1

The Evidence-Building Mandate

Do not copy, post, or distribute

Since the 1990s, numerous laws and executive actions calling for increased government accountability, learning, and improved performance in the United States have transformed how federal, tribal, state, and local governments collect data and employ evaluation. These activities, which in some cases have emerged as entire systems, generally emphasize the acts of goal-setting, measurement, and reporting. And since the early twenty-first century, efforts to promote “evidence-based policy” have been voiced from both inside and outside of government to steer public resources based on the results of program evaluations.

While the enthusiasm for evidence-based policy in concept is high, the production and use of evidence to inform decisions in federal, tribal, state, and local governments is uneven. In the United States, the implementation of performance measurement, evaluation, statistics, and policy research activities is often unintegrated and lacks a coordinated approach to support broader evidence-building goals and potential uses (NAPA 2020). Despite advances since 2016 from major federal initiatives and recommendations from the US Commission on Evidence-Based Policymaking to align existing resources to stregthen evaluation capacity, major gaps remain across the federal government.

Evaluation and performance activities are often siloed with little coordination within or across agencies, even as policymakers increasingly place expectations that more metrics or studies be produced and made transparent for government accountability purposes (Newcomer and Brass 2016). The various measurement activities are historically viewed by policymakers, managers, and agency staff as separate enterprises, with little interaction or integration. Performance measurement and reporting historically also does not benefit from evaluators’ or statistical agencies’

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

expertise, nor does it generally feed into evaluation initiatives. Despite the proliferation of data collection and measurement activities, in many forms and across disciplines, comparatively little emphasis has been placed from policymakers and program managers to use this information to improve the work of government and to expand the conception of use in complex, nonlinear decision-making processes.

Theories used to support efforts for improving capacity within government to use evaluation—such as accountability theory, organizational learning theory, and institutionalism theory—are greatly disconnected from modern practice. In reality, no single, recognized theory appears capable of adequately capturing the realities of modern government implementation nor the complexities of bureaucratic structures in the information age.

With some notable exceptions, including new efforts launched since 2017, major gaps exist for organizational- and individual-level capacity to produce and use evidence in the United States. Factors affecting capacity at the organizational level include integration of a coherent legal framework, adequate resources to produce and disseminate evidence, coordination across the evidence ecosystem, motivated leadership to support and sustain initiatives, a workforce to effectively execute, and an organizational culture that embodies the assimilation of an analogous evidence culture. At the individual level, capacity factors include personnel with the knowledge, skills, and attitudes consistent with principles, for example, proffered by the American Evaluation Association.

Despite calls for evidence to inform public policymaking and program management over the course of decades, different providers and different audiences have varying ideas about how to assess the quality and credibility of evidence. Nongovernmental organizations, such as the Cochrane Collaboration and the Campbell Collaboration, offer "evidence-based models" and draw attention to the need to learn from evaluation to inform public policy design and implementation. Promoters of scaling up evidence-based models have drawn attention to the need for rigorous designs. The federal government and other funders have embraced the notion of rigorous standards for evidence, as required in some federal programs by Congress and encouraged in previous guidance from the White House Office of Management and Budget (OMB), but there is still not a standard approach for developing and applying evaluation standards across government and related stakeholders.

Congress has highlighted the need for rigorous evaluation to support program management through legislative language inserted in a variety of laws. The Department of Health and Human Services' (HHS) Administration for Children and Families (ACF), for example, was required to create a clearinghouse of research and evaluation relevant for recipients of benefits from the Temporary Assistance for Needy Families Program to return to work. HHS opted to establish an advisory council to support design and

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

implementation of the resource in addition to soliciting public comments on the very design of the evidence standards used in the clearinghouse. Launched in 2020, the Clearinghouse of Proven and Promising Approaches to Move Welfare Recipients to Work was designed to address a range of potential system users.

To foster more consistency in evaluation work, in the Foundations for Evidence-Based Policymaking Act of 2018 Congress required OMB to issue Program Evaluation Standards for use across the federal government. After conducting an inclusive deliberative process, OMB issued evaluation standards in March 2020. The official federal evaluation standards are Relevance and Utility; Rigor; Independence and Objectivity; Transparency; and Ethics (OMB 2020). While these federal standards closely align with expectations and practices long-suggested by professional societies, such as the American Evaluation Association, many agencies are still developing the infrastructure and strategies to align the practices with expectations.

The environment in which public servants, contractors, and grantees conducting evaluations for government operate has become more complex, and even intimidating, amidst public dialogue about the need for more rigor, and strong evidence. Policymakers and commissioners of evaluations may interpret standards differently, and/or weigh the multiple criteria differently when judging the quality of study findings.

The myriad of policies and requirements shaping the supply of data and evaluations generated through a variety of legislative and executive requirements are difficult to distill and integrate. Public calls for evidence-based policymaking and data-driven decision-making are ubiquitous, but the signaling can be conflicting and overwhelming. Public managers and evaluators require a roadmap to navigate the evolving evidence ecosystem.

In this book, we provide just such a roadmap for evaluators doing business within, for, or in partnership with government, and for public servants who are expected to assess and use evidence generated by a large variety of evaluation approaches. "Evidence" has become a frequently used term in public policy discourse in the twenty-first century. We define the term here as data that are used with appropriate analysis to support a claim being made. Typically, the claims describe conditions targeted or affected by public policies and programs, and the data used may include administrative or survey data to support performance management or evaluation activities.

In this chapter, we first chronicle the history and current expectations of the "evidence-based policy" imperative. Then we describe the public sector's environment regarding the supply and demand for evidence, starting with the federal government's current legislative and executive requirements, and then a review of selected and relevant efforts made at the state and local governmental level. We conclude with clarifying what we mean by evidence-building capacity in government, and describe what the chapters to follow contain.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The Progression From Outputs to "Evidence-Based Policy"

Looking back over the last century in the United States, there has been a progression in strategies for framing and assessing the worth of government. The operative terms used within government and its observers have changed across time each building upon its successors, from outputs (1930s), to effectiveness (1960s), to outcomes (1980s), to results (1990s), to evidence-based policy (post 2000). Figure 1.1 chronicles the key events that inspired the use of these terms over time. Despite changes in the terms used, the focus has remained on measuring outputs, and, if possible, outcomes, and applying the information to support decision-making.

For over a century, local governments in the United States have been measuring tangible services or products (e.g., potholes filled or 911 calls received), that show the value of the local government to the residents. During the early 1900s, city governments were measuring what Clarence Ridley and Herbert Simon called outputs, and they commented on the inadequacy of mere output measures for informing administrative or policy decisions: "We can measure the miles of beat patrolled, the number of criminals apprehended, the number of finger-prints taken. But units such as these, however useful they may be, are not entirely adequate for our purposes. They tell us how much work has been done; but they do not tell how well it was done, nor whether the particular work undertaken was appropriate to the desired end" (Ridley and Simon 1938, p. 2). Many local governments have been measuring outputs, along with many other aspects of service delivery, such as efficiency, for decades.

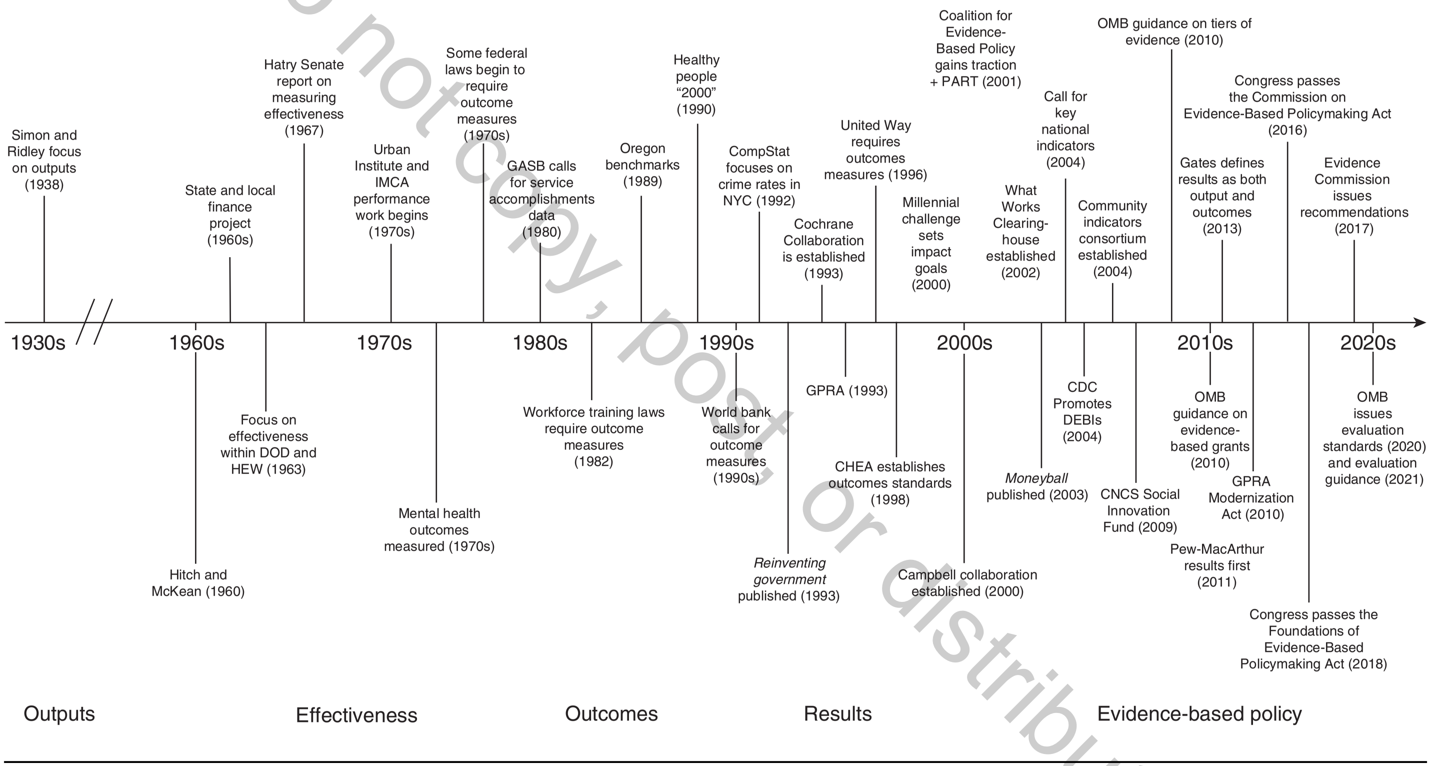
The term effectiveness as the desired value provided to the public came into use with the federal effort to link the "effectiveness" of programs to budget categories in the US Department of Defense (DoD) in 1961, a system that later became known as the Planning, Programming and Budgeting (PPB) System. When Robert McNamara took office as Secretary of Defense in 1961 he sought to evaluate military needs and use data to inform decisions on how to best meet those needs.

In an extremely impactful move, McNamara hired Charles Hitch as his Comptroller (Enthoven and Smith 1971, p. 33). At the time Hitch led a group of analysts at RAND, and was viewed as one of the leading authorities in the nation on program budgeting and the applications of economic analysis to defense problems (Enthoven and Smith 1971, p. 33). Along with Roland McKean, Hitch had published a book promoting systematic thinking and quantitative techniques, *The Economics of Defense in the Nuclear Age*, that had captured the attention of McNamara, among others (1960). Hitch set up a unit under him at DoD that focused on measuring "effectiveness" of policies and programs, and his staff included both Elmer Stats (later the head of what is now known as the Government

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

FIGURE 1.1 ● From Outputs to Evidence: Influential Events Across the Years


1930s
1960s
1970s
1980s
1990s
2000s
2010s
2020s
Simon and Ridley focus on outputs (1938)
State and local finance project (1960s)
Hatry Senate report on measuring effectiveness (1967)
Urban Institute and IMCA performance work begins (1970s)
Some federal laws begin to require outcome measures (1970s)
GASB calls for service accomplishments data (1980)
Oregon benchmarks (1989)
Healthy people "2000" (1990)
CompStat focuses on crime rates in NYC (1992)
Cochrane Collaboration is established (1993)
United Way requires outcomes measures (1996)
Millennial challenge sets impact goals (2000)
Coalition for Evidence-Based Policy gains traction + PART (2001)
What Works Clearing-house established (2002)
Call for key national indicators (2004)
Community indicators consortium established (2004)
OMB guidance on tiers of evidence (2010)
Gates defines results as both output and outcomes (2013)
Congress passes the Commission on Evidence-Based Policymaking Act (2016)
Evidence Commission issues recommendations (2017)
Hitch and McKean (1960)
Focus on effectiveness within DOD and HEW (1963)
Mental health outcomes measured (1970s)
Workforce training laws require outcome measures (1982)
World bank calls for outcome measures (1990s)
Reinventing government published (1993)
GPRA (1993)
CHEA establishes outcomes standards (1998)
Campbell collaboration established (2000)
Moneyball published (2003)
CDC Promotes DEBIs (2004)
CNCS Social Innovation Fund (2009)
OMB guidance on evidence-based grants (2010)
GPRA Modernization Act (2010)
Pew-MacArthur results first (2011)
Congress passes the Foundations of Evidence-Based Policymaking Act (2018)
OMB issues evaluation standards (2020) and evaluation guidance (2021)
Outputs
Effectiveness
Outcomes
Results
Evidence-based policy

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Accountability Office) and Harry Hatry—two pioneers in thinking about assessment of government performance.

An important legacy resulting from these initial PBB efforts was to spread the goal of measuring program effectiveness at the state and local levels of government (Hatry Interview, January 17, 2014). A pioneer study, called the 5-5-5 Study, was undertaken by a State and Local Finance project located at the George Washington University and funded by a Ford Foundation grant in 1968, to test the feasibility of implementing the PPB system in five states, five counties, and five cities. The grant money was used primarily to introduce PPB concepts, such as multiyear planning, to the 15 jurisdictions on the principles of PPB, and especially on how to develop criteria for assessing program effectiveness. In June 1969, the project's findings were published in a report, *Implementing PPB in State, City and County: A Report on the 5-5-5 Project* that was produced by Selma Muskin and her group of researchers—that included Harry Hatry. The quest to measure program effectiveness (the term used then, not outcomes) at the local level was pursued by both Mushkin and Hatry when they joined the newly established Urban Institute in the late 1960s, and Hatry's team at the Urban Institute shaped public discourse about how to measure the value of local governmental performance.

While still working on the 5-5-5 project in the mid-1960s, Hatry and his colleagues wrestled with the term "effectiveness" that they had inherited from PPBS since it seemed to imply causation. Hatry and this team sought to move away from using performance information for claims of causal impact of local government services. On July 21, 1967, Hatry presented a report to the Senate Subcommittee on Intergovernmental Relations of the Committee on Government Operations (90th Congress) entitled "Criteria for Evaluation in Planning State and Local Programs," which expressed the notion that attributing effectiveness was difficult.

Building on experience with PPB at the DoD, programs across the federal Department of Health, Education, and Welfare launched extensive efforts to understand outcomes in the 1960s and 1970s as part of the Great Society initiatives. New authorities, and even new federal agencies, sought to articulate and inform evolving policy approaches and solutions as welfare, social security, health, employment, disability, and environmental initiatives grew in scope and scale. The Urban Institute, along with a growing cadre of other analytical organizations and management consultancies, worked alongside government partners to study the effects of these programs. In many instances, Congress required these programs to conduct demonstration or pilot projects, paired with evaluations to ensure the causal outcomes claimed could be attributed to the project.

But in the 1970s, the term effectiveness, not outcomes, was still the operative term across the country. This was especially the case as some local governments (and their budget offices) moved to measure efforts and accomplishments with suggestions from the Urban Institute. Similarly, state

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

governments were working to implement the Great Society initiatives, which included requirements for performance indicators on new block grants for funding antipoverty programs that appealed to state officials to demonstrate effectiveness in implementation.

The use of the term *outcomes* became more common in the late 1970s, in part framed by work from the Urban Institute on mental health and other antipoverty programs (Schainblatt 1977). Work by staff at the Urban Institute and other contractors on completion of large, expensive, and national-scale program evaluations was intended to improve policymakers' understanding of the outcomes associated with antipoverty programs created under the Great Society. Many of these projects focused on work incentives and employment supports, and even relied on state capacity for implementing federal programs to improve the ability to compare outcomes across different policy strategies.

During the 1980s, performance measurement efforts housed in local and state budget offices were more likely to be embraced among certain groups, such as mental health professionals, who already had been collecting data from clients on their health outcomes. In addition, an interesting development was taking place somewhat independently, though not in a totally disconnected manner, in government accounting circles. Government accountants had traditionally examined the operations of government agencies through following financial transactions. Then in December 1980, the Financial Accounting Standards Board (FASB) published a *Statement of Financial Accounting Standards Concept No. 4 on Objectives of Financial Reporting by Nonbusiness Organizations* (1980).

The FASB concept paper first introduced the idea of measuring program accomplishments (read outcomes). In the statement FASB noted that: "Periodic measurement of the changes in the amount and nature of the net resources of a nonbusiness organization and information about the service efforts and accomplishments of an organization together represent the information most useful in assessing its performance" (p. xiv). And the statement went on to say "Ideally, financial reporting also should provide information about the service accomplishments of a nonbusiness organizations. Information about service accomplishments in terms of goods or services produced (outputs) and of program results may enhance significantly the value of information provided about service efforts. However, the ability to measure service accomplishments, particularly program results, is generally underdeveloped" (FASB 1980, pp. 25–26).

Eventually the Government Accounting Standards Board developed a series of reports on the need to report nonfinancial outcomes, such as its 1990 publication "Service Efforts and Accomplishments Reporting: Its Time Has Come." While "Accomplishments" is used in the title, the text when referring to such indicators used the work "outcome" for them. Since about 1990, many governmental accounting professionals have continued to highlight the need for more reporting of nonfinancial accomplishments,

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

and virtually every annual conference of the Association of Government Accountants has featured guidance on how to address the challenges of measuring accomplishments (outcomes).

In the 1990s, the Urban Institute worked with the International City Management Association (ICMA) to develop the ICMA Comparative Performance Measurement project. Based on input from staff in local government agencies, the project identified a number of output and outcome indicators for some basic services for all participating cities to report to ICMA. Participants could obtain information on how their government's performance on these indicators compared with other similarly sized local governments, though the names of the other local governments were not revealed.

During the time that the ICMA Project was gaining ground, and the book *Reinventing Government* by David Osborne and Ted Gaebler with its focus on outcomes and results was attracting attention across the country, the Urban Institute had also adopted the term "results," rather the "effectiveness" (Osborne and Gaebler 1992). By the time Harry Hatry and his team at the Urban Institute published a guidebook on measuring outcomes, *Performance Measurement: Getting Results, first edition* in 1999, the term "outcome" was ubiquitous at all levels of government, in part due to some other influential events.

Another performance-focused phenomenon that started in the 1990s at the local level of government also had ripple effects across local governments first, and then upward to some states and, finally, even the federal level. In 1994, Police Commissioner William Bratton introduced a data-driven management model in the New York City Police Department called CompStat, which involved the Chief holding weekly meetings with district commanders to review crime rates across the districts, and discuss changes and tactics to address rate increases. The use of the managerial tool was subsequently credited by its supporters with decreasing crime and increasing quality of life in New York City over the two decades (Kelling and Bratton 1998; Shane 2007). The CompStat model was adopted by other cities across the United States, and Bratton became a management consultant in much demand across the world. By 2000, over a third of police departments in the United States with 100 or more officers reported that they had implemented a "CompStat-like" program (Weisburd et al. 2003). Maryland's Governor Martin O'Malley used the model to create CitiStat when he was Mayor of Baltimore, and he implemented a state-level version when he became governor of Maryland in 2006, where he expanded his CitiStat model to review data on outputs (and some outcomes) to manage multiple government agencies (Fillichio 2005; Fenton 2007). The basic frameworks underlying CompStat, then CitiStat, and then StateStat models for data-informed reviews were adopted across the United States, especially in local governments. The model was adopted by the Barack Obama Administration for quarterly data reviews for federal agencies beginning in

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

2010, and the practice was added to requirements in the Government Performance and Results Modernization Act of 2010. Some federal agencies have adapted the model with some success.

Requirements for agencies to share data to demonstrate the worth of public services have been established through both executive action and legislation at the federal level of the US government. As noted earlier, the lead federal budget agency, now titled the Office of Management and Budget (OMB), started promoting measurement of effectiveness during the Lyndon Johnson administration as a result of the perceived positive effects of implementing a form of PPB in the DoD. After Johnson was defeated, the Richard Nixon and Gerald Ford Administrations turned more toward promoting internal process improvements, e.g., “Managing by Objectives,” and the Jimmy Carter Administration initiated efforts to introduce zero-based budgeting, but neither of these efforts required federal agencies to collect data and report the data publicly.

While some OMB program examiners had been routinely asking for evidence about how well programs were working as part of their examination of agency budget requests for some time, OMB officially started requiring agencies to include output measures in federal agency budget requests in 1992 under the leadership of OMB director Alice Rivlin. A call for measurement of program results also accompanied the executive initiative launched by President Clinton in March 23, 1993, and coordinated by Vice President Gore, commonly known as the National Performance Review (NPR), and then renamed the National Partnership for Reinvention. The NPR emphasized results-oriented management as well as reforms such as cutting “red tape” and outsourcing, and some of the language used was much in line with the principles espoused in the then very popular book, *Reinventing Government* by David Osborne and Ted Gaebler (Fox 1996; Breul and Kamensky 2008; and Moynihan 2008). Along with their claims that public leaders and managers should steer rather than row, not be rule-driven, and move from hierarchy to participation and teamwork, Osborne and Gaebler pushed for measuring outcomes and results rather than simply inputs.

The NPR effort was organized from the Clinton White House. The leaders recruited about 250 temporary staff on assignment from other executive agencies, and the teams reviewed the agencies and systems to which they had been assigned. In addition, Vice President Gore reached out to the public and the broader federal career workforce for stories of what was wrong and what needed to be fixed. Gore received more than 50,000 letters, and he went to listening sessions in several dozen federal agencies and around the country. The NPR teams developed findings and recommendations, and David Osborne was brought in to help craft the final report. The final report, *From Red Tape to Results: Creating a Government That Works Better and Costs Less* (Gore 1993), contained more than 1,200 recommendations (Breul and Kamensky 2008).

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Some NPR recommendations focused on improvements to the way the government worked, such as streamlining procurement and setting customer service standards, and other recommendations called for cutting the size of the federal government. NPR actions led to a decrease in the number of federal employees by more than 250,000, and elimination of a range of seemingly ineffective programs such as the wool and mohair subsidy (Breul and Kamensky 2008). The first set of recommendations appealed to government employees, while the second set was targeted to appeal to the general public. Critics have been quick to note that the most consequential result of the NPR was the reduction of the federal workforce, a move that led to a huge increase in the number of federal contractors in subsequent years with a whole new set of governance challenges (Fox 1996; Kettl 2005; and Light 2008).

The George W. Bush Administration directives to agencies stepped up the pressure to align programmatic performance goals and performance data with budget categories. New initiatives were included in Bush's President's Management Agenda (PMA). The PMA was similar to the NPR in intent, but focused its efforts on a limited number of areas—strategic management of human capital, competitive sourcing, financial performance, electronic government, and budget and performance integration. To insert performance data into the President's Budget, OMB employed the Program Assessment Rating Tool (PART) that assessed program performance and assigned scores to each of about 1,000 federal programs. President Bush also established agency performance improvement officers and OMB led meetings of these officers in a new Performance Improvement Council.

The legislative call for the provision of nonfinancial program performance and results data in agency financial statements was first established by the Chief Financial Officers Act of 1990. These reporting requirements were then expanded in the Government Management Reform Act of 1994. Perhaps the most important piece of legislation related to performance in the 1990s was the Government Performance and Results Act of 1993 (GPRA) that required most federal agencies to have strategic plans, performance goals, and performance reporting. The inclusion of the term "results" in the title of the law reflected the public dialogue inspired by the Osborne and Gaebler bestseller, as well as many other advocates of New Public Management reforms that included calls for managing by results, i.e., outcomes. Since the enactment of GPRA, dozens of federal laws have been passed that require agencies to measure performance in specific policy arenas, and GPRA reporting requirements were strengthened with the GPRA Modernization Act of 2010 (GAO 2008a, 2008b, 2013).

GPRA was the prime mover in requiring federal managers to develop performance measures, with direction and guidance from OMB staff who played a critical role in interpreting the law's requirements. OMB did not, however, provide stringent guidance or offer significant technical assistance

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

to agencies, leading representatives from some agencies to voice uncertainty about what measures would be perceived as appropriate. Federal managers were not convinced they would be granted sufficient authority or flexibility to change their way of doing business. Securing agreement among diverse stakeholders and service delivery partners on what to measure was difficult. Understandably, outcomes of government services or activities often were viewed as beyond the control of agency staff. More importantly, there was uncertainty about how performance data would be used in budgeting, and only spotty evidence existed that performance measures had much effect on congressional funding decisions.

OMB program examiners assumed more overt performance measurement oversight responsibilities through their role in reviewing and approving scores submitted by agencies with the PART tool during the Bush Administration, a tool which included information on program outcomes and goal attainment. Agencies submitted binders of materials for each aspect of the tool to back up claims and assertions. Disagreements between agency staff and OMB examiners were often elevated above the frontline examiners. Differences between OMB and agency staff regarding the appropriateness of measures, including which measures should be assessed in PART, created additional tensions for performance management implementation (Newcomer and Redburn 2008). However, experience with PART, which was primarily viewed as a performance management tool, was instrumental in advancing interest in evaluation across government, and also led to an expected increase in the production of evaluations about federal programs (Hart and Newcomer 2018).

With Congress' passage of welfare reform in 1996, that greatly modified federal policies on cash transfer social programs, discourse about measuring outcomes grew increasingly to focus on the need for more large-scale impact evaluations of the social policies. The renewed congressional interest in evaluations principally occurred regarding programs operated by the Departments of Labor, Education, parts of HHS, and the Social Security Administration. But even within those agencies, evaluation efforts were not always sufficiently resourced or supported.

Building on the experience the Bush administration had with PART, and the renewed interest with evaluating impact, the Obama Administration simultaneously advanced a performance framework alongside an evaluation focus. A key component of the Obama Administration's approach to the collection of performance data, or what they deemed their performance management framework, was to require all major Federal agencies to identify a limited number of high-priority performance goals reflecting the near-term implementation priorities of each agency's senior managers. The mandate given from the Obama performance team was to: "use goals to improve performance and accountability, measure and analyze performance to find what works, and deliver better results using frequent, data-driven reviews."

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Many, though not all, of the Obama Administration's high-priority performance goals were expressed as desired outcomes, such as one of their first and most publicized:

> Veterans' Homelessness: The Department of Housing and Urban Development and the Department of Veterans Affairs will jointly reduce homelessness among veterans. Together, the two agencies will reduce the number of homeless veterans to 59,000 in June, 2012. (OMB 2011c)

However, some of the high priority goals were focused on outputs, and few were supported with new funding. The need to direct more attention to using evidence and evaluation was even included in one high priority goal for the Department of Education:

> Evidence Based Policy: Measuring Effectiveness and Investing in What Works: Implementation of a comprehensive approach to using evidence to inform the Department's policies and major initiatives, including: Increase by 2/3 the number of Department discretionary programs that use evaluation, performance measures and other program data for continuous improvement; Implement rigorous evaluations for all of the Department's highest priority programs and initiatives; and Ensure all newly authorized Department discretionary programs include a rigorous evaluation component. (OMB 2011c)

During the first term of the Obama Administration both the terms "results" and "evidence" were used when referring to the ongoing performance data being collected (see Hatry and Davies 2011; CRS 2012; GAO 2013). Both the use of high priority goals and quarterly reviews that were originally Obama OMB initiatives were incorporated in the GPRA Modernization Act of 2010.

Obama's OMB also promoted the use of data analytics, that is sophisticated analyses of performance data and administrative data, to inform decision-making, reflecting the widespread popularity of the use of data analytics in Michael Lewis' bestselling book *Moneyball* (2003). While not specifically integrated with the performance or evaluation activities, in the Obama Administration's first term the OMB issued guidance to agencies promoting evaluation and even hosted a competition among agencies to award new funding in the annual budget for certain evaluation activities, although the new funding was never distributed. In 2013, OMB called on agencies to increasingly produce evaluations to support budgeting decisions. Despite the enthusiasm for evaluation projects expressed at the time, there is little evidence to suggest the initial guidance and early calls from OMB substantially improved the evaluation infrastructure or capacity in government.

One activity that gained more traction in the Obama Administration was the encouragement of rapid cycle evaluation, which increased in 2015 in

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

the spirit of *Moneyball*. A bipartisan group of advocates of using data analytics in government published *Moneyball for Government* (2014), translating Michael Lewis' work into more government-salient efforts. Proponents called upon agencies to not only view evaluation as entailing large-scale, national, and costly projects, but rather to use existing data collection small-scale initiatives to test out how innovative practices could yield big changes in how government operates. The rapid-cycle evaluation approaches were, helpfully, not tied to budget reporting which enabled a greater focus on a range of evaluative approaches for administrative and operational issues. In addition, the Social and Behavioral Sciences Team (now the Office of Evaluation Sciences at GSA) offered technical assistance and support for projects, largely staffed by academics from outside government who offered methodological expertise to design and implement internal, often confidential, evaluation projects in federal agencies.

The Current Environment Regarding the Demand for Evidence

A confluence of influential events at the turn of the century heralded an increased public enchantment with the term "evidence-based," imagined as an idealized approach for translating a body of research evidence into decision-making processes and actions. In 2000, North American and European social scientists established the Campbell Collaboration to share evidence-based models in the social policy arena much like the Cochrane Collaboration had done for medical interventions. Both models relied on approaches for conducting systematic reviews of multiple studies on similar issues and programs to distill and translate the meaning across studies.

In the United States, more attention and focus was placed over the same period in expanding the capabilities for individual studies, often requested by policymakers in legislative requirements. The Coalition for Evidence-Based Policy became a proponent of using rigorous evidence of efficacy obtained from random control trials starting in 2001, building on the expectations of some congressional and OMB staff that human services and employment programs evaluated at the national level should specifically rely on this approach. The efforts of the Coalition encouraged an uptick in the availability of experimental evaluations for a small number of large social services and human services programs. As the availability increased, so too did the recognition that in the United States improved mechanisms were needed to ensure decision-makers and practitioners could easily access relevant information to apply in their specific contexts. The What Works Clearinghouse at the US Department of Education in 2002 was one of the initial publicized commitments made jointly by social scientists and by government to advance the systematic collection and analysis of research to inform decision-making in the public sector (GAO 2009). Since the launch

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

of the What Works Clearinghouse, numerous others have emerged from the US federal government related to fields including employment services and criminal justice.

Collectively these efforts attempt to bridge the supply and demand of evaluation, ensuring that once evaluations are completed and available that users can access relevant information to support decision-making in the appropriate context. In fact, this is the idealized model of evidence-based policy and practice, where research and evaluation are used responsibly to inform decisions about implementation of policies, programs, and activities. The model contains two conceptual foci: evidence building and use. The first is the group of activities where researchers and evaluators conduct methodologically rigorous evaluations and studies. Then, these researchers and evaluators produce study deliverables that clearly explain the research methods, key findings, and policy implications. At this stage, the researchers and evaluators can serve as trusted intermediaries to share and disseminate the findings and any recommendations, or this can also be conducted through knowledge brokers that serve as third-party intermediaries. This brokering phase in the decision model is when evidence building and use can occur with some overlap, perhaps even leading to iterative dialogues about what evidence is needed or can be improved upon to support decision-making needs.

When it comes to evidence use, policymakers need to assess the quality of the evidence and understand the policy implications of the findings. While third-party brokers can facilitate this step, policymakers and their staff can also fulfill this role independently. Then, these educated policymakers review the policy-oriented findings and recommendations, and enact policies or programs as per the findings and recommendations. Importantly, the use of evidence comes in many forms. Direct uses occur when the evidence from a study of collection of studies suggests an intervention produces intended results, with "evidence-based interventions" available for adoption by policymakers that are then applied in practice. Other types of uses can be more indirect, leading to requests for changes in funding or may even inform dialogues about the nature of a problem and the options available for solving them (Hart and Yohannes 2019). In any case, the goal of evidence use is to improve a desired outcome, which can then itself be reviewed and evaluated to provide feedback for future evidence building.

This vision of evidence-based policymaking is one that has become increasingly popular across the United States in recent years. In addition to the federal government taking up the evidence-based policy goal, leading foundations in the United States and funders of international development across the world have been investing resources to support and disseminate evidence-based practices. For example, in 2001, the Pew-MacArthur Results First Initiative was started to promote the use of evidence by US state governments. Calls for evidence of impact in the international development arena were increasingly voiced after 2001, with

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

the publicly funded Millennium Challenge Corporation in the United States established in 2004 that embodies an evidence-based decision-making framework for determining how to allocate project funding across the world. In the nonprofit space, a leader in developing a rigorous evidence base, the International Initiative for Impact Evaluation (3ie), was established in 2008 to fund impact evaluations in development work, reflecting the worldwide enchantment with the systematic collection of performance data and the assessment of impact to ultimately inform decision-making by both development funders and implementers. 3ie is a US nonprofit organization with an office in Washington, DC, and programs operating in Delhi and London under the auspices of the Global Development Network and London International Development Centre, respectively, and it is but one of many other organizations seeking evidence about "what works" in fostering development (Lipskey and Noonan 2009). JPAL North America similarly set out in 2003 to test antipoverty interventions around the world, and its founders were awarded the Nobel Prize in Economics in 2019, recognizing the expansive application of the evidence-based model had on policymaking and efforts to alleviate global poverty.

While the supply of evidence about public programs has grown substantially over the years, the alignment between the supply and immediate use has remained tenuous. The demonstration projects and evaluations of large antipoverty programs from the Great Society initiative and the 1990s welfare reform led to a proliferation of "evidence," which some interpreted to suggest the programs did not operate as effectively as they could, but findings did not inform policymaking (see Gueron and Rolston 2013). There have not been many examples of performance data informing budgetary decisions, even when the George W. Bush Administration made an attempt to apply evidence, both performance information and evaluations, to inform funding decisions and to hold programs accountable (Hart and Newcomer 2018).

Since the 2000s, the federal OMB has advocated for more rigorous evaluation work to supply strong evidence on the extent to which specific programs work, i.e., produce results. During the Obama Administration, multiple agencies worked in collaboration to develop a tiered evidence framework, intended to communicate which evaluation research designs are deemed more likely to produce valid information from evaluation studies and also how many of such studies are needed to reach convergence on the idea that a particular intervention is effective. HHS, DOL, Education, and the National Science Foundation collectively presented common guidelines for such a framework in 2013. Since that time, tiered evidence frameworks have also appeared in federal legislation. Notably, in 2018, the Family First Prevention Services Act required HHS to apply different thresholds of proven, promising, and ineffective tiers to making funding determinations, and to determine the amount of grant funding that could be allocated for child welfare programs across the country.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Without question, the Obama White House, including OMB, publicly voiced support (moral if not financial) for rigorous program evaluation more prominently than previous administrations. A series of memoranda from OMB between 2009 and 2013 signaled that performance measurement and evaluation were to be used to produce "evidence on what works" (OMB 2010a, 2010b, 2010c, 2010d, 2011a, 2011b, 2012, 2013). OMB established a cross-agency work group to coordinate federal evaluation policy matters, the Interagency Council on Evaluation Policy, that signaled to agencies that evidence-based programs were more likely to receive funding in the President's Budget Request; focused on improving access to data and linking of data across program and agencies; called for more collaborative evaluations both across agencies and across service providers in different sectors; established new expectations for evaluations and offered training on evaluation expectations to agency staff (Newcomer and Brass 2013; Hart and Newcomer 2018).

Federal agencies independently, or with OMB support, have launched an extensive set of evaluation efforts over the last decade. Prior to passage of the Evidence Act in 2018, several agencies established Chief Evaluation Officers, such as at the US Department of Labor and the Centers for Disease Control. Some agencies established their own evaluation standards, e.g., the ACF in HHS, USAID, and the CDC. And federal support was given to the National Academics of Science, Engineering, and Medicine's Committee on National Statistics to facilitate dialogue about evaluation standards and principles for the federal government.

The role of OMB for coordinating evidence and evaluation activities across government has been influential. For more than a decade, the federal OMB has directed agencies to consider available evidence in proposing annual budget requests. OMB signaling to agency staff has elevated the role of evidence and the emphasis on using evidence in policymaking. OMB included a chapter on the role of evidence in the President's Budget each year, from 2010 to 2019. Generally, there was even notable continuity in the language used by OMB across the Obama and Trump administrations (2010–2019), a reflection of the career OMB staff influence in promoting evidence use across administrations.

Another sign of interest in promoting the use of evidence was the widespread support for the US Commission on Evidence-Based Policymaking across the executive and legislative branches of the federal government. Established by law in 2016, during the Obama Administration, the Evidence Commission was charged with developing a national strategy for better using data already collected by government to generate insights relevant for decision-making within an eighteen-month period. While the commission's formal charge focused principally on data issues, the Evidence Commission's unanimous recommendations expanded greatly on the role rigorous evidence could and should have in informing policymaking. The very establishment of the commission by Congress was a signal about

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

demand for evidence from federal policymakers who recognized that many of the questions they were asking about important policies and programs could not always be answered, as the supply of the evidence was not adequate.

The Evidence Commission's final report, presented to Congress and the Trump Administration during its first year, offered twenty-two recommendations that wove together a strategy for improving data access, strengthening privacy protections, and promoting evaluation capacity. Of note, the recommendations explicitly encouraged federal agencies to establish senior leaders for data management, evaluation officers, and engage in planning processes for evidence building (i.e., learning agendas and assessments).

The Evidence Commission also recognized the breadth of types of evidence relevant for policymaking, ranging from descriptive statistics formulated as national economic indicators to random control trials as one form of evaluation. The commission explicitly called for the use of a portfolio of evidence: "The Congress and the President should provide sufficient and appropriate authority for departments to design programs and policies that enable a portfolio of evidence to support continuous learning and information needed to ensure accountability" (CEP 2017, p. 102). The message increasingly resonated with OMB, and was reflected in guidance documents OMB has provided to the agencies. The main message is that all agencies need a portfolio of evidence that includes a variety of data and studies, including the GPRA performance data, impact evaluations, implementation studies, and economic analyses that can assess the value that the federal programs and policies provide.

In 2017, shortly after the Evidence Commission issued its recommendations, legislation began weaving its way through Congress to implement half of the recommendations. The Foundations for Evidence-Based Policymaking Act of 2018 (P.L. 115-435) drastically shifted the expectations for federal agencies to develop and support infrastructure to engage in generating relevant evidence. Signed into law in January 2019, the Evidence Act requires the twenty-four largest federal agencies to establish evaluation officer positions and produce strategic plans for evidence building, requires every agency in the Executive Branch to have a chief data officer, and requires the publication of government-wide evaluation standards and practices alongside formal infrastructure to recognize evaluation as an occupation for federal employees. The Evidence Act also allows for improvements in the privacy framework and infrastructure used for sharing and linking confidential data used for building evidence. At the same time, the new law promotes efforts to improve transparency in government, specifically calling on agencies to vastly improve and expand open data efforts. Collectively, the Evidence Act's emphasis on producing evidence for government agencies more efficiently by enabling capacity and reducing barriers provides unique opportunities to align long-standing performance,

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

evaluation, and statistical activities to better coordinate across the evaluation ecosystem to generate the portfolio of evidence envisioned by the Evidence Commission. The role of the Evidence Act for government today has since been reinforced with continued emphasis on its implementation, and upon the use of evidence in decision-making by the Biden Administration (Biden 2021; OMB 2021).

Table 1.1 lists the key initiatives addressing the use of "evidence" in the last two decades in the United States, and clarifies their scope and impact on the production of data. When assessing the impact of the various laws, it should be noted that while agencies can document that there are data and studies produced, there have been no systematic assessment of their impact on informing decision-making, nor on improving government programs and policies. There are some promising cases studies of the use of evidence across a host of agencies, ranging beyond the scope of well-documented

TABLE 1.1 ● Current Forces Affecting the Supply and Demand for Evidence in Government

Authority	Scope	Impact
Government Performance and Results Act, and the GPRA Modernization Act of 2010	All federal executive agencies and grantees, e.g., state and local governments, tribal authorities, and funded nonprofit service providers	• Agencies are required to collect data on outputs, and some outcomes • Some grantees must use evidence-based models provided to them
US Commission on Evidence-Based Policymaking	Presented recommendations to the President and Congress that largely focused on the federal government	• The Commission outlined a clear vision and the benefits of using evidence in decision-making • The unanimous recommendations from the Evidence Commission provided a framework for improving data accessibility and use for evidence building • A portion of Evidence Commission's recommendations provided the basis for the Evidence Act

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Authority	Scope	Impact
Foundations for Evidence-Based Policymaking Act, or Evidence Act (including the OPEN Government Data Act and the Confidential Information Protection and Statistical Efficiency Act)	Different parts of the law apply to different agencies. While the open data and broad data mandates apply to all federal agencies, directives about capacity-building focus on the 24 major departments and agencies, and statistical data sharing authorities are limited to select agencies	• Under Evidence Act agencies are required to develop plans that involve collecting evidence on programs • Part of the Evidence Act focuses on establishing processes and leadership positions to bolster capacity for evidence production and use • Major components of the Evidence Act include directives to promote open data, secure data sharing, and accessibility of information for evidence building
Digital Accountability and Transparency Act, Congressional Budget Justification Transparency Act, Grant Reporting Efficiency and Agreements Transparency Act	All federal executive agencies	• Improves standardization of information about government spending • Includes expectations of transparency and openness about budget requests, spending summaries, and grant reports • Provides a model for sector- or topic-specific authorities that can encourage evidence building
Agency Specific Laws and Regulations for Topical Issues (e.g., EPA's Emergency Planning and Right to Know Act, Social Impact Partnership to Pay for Results Act)	Varies, individual agencies or policy issues	• Individual laws and authorities can outline specific deliverables and evidence needs required or expected by Congress

(Continued)

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.1 ● *(Continued)*

Authority	Scope	Impact
Federal Office of Management and Budget Guidance	Federal executive agencies and grantees, e.g., state and local governments, tribal authorities, nonprofit service providers	• OMB is providing phased guidance on how to implement Evidence Act provisions • OMB provided evaluation standards in 2020 that apply to the major departments and agencies, and their activities
State Laws Directing Funding to Evidence-Based Solutions	A select number of states and policy areas within those states	• Some grantees must use evidence-based models provided to them
Local Government Performance Measurement Systems	Some well-resourced local governments, e.g., NYC, Montgomery County, MD	• Agencies in the jurisdictions are required to collect data on outputs, efficiency, and some outcomes
Directives from Other Funders, e.g., World Bank, Pew Charitable Trusts, Bill and Melinda Gates Foundation, Arnold Ventures	All entities funded by that specific funder	• Some loans and grants require reporting of data on results and impact, with some guidance to use rigorous evaluation techniques, e.g., random-control trials

social service programs to include environmental policy, food safety, homelessness supports, and much more (Hart and Yohannes 2019). Notwithstanding the success stories, the US Government Accountability Office has reported generic uses of data and evaluations via surveys of federal managers since about 2001, and the results are not universally encouraging about the level of learning and improvements made due to the use of evidence (GAO 2018). How and when evidence informs internal learning and management within government presents a valuable target for research that some have only begun to explore, and we only scratch the surface of this puzzle in this book.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

What Does Evidence-Building Capacity Entail?

While evidence pertinent to government may be provided from many sources, we focus on the data generated and used for the express purpose of informing public sector decision-making. We view *evidence-building capacity (EBC)* within government as including both the supply and demand for evidence to inform deliberations about public policies. We define EBC as:

The motivation and infrastructure to do the following:

- develop *relevant questions* about an organization's programs and policies,
- collect and generate (or access if already collected by other agencies) *data* to address the questions, manage and protect data, analyze and interpret the data, and
- provide *relevant insights* from the data to inform management and stakeholders for policymaking.

By motivation we mean that someone in public organizations should be interested enough to ask questions about how public policies and programs are operating, and achieving desired results. Infrastructure refers to staff, data, data systems, and analytical capacity to collect, analyze, and interpret data to address relevant questions about policies and programs.

While we hope that motivation to ask for evidence translates to using the evidence, assuming it is of sufficient quality, to inform decision-making, we realize that use is not always guaranteed. Many criteria and values are brought into play in decision-making and policymaking within the public sector, and strong evidence about conditions or results is not likely to overcome political values. Political leaders tend to have short time frames and may impose policy mandates with little patience for learning from evidence, or from previous efforts to employ evidence-informed approaches.

Data are routinely collected by government agencies and their agents, e.g., grantees, to meet reporting requirements, but they may not be used to address questions that the managers or leaders care about. Reporting data in the exercise of showing accountability for resources tends not to enable nor encourage learning in a forward-looking sense. Time and staffing constraints mean that presenting data to demonstrate accountability upward tends to detract from time for managers to spend analyzing data and findings to learn how to improve programs. And there are typically insufficient incentives for program managers to spend time learning from data or evaluation studies.

Simply ensuring that there is adequate EBC within any public agency does not automatically ensure that public servants will take actions to improve policies and programs. But we assert that learning from evidence to

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

improve government is a desirable and worthy goal. Learning from evidence within government can be incentivized and facilitated. We talk more about that in this book.

Conclusion

In this introductory chapter, we chronicled the history of the "evidence-based policy" imperative for government, and we described the forces shaping the demand for evidence to depict the value and worth of government. We provided an overview of the public sector's environment regarding the supply and demand for evidence, starting with the federal government's current legislative and executive requirements, and a review of relevant efforts made at the state and local governmental level and by foundations. We concluded with clarifying what we mean by evidence-building capacity in government.

In the chapters that follow, we explore in more detail evidence building, starting with a deep dive into how the quality of evidence may be assessed in Chapter 2. Then in Chapter 3, we describe the value and practical uses of evaluative thinking for measurement and evaluation work in government. We describe the processes and benefits of developing an especially important evidence-building tool—learning agendas for government agencies—in Chapter 4. We explain the challenges and opportunities within government to first generate evidence in Chapter 5, and then to stimulate learning from evidence in Chapter 6. We conclude by providing a set of recommendations on how to sustain efforts to build and maintain evidence building capacity in government.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Exercises

1. You have been asked to brief a new political appointee working in the U.S. Department of Labor on what the following terms mean in the context of government: outputs, outcomes, effectiveness, and evidence-based policy. Please explain each term and offer examples.
2. Explain when and how two bestselling books, *Reinventing Government* by David Osborne and Ted Gaebler and *Moneyball* by Michel Lewis, affected thinking about how government should work in the federal level of government in the United States.

Resources for Additional Learning

Newcomer, Kathryn E., Harry P. Hatry, and Joseph S. Wholey. 2015. *Handbook of Practical Program Evaluation.* San Francisco, CA: Jossey-Bass.

Betterevaluation.org.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

Copyright ©2022 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.