*, Basingstoke, UK: Palgrave Macmillan, 2005.

Attorney General of the State of New York, *In the Matter of Connecticut General Life Insurance Company and CIGNA HealthCare of New York, Inc., Agreement Concerning Physician Performance Measurement, Reporting and Tiering Programs*, October 29, 2007. As of June 7, 2010: http://www.oag.state.ny.us/media_center/2007/oct/CIGNA%20Settlement%20Final.pdf

Bae, Chang-Hee Christine, "Transportation and the Environment," in Susan Hanson and Genevieve Giuliano, eds., *The Geography of Urban Transportation*, 3rd ed., New York: Guilford Press, 2004, pp. 356–381.

Beaton, Albert E., Ina V. S. Mullis, Michael O. Martin, Eugenio J. Gonzalez, Dana L. Kelly, and Teresa A. Smith, *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*, Chestnut Hill, Mass.: TIMSS International Study Center, Boston College, 1996.

BEES—*See* Board on Energy and Environmental Systems.

Berman, Paul, and Milbrey Wallin McLaughlin, *Federal Programs Supporting Educational Change*, Vol. VIII: *Implementing and Sustaining Innovations*, Santa Monica, Calif.: RAND Corporation, R-1589/8-HEW, 1978. As of June 21, 2010: http://www.rand.org/pubs/reports/R1589.8/

Bishop, John H., "The Effect of National Standards and Curriculum-Based Exams on Achievement," *American Economic Review*, Vol. 87, No. 2, May 1997, pp. 260–264.

———, *Do Curriculum-Based External Exit Exam Systems Enhance Student Achievement?* Philadelphia, Pa.: Consortium for Policy Research in Education, University of Pennsylvania, Graduate School of Education, research report RR-40, 1998.

Bishop, John, Ferran Mane, and Michael Bishop, *Is Standards-Based Reform Working? And for Whom?* Ithaca, N.Y.: Center for Advanced Human Resource Studies, Cornell University, 2001. As of June 6, 2010: http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1073&context=cahrswp

Board on Energy and Environmental Systems, National Research Council, *Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards*, Washington, D.C.: National Academy Press, 2002. As of June 6, 2010: http://www.nap.edu/openbook.php?isbn=0309076013

Bohrnstedt, George W., and Brian M. Stecher, eds., *Class Size Reduction in California: Early Evaluation Findings, 1996–1998*, Sacramento, Calif.: CSR Research Consortium, June 1999.

———, eds., *What We Have Learned About Class Size Reduction in California*, Palo Alto, Calif.: American Institutes for Research, 2002. As of June 6, 2010: http://www.classize.org/techreport/CSRYear4_final.pdf

Booher-Jennings, Jennifer, "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal*, Vol. 42, No. 2, Summer 2005, pp. 231–268.

Brudney, Jeffrey L., F. Ted Hebert, and Deil S. Wright, "Reinventing Government in the American States: Measuring and Explaining Administrative Reform," *Public Administration Review*, Vol. 59, No. 1, 1999, pp. 19–30.

Burchinal, Margaret R., Joanne E. Roberts, Laura A. Nabors, and Donna M. Bryant, "Quality of Center Child Care and Infant Cognitive and Language Development," *Child Development*, Vol. 67, No. 2, April 1996, pp. 606–620.

Cambridge Systematics, *A Guidebook for Performance-Based Transportation Planning*, Washington, D.C.: National Academy Press, National Cooperative Highway Research Program report 446, 2000.

Camm, Frank, Jeffrey A. Drezner, Bech E. Lachman, and Susan A. Resetar, *Implementing Proactive Environmental Management: Lessons Learned from Best Commercial Practice*, Santa Monica, Calif.: RAND Corporation, MR-1371-OSD, 2001. As of June 10, 2010: http://www.rand.org/pubs/monograph_reports/MR1371/

Camm, Frank, and Brian M. Stecher, *Analyzing the Operation of Performance-Based Accountability Systems for Public Services*, Santa Monica, Calif.: RAND Corporation, TR-853, 2010. As of August 2010: http://www.rand.org/pubs/technical_reports/TR853/

Campbell, Frances A., and Craig T. Ramey, "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention," *American Educational Research Journal*, Vol. 32, No. 4, Winter 1995, pp. 743–772.

Campbell, Stephen, David Reeves, Evangelos Kontopantelis, Elizabeth Middleton, Bonnie Sibbald, and Martin Roland, "Quality of Primary Care in England with the Introduction of Pay for Performance," *New England Journal of Medicine*, Vol. 357, No. 2, July 12, 2007, pp. 181–190.

Campbell, Stephen M., David Reeves, Evangelos Kontopantelis, Bonnie Sibbald, and Martin Roland, "Effects of Pay for Performance on the Quality of Primary Care in England," *New England Journal of Medicine*, Vol. 361, No. 4, July 23, 2009, pp. 368–378.

Carman, Joanne G., "Evaluation Practice Among Community-Based Organizations: Research into the Reality," *American Journal of Evaluation*, Vol. 28, No. 1, 2007, pp. 60–75.

Carnoy, Martin, and Susanna Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis*, Vol. 24, No. 4, 2002, pp. 305–331.

Carr, David K., and Ian D. Littman, *Excellence in Government: Total Quality Management in the 1990s*, Arlington, Va.: Coopers and Lybrand, 1990.

Casalino, Lawrence P., Arthur Elster, Andy Eisenberg, Evelyn Lewis, John Montgomery, and Diana Ramos, "Will Pay-for-Performance and Quality Reporting Affect Health Care Disparities?" *Health Affairs*, Vol. 26, No. 3, 2007, pp. w405–w414.

CBO—*See* Congressional Budget Office.

Center on Education Policy, *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act*, Washington, D.C., 2006. As of June 6, 2010: http://www.cep-dc.org/index.cfm?fuseaction=page.viewPage&pageID=540&nodeID=1

———, *Are Achievement Gaps Closing and Is Achievement Rising for All?* Washington, D.C., 2009. As of June 6, 2010: http://www.cep-dc.org/document/docWindow.cfm?fuseaction=document.viewDocument&documentid=292&documentFormatId=4388

Chandler, Alfred D., *The Visible Hand: The Managerial Revolution in American Business*, Cambridge, Mass.: Belknap Press, 1977.

Chen, Huey-tsyh, *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness*, Thousand Oaks, Calif.: Sage Publications, 2005.

Clarke-Stewart, K. Alison, Deborah Lowe Vandell, Margaret Burchinal, Marion O'Brien, and Kathleen McCartney, "Do Regulable Features of Child-Care Homes Affect Children's Development?" *Early Childhood Research Quarterly*, Vol. 17, No. 1, March 2002, pp. 52–86.

Cohen, Steven, and Ronald Brand, *Total Quality Management in Government: A Practical Guide for the Real World*, San Francisco, Calif.: Jossey-Bass, 1993.

Congressional Budget Office, *Reducing Gasoline Consumption: Three Policy Options*, Washington, D.C., November 2002. As of June 6, 2010: http://purl.access.gpo.gov/GPO/LPS24735

———, *The Economic Costs of Fuel Economy Standards Versus a Gasoline Tax*, Washington, D.C., December 2003. As of June 6, 2010: http://purl.access.gpo.gov/GPO/LPS72686

Cook, Thomas J., and Judson J. Lawrie, *Use of Performance Standards and Measures for Public Transportation Systems*, Raleigh, N.C.: North Carolina Department of Transportation, Research and Analysis Group, September 2004. As of June 6, 2010: http://www.ncdot.org/doh/preconstruct/tpb/research/download/2004-10FinalReport.pdf

Crosby, Philip B., *Quality Is Free: The Art of Making Quality Certain*, New York: McGraw-Hill, 1979.

Curristine, Teresa, "Performance and Accountability: Making Government Work," *OECD Observer*, No. 252–253, November 2005, pp. 11–12. As of June 6, 2010:
http://www.oecdobserver.org/news/fullstory.php/aid/1697/Performance_and_accountability:_Making_government_work.html

Curtin, K., H. Beckman, G. Pankow, Y. Milillo, and R. A. Green, "Return on Investment in Pay for Performance: A Diabetes Case Study," *Journal of Healthcare Management*, Vol. 51, No. 6, November–December 2006, pp. 365–374, discussion pp. 375–376.

Damberg, Cheryl L., Melony E. Sorbero, Ateev Mehrotra, Stephanie Teleki, Susan Lovejoy, and Lily Bradley, *An Environmental Scan of Pay for Performance in the Hospital Setting: Final Report*, Santa Monica, Calif.: RAND Corporation, WR-474-ASPE/CMS, November 2007. As of June 6, 2010:
http://aspe.hhs.gov/health/reports/08/payperform/PayPerform07.html

Damberg, Cheryl L., Kristiana Raube, Stephanie S. Teleki, and Erin dela Cruz, "Taking Stock of Pay-for-Performance: A Candid Assessment from the Front Lines," *Health Affairs*, Vol. 28, No. 2, 2009, pp. 517–525.

Davenport, Thomas H., *Process Innovation: Reengineering Work Through Information Technology*, Boston, Mass.: Harvard Business School Press, 1993.

Dee, Thomas S., and Brian Jacob, *The Impact of No Child Left Behind on Student Achievement*, Cambridge, Mass.: National Bureau of Economic Research, working paper 15531, 2009. As of June 21, 2010:
http://www.nber.org/papers/w15531

Dixon, Lloyd, and Steven Garber, *Fighting Air Pollution in Southern California by Scrapping Old Vehicles*, Santa Monica, Calif.: RAND Corporation, MR-1256-PPIC/ICJ, 2001. As of June 10, 2010:
http://www.rand.org/pubs/monograph_reports/MR1256/

Doran, Tim, Catherine Fullwood, Evangelos Kontopantelis, and David Reeves, "Effect of Financial Incentives on Inequalities in the Delivery of Primary Clinical Care in England: Analysis of Clinical Activity Indicators for the Quality and Outcomes Framework," *Lancet*, Vol. 372, No. 9640, August 2008, pp. 728–736.

Ellig, Jerry, "Scoring Government Performance Reports: Telling It Like It Is, or Not?" *Public Manager*, Vol. 36, No. 2, June 22, 2007, pp. 3–8.

EPA—*See* U.S. Environmental Protection Agency.

Fairbrother, Gerry, Stephen Friedman, Karla L. Hanson, and Gary C. Butts, "Effect of the Vaccines for Children Program on Inner-City Neighborhood Physicians," *Archives of Pediatrics and Adolescent Medicine*, Vol. 151, No. 12, December 1997, pp. 1229–1235.

Fairbrother, Gerry, Karla L. Hanson, Stephen Friedman, and Gary C. Butts, "The Impact of Physician Bonuses, Enhanced Fees, and Feedback on Childhood Immunization Coverage Rates," *American Journal of Public Health*, Vol. 89, No. 2, February 1999, pp. 171–175.

Fairbrother, Gerry, Michele J. Siegel, Stephen Friedman, Pierre D. Kory, and Gary C. Butts, "Impact of Financial Incentives on Documented Immunization Rates in the Inner City: Results of a Randomized Controlled Trial," *Ambulatory Pediatrics*, Vol. 1, No. 4, July–August 2002, pp. 206–212.

Feller, Irwin, "Performance Measurement Redux," *American Journal of Evaluation*, Vol. 23, No. 4, 2002, pp. 435–452.

Ferlie, Ewan, "The New Public Management in the United Kingdom: Origins, Implementation, and Prospects," paper presented at the International Seminar on Managerial Reform of the State, Brasilia, Brazil, November 1998.

Fielding, Gordon J., S. R. Mundle, and J. Misner, "Performance-Based Funding-Allocation Guidelines for Transit Operators in Los Angeles County," *Transportation Research Record*, Vol. 857, 1982, pp. 14–18.

Finn, Jeremy D., "Class-Size Reduction in Grades K–3," in Alex Molnar, ed., *School Reform Proposals: The Research Evidence*, Greenwich, Conn.: Information Age Pub., 2002, pp. 15–24.

Francis, David O., Howard Beckman, John Chamberlain, Greg Partridge, and Robert A. Greene, "Introducing a Multifaceted Intervention to Improve the Management of Otitis Media: How Do Pediatricians, Internists, and Family Physicians Respond?" *American Journal of Medical Quality*, Vol. 21, No. 2, 2006, pp. 134–143.

Friedberg, Mark W., Ateev Mehrotra, and Jeffrey A. Linder, "Reporting Hospitals' Antibiotic Timing in Pneumonia: Adverse Consequences for Patients?" *American Journal of Managed Care*, Vol. 15, No. 2, February 2009, pp. 137–144.

Fryer, Roland G. Jr., *Financial Incentives and Student Achievement: Evidence from Randomized Trials*, Cambridge, Mass.: National Bureau of Economic Research, working paper 15898, April 2010. As of June 6, 2010: http://papers.nber.org/papers/15898

Fung, Constance H., Yee-Wei Lim, Soeren Mattke, Cheryl Damberg, and Paul G. Shekelle, "Systematic Review: The Evidence That Publishing Patient Care Performance Data Improves Quality of Care," *Annals of Internal Medicine*, Vol. 148, No. 2, January 15, 2008, pp. 111–123.

Gayer, Ted, "The Fatality Risks of Sport-Utility Vehicles, Vans, and Pickups Relative to Cars," *Journal of Risk and Uncertainty*, Vol. 28, No. 2, March 2004, pp. 103–133.

George, Stephen, and Arnold Weimerskirch, *Total Quality Management: Strategies and Techniques Proven at Today's Most Successful Companies*, New York: Wiley, 1994.

Gilmour, John B., and David E. Lewis, "Assessing Performance Budgeting at OMB: The Influence of Politics, Performance, and Program Size," *Journal of Public Administration Research and Theory*, Vol. 16, No. 2, 2006a, pp. 169–186.

———, "Does Performance Budgeting Work? An Examination of the Office of Management and Budget's PART Scores," *Public Administration Review*, Vol. 66, No. 5, September–October 2006b, pp. 742–752.

Gormley, William T., Jr., *Everybody's Children: Child Care as a Public Problem*, Washington, D.C.: Brookings Institution, 1995.

———, personal communication with the authors, 2010.

Gormley, William T., Jr., and Cristina Boccuti, "HCFA and the States: Politics and Intergovernmental Leverage," *Journal of Health Politics, Policy and Law*, Vol. 26, No. 3, 2001, pp. 557–580.

Gormley, William T., Jr. and David Leo Weimer, *Organizational Report Cards*, Cambridge, Mass.: Harvard University Press, 1999.

Governmental Accounting Standards Board and National Academy of Public Administration, *Report on Survey of State and Local Government Use and Reporting of Performance Measures*, Washington, D.C., 1997.

Grady, Kathleen E., Jeanne Parr Lemkau, Norma R. Lee, and Cheryl Caddell, "Enhancing Mammography Referral in Primary Care," *Preventive Medicine*, Vol. 26, No. 6, November 1997, pp. 791–800.

Greene, Robert A., Howard Beckman, John Chamberlain, Greg Partridge, Marla Miller, Diane Burden, and Jamie Kerr, "Increasing Adherence to a Community-Based Guideline for Acute Sinusitis Through Education, Physician Profiling, and Financial Incentives," *American Journal of Managed Care*, Vol. 10, No. 10, October 2004, pp. 670–678.

Hamilton, Laura S., "Assessment as a Policy Tool," Santa Monica, Calif.: RAND Corporation, RP-1163, 2004. (Reprinted from *Review of Research in Education*, Vol. 27, No. 1, 2003, pp. 25–68.) As of June 6, 2010: http://www.rand.org/pubs/reprints/RP1163/

———, personal communication with the authors, December 10, 2009.

Hamilton, Laura S., Brian M. Stecher, Julie A. Marsh, Jennifer Sloan McCombs, Abby Robyn, Jennifer Russell, Scott Naftel, and Heather Barney, *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States*, Santa Monica, Calif.: RAND Corporation, MG-589-NSF, 2007. As of June 6, 2010: http://www.rand.org/pubs/monographs/MG589/

Hannaway, Jane, and Laura Hamilton, *Accountability Policies: Implications for School and Classroom Practices*, Washington, D.C.: Urban Institute, October 16, 2008. As of June 6, 2010:
http://www.urban.org/url.cfm?ID=411779

Hanushek, Eric A., and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*, Vol. 24, No. 2, Spring 2005, pp. 297–327.

Harry, Mikel J., *The Nature of Six Sigma Quality*, Schaumburg, Ill.: Motorola University Press, 1988.

Hartman, Ronald J., Elaine M. Kurtz, and Alan B. Winn, *The Role of Performance-Based Measures in Allocating Funding for Transit Operations*, Washington, D.C.: National Academy Press, 1994.

Hatry, Harry P., with Joseph S. Wholey, *Performance Measurement: Getting Results*, Washington, D.C.: Urban Institute Press, 1999.

Hatry, Harry P., *Governing for Results: Improving Federal Government Performance and Accountability: Suggestions for the New Federal Administration*, Washington, D.C.: Urban Institute, October 15, 2008. As of June 7, 2010:
http://www.ombwatch.org/files/performance/HATRYgoverningforresults.pdf

Hatry, Harry P., Philip S. Schaenman, Donald M. Fisk, John R. Hall Jr., and Louise Snyder, *How Effective Are Your Community Services? Procedures for Performance Measurement*, Washington, D.C.: International City Management Association and Urban Institute, 2006.

Herb, Jeanne, Jennifer Sullivan, Mark Stoughton, and Allen White, *The National Environmental Performance Partnership System: Making Good on Its Promise?* Washington, D.C.: National Academy of Public Administration, 2000. As of June 7, 2010:
http://www.napawash.org/pc_economy_environment/epafile12.pdf

Hess, Frederick M., "Science and Nonscience: The Limits of Scientific Research," *On the Issues*, May 2005. As of June 21, 2010:
http://www.aei.org/docLib/20050505_OTIHess_g.pdf

Hibbard, Judith H., Jean Stockard, and Martin Tusler, "Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts?" *Health Affairs*, Vol. 22, No. 2, March–April 2003, pp. 84–94.

———, "Hospital Performance Reports: Impact on Quality, Market Share, and Reputation," *Health Affairs*, Vol. 24, No. 4, 2005, pp. 1150–1160.

Hibbard, Judith H., P. Slovic, E. Peters, and M. L. Finucane, "Strategies for Reporting Health Plan Performance Information to Consumers: Evidence from Controlled Studies," *Health Services Research*, Vol. 37, No. 2, April 2002, pp. 291–313.

Hillman, Alan L., Kimberly Ripley, Neil Goldfarb, Isaac Nuamah, Janet Weiner, and Edward Lusk, "Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care," *American Journal of Public Health*, Vol. 88, No. 11, November 1998, pp. 1699–1701.

Hillman, Alan L., Kimberly Ripley, Neil Goldfarb, Janet Weiner, Isaac Nuamah, and Edward Lusk, "The Use of Physician Financial Incentives and Feedback to Improve Pediatric Preventive Care in Medicaid Managed Care," *Pediatrics*, Vol. 104, No. 1, October 1999, pp. 931–935.

Hitch, Charles Johnston, and Roland N. McKean, *The Economics of Defense in the Nuclear Age*, Santa Monica, Calif.: RAND Corporation, R-346, 1960. As of June 7, 2010:
http://www.rand.org/pubs/reports/R346/

Ho, Alfred Tat-Kei, "Accounting for the Value of Performance Measurement from the Perspective of Midwestern Mayors," *Journal of Public Administration Research and Theory*, Vol. 16, No. 2, 2006, pp. 217–237.

Howes, Carollee, "Relations Between Early Child Care and Schooling," *Developmental Psychology*, Vol. 24, No. 1, January 1988, pp. 53–57.

Huesing, Tina, "Six Sigma Through the Years," briefing, Motorola, October 20, 2008. As of June 6, 2010:
http://sigmaexperts.com/presentations/Six_Sigma_Through_the_Years.pdf

Hughes, Adam, and J. Robert Shull, *PART Backgrounder*, Washington, D.C.: OMB Watch, April 2005.

Iezzoni, Lisa I., "100 Apples Divided by 15 Red Herrings: A Cautionary Tale from the Mid-19th Century on Comparing Hospital Mortality Rates," *Annals of Internal Medicine*, Vol. 124, No. 12, June 15, 1996, pp. 1079–1085.

Imbens, Guido, and Thomas Lemieux, *Regression Discontinuity Designs: A Guide to Practice*, Cambridge, Mass.: National Bureau of Economic Research, working paper 13039, 2007. As of June 6, 2010:
http://papers.nber.org/papers/13039

Institute of Medicine, Committee on Quality of Health Care in America, *Crossing the Quality Chasm: A New Health System for the 21st Century*, Washington, D.C.: National Academy Press, 2001. As of June 7, 2010:
http://www.nap.edu/catalog/10027.html

IOM—*See* Institute of Medicine.

Jacob, Brian Aaron, *Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools*, Cambridge, Mass.: National Bureau of Economic Research, working paper 8968, June 2002. As of June 21, 2010:
http://papers.nber.org/papers/W8968.pdf

Jaeger, Richard M., "The Final Hurdle," in Gilbert R. Austin and Herbert Garber, eds., *The Rise and Fall of National Test Scores*, New York: Academic Press, 1982.

Johnston, Robert A., "The Urban Transportation Planning Process," in Susan Hanson and Genevieve Giuliano, eds., *The Geography of Urban Transportation*, 3rd ed., New York: Guilford Press, 2004, pp. 115–140.

Kagan, L. S., "Buckets, Banks, and Hearts: Aligning Early Childhood Standards and Systems," presentation, Quality Rating and Improvement Systems: Creating the Next Generation of QRISs conference, St. Paul, Minn., June 4, 2008.

Kaplan, Robert S., and David P. Norton, *The Balanced Scorecard: Translating Strategy into Action*, Boston, Mass.: Harvard Business School Press, 1996.

———, *Alignment: Using the Balanced Scorecard to Create Corporate Synergies*, Boston, Mass.: Harvard Business School Press, 2006.

Karoly, Lynn A., Bonnie Ghosh-Dastidar, Gail L. Zellman, Michal Perlman, and Lynda Fernyhough, *Prepared to Learn: The Nature and Quality of Early Care and Education for Preschool-Age Children in California*, Santa Monica, Calif.: RAND Corporation, TR-539-PF/WKKF/PEW/NIEER/WCJVSF/LAU, 2008. As of June 21, 2010:
http://www.rand.org/pubs/technical_reports/TR539/

Karoly, Lynn A., M. Rebecca Kilburn, and Jill S. Cannon, *Early Childhood Interventions: Proven Results, Future Promises*, Santa Monica, Calif.: RAND Corporation, MG-341-PNC, 2005. As of June 21, 2010:
http://www.rand.org/pubs/monographs/MG341/

Keating, Edward G., and Elvira N. Loredo, *Valuing Programmed Depot Maintenance Speed: An Analysis of F-15 PDM*, Santa Monica, Calif.: RAND Corporation, TR-377-AF, 2006. As of June 10, 2010:
http://www.rand.org/pubs/technical_reports/TR377/

Keehley, Pat, "FQI Highlights Quality Management," *Public Administration Times*, July 1, 1990, p. 3.

Kettl, Donald F., *Reinventing Government: A Fifth-Year Report Card*, Washington, D.C.: Center for Public Management, Brookings Institution, 1998.

———, *The Global Public Management Revolution: A Report on the Transference of Governance*, 2nd ed., Washington, D.C.: Brookings Institution, 2005.

Kettl, Donald F., and John J. DiIulio, *Inside the Reinvention Machine: Appraising Governmental Reform*, Washington, D.C.: Brookings Institution, 1995.

Kohn, Linda T., Janet Corrigan, and Molla S. Donaldson, *To Err Is Human: Building a Safer Health System*, Washington, D.C.: National Academy Press, 1999. As of June 7, 2010:
http://www.nap.edu/catalog/9728.html

Koretz, Daniel M., "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources*, Vol. 37, No. 4, Fall 2002, pp. 752–777.

Koretz, Daniel M., and Laura S. Hamilton, "Testing for Accountability in K–12," in Robert L. Brennan, ed., *Educational Measurement*, Westport, Conn.: Praeger Publishers, 2006, pp. 531–578.

Koretz, Daniel M., K. Mitchell, S. Barron, and S. Keith, *Final Report: Perceived Effects of the Maryland School Performance Assessment Program*, Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California, Los Angeles, March 1996.

Kotter, John P., *Leading Change*, Boston, Mass.: Harvard Business School Press, 1996.

Kouides, Ruth W., Nancy M. Bennett, Bonnie Lewis, Joseph D. Cappuccio, William H. Barker, F. Marc LaForce, and the primary-care physicians of Monroe County, "Performance-Based Physician Reimbursement and Influenza Immunization Rates in the Elderly," *American Journal of Preventive Medicine*, Vol. 14, No. 2, February 1998, pp. 89–95.

Laffont, Jean-Jacques, and Jean Tirole, *A Theory of Incentives in Procurement and Regulation*, Cambridge, Mass.: MIT Press, 1993.

Lane, Suzanne, Carol S. Parke, and Clement A. Stone, "The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance," *Educational Assessment*, Vol. 8, No. 4, November 2002, pp. 279–315.

Le Floch, Kerstin Carlson, Jose Felipe Martinez, Jennifer O'Day, Brian M. Stecher, James Taylor, Andrea Cook, Georges Vernez, Beatrice Birman, and Michael Garet, *State and Local Implementation of the No Child Left Behind Act*, Vol. III: *Accountability Under NCLB: Interim Report*, Santa Monica, Calif.: RAND Corporation, RP-1303, 2007. (Reprinted with permission from "State and Local Implementation of the No Child Left Behind Act: Volume III—Accountability Under NCLB: Interim Report," by the U.S. Department of Education, Office of Planning, Evaluation and Development, Policy and Program Studies Service, Washington D.C., 2007.) As of June 6, 2010: http://purl.access.gpo.gov/GPO/LPS117062

Levine, Arnold, and Jeffrey Luck, *The New Management Paradigm: A Review of Principles and Practices*, Santa Monica, Calif.: RAND Corporation, MR-458-AF, 1994. As of June 10, 2010: http://www.rand.org/pubs/monograph_reports/MR458/

Light, Paul Charles, *The Tides of Reform: Making Government Work, 1945–1995*, New Haven, Conn.: Yale University Press, 1997.

Lindenauer, Peter K., Denise Remus, Sheila Roman, Michael B. Rothberg, Evan M. Benjamin, Allen Ma, and Dale W. Bratzler, "Public Reporting and Pay for Performance in Hospital Quality Improvement," *New England Journal of Medicine*, Vol. 356, No. 5, February 1, 2007, pp. 486–496.

Liner, Blaine, Harry P. Hatry, Elisa Vinson, Ryan Allen, Pat Dusenbury, Scott Bryant, and Ron Snell, *Making Results-Based State Government Work*, Washington, D.C.: Urban Institute, 2001.

Lipsey, R. G., and Kelvin Lancaster, "The General Theory of Second Best," *Review of Economic Studies*, Vol. 24, No. 1, 1956–1957, pp. 11–32.

Lynn, Laurence E. Jr., *Public Management: Old and New*, New York: Routledge, 2006.

Martin, Timothy Carl, *The Corporate Average Fuel Economy Standards: A History of Compliance and Analysis of Manufacturer Response in the Short Run*, Waltham, Mass.: Brandeis University, International School of Business, doctoral dissertation, 2005.

Martin, V., and M. H. Jobin, "Results-Based Management: Comparison of the Management Framework of Eight Jurisdictions," *Canadian Public Administration—Administration Publique du Canada*, Vol. 47, No. 3, 2004, pp. 304–331.

McCall, Martha S., G. Gage Kingsbury, and Allan Olson, *Individual Growth and School Success: A Technical Report from the NWEA Growth Research Database*, Lake Oswego, Oreg.: Northwest Evaluation Association, 2004.

McCarthy, James E., *Transportation Conformity Under the Clean Air Act: In Need of Reform?* Washington, D.C.: Congressional Research Service, January 8, 2004. As of June 6, 2010:
http://www.ncseonline.org/NLE/CRSreports/04apr/RL32106.pdf

McGlynn, Elizabeth A., Steven M. Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, and Eve A. Kerr, "The Quality of Health Care Delivered to Adults in the United States," *New England Journal of Medicine*, Vol. 348, No. 26, June 26, 2003, pp. 2635–2645.

Melkers, Julia, and Katherine Willoughby, "Models of Performance-Measurement Use in Local Governments: Understanding Budgeting, Communication, and Lasting Effects," *Public Administration Review*, Vol. 65, No. 2, March 2005, pp. 180–190.

Meltsner, Arnold J., *Policy Analysts in the Bureaucracy*, Berkeley, Calif.: University of California Press, 1986.

Metzenbaum, Shelley H., *Strategies for Using State Information: Measuring and Improving Performance*, Washington, D.C.: IBM Center for the Business of Government, December 2003. As of June 7, 2010: http://www.businessofgovernment.org/report/strategies-using-state-information-measuring-and-improving-performance

Miles, Matthew B., and A. Michael Huberman, *Qualitative Data Analysis: A Sourcebook of New Methods*, Beverly Hills, Calif.: Sage Publications, 1984.

Mishan, E. J., *Cost-Benefit Analysis*, New York: Praeger, 1976.

Mohan, Erin, Grace Reef, and Mousumi Sarkar, *Breaking the Piggy Bank: Parents and the High Price of Child Care*, Arlington, Va.: National Association of Child Care Resource and Referral Agencies, 2006. As of June 21, 2010: http://www.naccrra.org/docs/policy/Breaking%20the%20Piggy%20Bank_FINAL(printer).pdf

Moon, M. Jae, and Peter deLeon, "Municipal Reinvention: Managerial Values and Diffusion Among Municipalities," *Journal of Public Administration Research and Theory*, Vol. 11, No. 3, 2001, pp. 327–351.

Moore, Nancy Y., Laura H. Baldwin, Frank Camm, and Cynthia R. Cook, *Implementing Best Purchasing and Supply Management Practices: Lessons from Innovative Commercial Firms*, Santa Monica, Calif.: RAND Corporation, DB-334-AF, 2002. As of June 7, 2010: http://www.rand.org/pubs/documented_briefings/DB334/

Morrow, Robert W., Anne D. Gooding, and Colleen Clark, "Improving Physicians' Preventive Health Care Behavior Through Peer Review and Financial Incentives," *Archives of Family Medicine*, Vol. 4, No. 2, 1995, pp. 165–169.

Moynihan, Donald P., "Managing for Results in State Government: Evaluating a Decade of Reform," *Public Administration Review*, Vol. 66, No. 1, January 2006, pp. 77–89.

Moynihan, Donald P., and Patricia Wallace Ingraham, "Integrative Leadership in the Public Sector: A Model of Performance-Information Use," *Administration and Society*, Vol. 36, No. 4, 2004, pp. 427–453.

Mullen, Kathleen J., Richard G. Frank, and Meredith B. Rosenthal, "Can You Get What You Pay for? Pay-for-Performance and the Quality of Healthcare Providers," *RAND Journal of Economics*, Vol. 41, No. 1, Spring 2010, pp. 64–91.

Nathan, Richard P., "Presidential Address: 'Complexifying' Performance Oversight in America's Governments," *Journal of Policy Analysis and Management*, Vol. 24, No. 2, Spring 2005, pp. 207–480.

National Association of County and City Health Officials, *Federal Funding for Public Health Emergency Preparedness: Implications and Ongoing Issues for Local Health Departments*, Washington, D.C., August 1, 2007.

National Commission on Excellence in Education, *A Nation at Risk: The Imperative for Educational Reform—A Report to the Nation and the Secretary of Education*, United States Department of Education, Washington, D.C., 1983. As of June 7, 2010:
http://purl.access.gpo.gov/GPO/LPS3244

National Committee for Quality Assurance, *The State of Health Care Quality*, Washington, D.C., 2009.

National Institute of Child Health and Human Development Early Child Care Research Network, "The Relation of Child Care to Cognitive and Language Development," *Child Development*, Vol. 71, No. 4, July–August 2000, pp. 960–980.

National Transportation Policy Project, *Performance Driven: A New Vision for U.S. Transportation Policy*, Washington, D.C.: Bipartisan Policy Center, June 9, 2009.

NCQA—*See* National Committee for Quality Assurance.

Neal, Derek A., and Diane Schanzenbach, *Left Behind by Design: Proficiency Counts and Test-Based Accountability*, Cambridge, Mass.: National Bureau of Economic Research, working paper 13293, 2007. As of June 21, 2010:
http://papers.nber.org/papers/13293

NICHD ECCRN—*See* National Institute of Child Health and Human Development Early Child Care Research Network.

Nightingale, Florence, *Notes on Hospitals*, London: Longman, Green, Longman, Roberts, and Green, 1863.

Novick, David, ed., *Program Budgeting: Program Analysis and Federal Budget*, Cambridge, Mass.: Harvard University Press, 1965.

Office of Homeland Security, *National Strategy for Homeland Security*, Washington, D.C., July 2002. As of June 14, 2010:
http://purl.access.gpo.gov/GPO/LPS20641

Ohno, Taiichi, and Norman Bodek, *Toyota Production System: Beyond Large-Scale Production*, University Park, Ill.: Productivity Press, 1988.

Palladium Group, undated home page. As of June 10, 2010:
http://www.thepalladiumgroup.com

Pande, Peter S., and Robert P. Neuman, *The Six Sigma Way: How GE, Motorola, and Other Top Companies Are Honing Their Performance*, New York: McGraw-Hill, 2000.

Pearson, Steven D., Eric C. Schneider, Ken P. Kleinman, Kathryn L. Coltin, and Janice A. Singer, "The Impact of Pay-for-Performance on Health Care Quality in Massachusetts, 2001–2003," *Health Affairs*, Vol. 27, No. 4, 2008, pp. 1167–1176.

Peisner-Feinberg, Ellen S., and Margaret R. Burchinol, "Relations Between Preschool Children's Child-Care Experiences and Concurrent Development: The Cost, Quality, and Outcomes Study," *Merrill-Palmer Quarterly*, Vol. 43, No. 3, July 1997, pp. 451–477.

Peisner-Feinberg, Ellen S., Margaret R. Burchinal, Richard M. Clifford, Mary L. Culkin, Carollee Howes, Sharon Lynn Kagan, and Noreen Yazejian, "The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Developmental Trajectories Through Second Grade," *Child Development*, Vol. 72, No. 5, September–October 2001, pp. 1534–1553.

Podgursky, Michael J., and Matthew G. Springer, "Teacher Performance Pay: A Review," *Journal of Policy Analysis and Management*, Vol. 26, No. 4, 2007, pp. 551–573.

Poister, Theodore H., and Gregory Streib, "Performance Measurement in Municipal Government: Assessing the State of the Practice," *Public Administration Review*, Vol. 59, No. 4, July–August 1999, pp. 325–335.

Pressman, Jeffrey L., and Aaron B. Wildavsky, *Implementation: How Great Expectations in Washington Are Dashed in Oakland*, Berkeley, Calif.: University of California Press, 1984.

Public Law 88-157, Federal-Aid Highway Amendments Act of 1963, October 24, 1963.

Public Law 88-206, Clean Air Act of 1963, December 17, 1963.

Public Law 89-10, Elementary and Secondary Education Act, April 11, 1965.

Public Law 91-604, Clean Air Act Extension of 1970, December 31, 1970.

Public Law 94-163, Energy Policy and Conservation Act, December 22, 1975.

Public Law 95-95, Clean Air Act Amendments of 1977, August 8, 1977.

Public Law 95-599 Federal-Aid Highway Act of 1978, November 6, 1978.

Public Law 101-549, Clean Air Act Amendments of 1990, November 15, 1990.

Public Law 101-576, Chief Financial Officers Act of 1990, November 15, 1990.

Public Law 103-62, Government Performance and Results Act of 1993, August 3, 1993.

Public Law 103-382, Improving America's Schools Act of 1994, October 20, 1994.

Public Law 107-110, No Child Left Behind Act of 2001, January 8, 2002. As of June 7, 2010:
http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ110.107.pdf

Public Law 108-173, Medicare Prescription Drug Improvement and Modernization Act of 2003, December 8, 2003.

Public Law 109-417, Pandemic and All-Hazards Preparedness Act, December 20, 2006. As of June 7, 2010:
http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s3678enr.txt.pdf

Public Law 110-140, Energy Independence and Security Act of 2007, December 19, 2007. As of June 14, 2010:
http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ140.110

Radin, Beryl A., "A Comparative Approach to Performance Management: Contrasting the Experience of Australia, New Zealand, and the United States," *International Journal of Public Administration*, Vol. 26, No. 12, January 2003, pp. 1355–1376.

Ramey, Craig T., and Sharon Landesman Ramey, "Early Learning and School Readiness: Can Early Intervention Make a Difference?" in Norman F. Watt, Catherine C. Ayoub, Robert H. Bradley, Jini E. Puma, and W. A. LeBouef, eds., *The Crisis in Youth Mental Health: Critical Issues and Effective Programs*, Vol. 4: *Early Intervention Programs and Policies*, Westport, Conn.: Praeger, 2006, pp. 291–318.

Reiter, Kristin L., Tammie A. Nahra, Jeffrey A. Alexander, and John R. C. Wheeler, "Hospital Responses to Pay-for-Performance Incentives," *Health Services Management Research*, Vol. 19, No. 2, 2006, pp. 123–134.

Resnick, D. P., "History of Educational Testing," in Alexandra K. Wigdor, ed., *Ability Testing: Uses, Consequences, and Controversies*, Part 2: *Documentation Section, 1982*, Washington, D.C.: National Academies Press, 1982, pp. 173–194.

Rosenthal, Meredith B., Richard G. Frank, Zhonghe Li, and Arnold M. Epstein, "Early Experience with Pay-for-Performance: From Concept to Practice," *Journal of the American Medical Association*, Vol. 294, No. 14, 2005, pp. 1788–1793.

Roski, Joachim, Robert Jeddeloh, Larry An, Harry Lando, Peter Hannan, Carmen Hall, and Shu-Hong Zhu, "The Impact of Financial Incentives and a Patient Registry on Preventive Care Quality: Increasing Provider Adherence to Evidence-Based Smoking Cessation Practice Guidelines," *Preventive Medicine*, Vol. 36, No. 3, March 2003, pp. 291–299.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman, *Evaluation: A Systematic Approach*, Thousand Oaks, Calif.: Sage Publications, 2004.

Rothman, Robert, J. B. Slattery, J. L. Vranek, and L. B. Resnick, *Benchmarking and Alignment of Standards and Testing*, Los Angeles, Calif.: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California, Los Angeles, May 2002.

Seid, Michael, Debra Lotstein, Valerie L. Williams, Christopher Nelson, Kristin J. Leuschner, Allison Diamant, Stefanie Stern, Jeffrey Wasserman, and Nicole Lurie, "Quality Improvement in Public Health Emergency Preparedness," *Annual Review of Public Health*, Vol. 28, 2007, pp. 19–31.

Senge, Peter M., *The Fifth Discipline: The Art and Practice of the Learning Organization*, New York: Doubleday/Currency, 1990.

Sorbero, Melony E., Cheryl L. Damberg, R. Shaw, *Assessment of Pay-for-Performance Options for Medicare Physician Services: Final Report*, Santa Monica, Calif.: RAND Corporation, unpublished research, 2006.

Stecher, Brian M., "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practices," in Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein, eds., *Making Sense of Test-Based Accountability in Education*, Santa Monica, Calif.: RAND Corporation, MR-1554-EDU, 2002, pp. 79–100. As of June 6, 2010:
http://www.rand.org/pubs/monograph_reports/MR1554/

Stecher, Brian M., Scott Epstein, Laura S. Hamilton, Julie A. Marsh, Abby Robyn, Jennifer Sloan McCombs, Jennifer Russell, and Scott Naftel, *Pain and Gain: Implementing No Child Left Behind in Three States, 2004–2006*, Santa Monica, Calif.: RAND Corporation, MG-784-NSF, 2008. As of June 22, 2010:
http://www.rand.org/pubs/monographs/MG784/

Sterck, Miekatrien, "The Impact of Performance Budgeting on the Role of the Legislature: A Four-Country Study," *International Review of Administrative Sciences*, Vol. 73, No. 2, 2007, pp. 189–203.

Strong, Kelly, *Performance Effectiveness of Design-Build, Lane Rental, and A + B Contracting Techniques*, St. Paul, Minn.: Minnesota Local Road Research Board, MN/RC-2006-09, March 2006. As of June 6, 2010:
http://www.lrrb.org/detail.aspx?productid=2030

Swiss, James E., "Adapting Total Quality Management (TQM) to Government," *Public Administration Review*, Vol. 52, No. 4, July–August 1992, pp. 356–362.

Taylor, Brian D., "The Geography of Urban Transportation Finance," in Susan Hanson and Genevieve Giuliano, eds., *The Geography of Urban Transportation*, 3rd ed., New York: Guilford Press, 2004, pp. 294–331.

Taylor, James, Brian Stecher, J. O'Day, S. Naftel, and K. C. LeFloch, *State and Local Implementation of the No Child Left Behind Act*, Vol. IX: *Accountability Under NCLB: Final Report*, Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, Policy and Program Studies Services, forthcoming.

Teleki, Stephanie S., Cheryl L. Damberg, Chau Pham, and Sandra H. Berry, "Will Financial Incentives Stimulate Quality Improvement? Reactions from Frontline Physicians," *American Journal of Medical Quality*, Vol. 21, No. 6, 2006, pp. 367–374.

Thistlethwaite, Donald L., and Donald T. Campbell, "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, Vol. 51, No. 6, December 1960, pp. 309–317.

Thompson, Frank J., "Reinvention in the States: Ripple or Tide?" *Public Administration Review*, Vol. 62, No. 3, May–June 2002, pp. 362–367.

Thornburg, Kathy R., Wayne A. Mayfield, Jacqueline S. Hawks, and Kathryn L. Fuger, *The Missouri Quality Rating System School Readiness Study*, Columbia, Mo.: Center for Family Policy and Research, October 2009. As of June 6, 2010: http://cfpr.missouri.edu/MOQRSreport.pdf

Trochim, William M. K., "The Nonequivalent Groups Design," *Research Methods Knowledge Base*, last revised October 20, 2006a. As of June 14, 2010: http://www.socialresearchmethods.net/kb/quasnegd.php

———, "The Regression-Discontinuity Design," *Research Methods Knowledge Base*, last revised October 20, 2006b. As of June 14, 2010: http://www.socialresearchmethods.net/kb/quasird.htm

U.S. Code, Title 23, Highways.

U.S. Environmental Protection Agency, "2008 Program Implementation Summary: National Environmental Performance Partnership System (NEPPS)," briefing, Washington, D.C.: Office of Congressional and Intergovernmental Relations, undated.

———, "The Green Book Nonattainment Areas for Criteria Pollutants," last updated June 16, 2010. As of June 22, 2010: http://epa.gov/airquality/greenbk/

Viscusi, W. Kip, and Ted Gayer, "Safety at Any Price?" *Regulation*, Vol. 25, No. 3, 2002, pp. 54–63.

Wang, Xiaohu, "Assessing Administrative Accountability: Results from a National Survey," *American Review of Public Administration*, Vol. 32, No. 3, 2002, pp. 350–370.

Weick, Karl E., "Educational Organizations as Loosely Coupled Systems," *Administrative Science Quarterly*, Vol. 21, No. 1, March 1976, pp. 1–9.

Weikart, David P., J. T. Bond, and J. T. McNeil, *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results Through Fourth Grade*, Ypsilanti, Mich.: High/Scope Educational Research Foundation, 1978.

Welch, Jack, with John A. Byrne, *Jack: Straight from the Gut*, New York: Warner Books, 2001.

Wheeler, J. R. C., B. White, S. Rauscher, T. A. Nahra, K. L. Reiter, K. M. Curtin, and C. L. Damberg, "Pay-for-Performance as a Method to Establish the Business Case for Quality," *Journal of Health Care Finance*, Vol. 33, No. 4, June 2007, pp. 17–31.

Wholey, Joseph S., *Evaluation: Promise and Performance*, Washington, D.C.: Urban Institute, 1979.

Wholey, Joseph S., Harry P. Hatry, and Kathryn E. Newcomer, eds., *Handbook of Practical Program Evaluation*, 2nd ed., San Francisco, Calif.: Jossey-Bass, 2004.

Williams, Daniel W., "Measuring Government in the Early Twentieth Century," *Public Administration Review*, Vol. 63, No. 6, November–December 2003, pp. 643–659.

Winters, Marcus A., Gary W. Ritter, Joshua H. Barnett, and Jay P. Greene, *An Evaluation of Teacher Performance Pay in Arkansas*, Fayetteville, Ark.: University of Arkansas, Department of Education Reform, March 2007. As of June 6, 2010: http://www.heartland.org/policybot/results/20771/An_Evaluation_of_Teacher_Performance_Pay_in_Arkansas.html

Womack, James P., and Daniel T. Jones, *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*, New York: Free Press, 2003.

Womack, James P., Daniel T. Jones, and Daniel Roos, *The Machine That Changed the World: How Japan's Secret Weapon in the Global Auto Wars Will Revolutionize Western Industry*, New York: HarperPerennial, 1991.

Zellman, Gail L., personal communication with the authors, 2010.

Zellman, Gail L., and Anne S. Johansen, "Investment or Overkill: Should Military Child Development Centers Be Accredited?" *Armed Forces and Society*, Vol. 23, No. 2, 1996, pp. 249–268.

Zellman, Gail L., and Michal Perlman, *Child-Care Quality Rating and Improvement Systems in Five Pioneer States: Implementation Issues and Lessons Learned*, Santa Monica, Calif.: RAND Corporation, MG-795-AECF/SPF/UWA, 2008. As of June 6, 2010: http://www.rand.org/pubs/monographs/MG795/

Zellman, Gail L., Michal Perlman, Vi-Nhuan Le, and Claude Messan Setodji, *Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child-Care Quality*, Santa Monica, Calif.: RAND Corporation, MG-650-QEL, 2008. As of June 6, 2010: http://www.rand.org/pubs/monographs/MG650/

Zippay, Allison, and Anu Rangarajan, "Child Care 'Packaging' Among TANF Recipients: Implications for Social Work," *Child and Adolescent Social Work Journal*, Vol. 24, No. 2, April 2007, pp. 153–172.