PART 3

Government Analytics Using Administrative Data

CHAPTER 9

Creating Data Infrastructures for Government Analytics

Khuram Farooq and Galileu Kim

SUMMARY

Creating the conditions for effective data analytics in the public sector requires reforming management information systems (MIS). This chapter outlines a road map to guide practitioners in the development of analytically driven information systems, drawing on the experience of World Bank practitioners and government officials. The conceptual framework and cases presented focus on human resources management information systems (HRMIS), foundational data infrastructures that provide information on personnel and compensation. However, the chapter articulates broader lessons for designing and managing information systems for government analytics. The chapter first discusses the stages in reforming an HRMIS, outlining key decision points and trade-offs involved in building public sector data architectures. It then demonstrates how this framework can be applied in practice and the associated risks, drawing on case studies of analytical transformations of HRMIS in Luxembourg, Brazil, and the United States.

ANALYTICS IN PRACTICE

- Many government information systems, such as human resources management information systems (HRMIS), can be thought of in terms of the interaction of distinct data modules. Thinking about data within these "islands" of like data allows the analyst to define more precisely the reforms such modules might need and identify interdependencies between them. For example, an HRMIS can be thought of as multiple modules, including payroll and personnel data, that interact to provide insights any single type of data could not. As a result, reforms must be precise regarding which modules will be modified and why. The implementation team needs to be aware of how modules depend on one another in order to understand how the flow of data within the information system might disrupt the use and interpretation of other modules in the system.

Khuram Farooq is a senior financial management specialist in the World Bank's Governance Global Practice. Galileu Kim is a research analyst in the World Bank's Development Impact Evaluation (DIME) Department.

- HRMIS data modules are not created equal: rather, some provide the foundation upon which all other systems rely. Reforms should prioritize the comprehensiveness and quality of foundational modules in human resource management and then transition to developing more complex analytical modules. In HRMIS reforms, foundational modules—including payroll compliance and basic personnel information—should take precedence over other layers, such as talent management and analytics. Without a quality foundation, other modules will produce imprecise or inaccurate analytics. Analytical modules, including dashboards and reports, require that foundational modules be set in place and that their data be accurate.

- However, almost any government data system that has prioritized accurate measurement can be useful. Thus, small reforms targeted at strengthening basic data quality can generate analytical insights even when other foundational modules need further reform. Most developing countries are in the process of building foundational HRMIS modules, such as payroll or basic information on HR headcount. Accurate measurement of these data can be useful. Even if these foundational modules are incomplete in terms of the wider vision of the implementation team, it is still possible to develop analytics reports from foundational modules, such as wage bill analysis and sectorwise employment. Analytical reports can be produced, even if manually, without implementing an analytics module. Though data analysis and visualization might be narrow in scope, this approach can provide quicker results and build even greater political will for further reform.

- HRMIS reform processes should be informed by the policy objectives of practitioners, such as improving the budgetary compliance of the wage bill. This facilitates political commitment to HRMIS reforms and ensures their policy relevance, although institutional coordination should be secured as well. HRMIS reforms should be anchored in problems the government considers policy priorities to secure commitment from political leadership. These problems typically include wage bill controls, identifying ghost workers, and providing analytics for decision-making about workforce composition and planning. Since HRMIS reforms are often cross-cutting, institutional coordination among the various government agencies involved in public administration is critical. An institutional mandate and the inclusion of key stakeholders may facilitate this effort.

- The reform process should sequentially strengthen the maturity of the wider system, with defined stages guiding the implementation process. A sequential approach to HRMIS reforms is illustrated in the figures throughout this chapter. They imply the preparation of a rollout strategy—however limited—that plans for the political obstacles ahead and considers the constraints of the existing legal framework. Implementation requires both repeated testing of the solution and creating accessible technical support for users of the HRMIS. Finally, monitoring the reform includes credibly shutting down legacy systems and tracking the use of new solutions.

- A gradual and flexible approach can enhance the sustainability and future development of the HRMIS, due to unexpected data coverage and quality issues. Because HRMIS and other public sector data systems are so complex, unexpected challenges may arise along the way. Data coverage may be incomplete, requiring that additional information be integrated from other modules. Data quality may also be compromised because incorrect human resources (HR) records may be widespread. Therefore, reform teams should build contingency plans from the start, make choices that provide them with multiple options, and be ready to adapt their plan even during the implementation phase.

- The design of the reform should carefully consider the trade-offs involved in choosing different specifications. Design choices have different implications for reform, regarding both the breadth of the reform and its sustainability. For instance, outsourcing the solution to private firms for the implementation of HRMIS reforms may reduce the need to build in-house capacity to develop the software and accelerate the reform timeline, but this choice may still require building capacity for maintenance in the long run. Building internal capacity and managing these operational trade-offs is at the heart of a public service that is most likely to capitalize on technological progress.

INTRODUCTION

The *World Development Report 2021* (World Bank 2021b) outlines the potential for data to improve developmental outcomes. However, as the report highlights, the creative potential of data can only be tapped by embedding data in systems, particularly information systems, that produce value from them. In other words, data must be harnessed into public intent data, defined as "data collected with the intent of serving the public good by informing the design, execution, monitoring, and evaluation of public policy" (World Bank 2021b, 54).

For data to serve this purpose, robust data infrastructures are necessary. Governments across the world collect vast amounts of data, particularly through statistical offices—when gathering information on society—and through the internal use of management information systems (MIS) (Bozeman and Bretschneider 1986).[1] These forms of data infrastructure are used to generate measures in multiple policy domains described elsewhere in *The Government Analytics Handbook*: from budget planning (chapter 11) to customs (chapter 14) and public procurement (chapter 12). Government-owned data infrastructure can generate data analytics to support policy making through the application of statistical techniques (Runkler 2020).

However, data infrastructures that provide analytical insights are still at various levels of maturity in the public sector in general and developing countries in particular. A 2016 report highlights that governments only explore 10–20 percent of the potential value of analytics, in contrast to 40–50 percent in private sector retail (Henke et al. 2016). Multiple factors account for the relative underdevelopment of analytics in public administration. In contrast to the private sector, governments respond to multidimensional demands and diverse stakeholders (Newcomer and Caudle 1991). Siloed and legacy systems inhibit data integration and analytics pipelines (Caudle, Gorr, and Newcomer 1991).

Promoting the use of data analytics in the public sector requires a combination of both technological innovation and organizational change, the analog complements to data analytics (World Bank 2016). In particular, the development of data analytics within the public sector requires a coordinated effort to both transform how data are stored and analyzed and embed these analytical insights into the decision-making processes of public sector agencies. These reforms are often part of a larger digitalization strategy (World Bank 2021a). It is these reforms in data infrastructure that make possible the use of data analytics in the public sector, often led by a public sector reform team.

This chapter provides a road map to the implementation of analytically driven data infrastructures in the public sector, drawing on the experiences of World Bank practitioners and government officials across the world. The substantive focus is on human resources management information systems (HRMIS), a core function within public administration.[2] The conceptual framework outlined provides a foundational perspective on data analytics in the public sector, exploring the established domain of human resource management to illustrate key design decisions in the transformation of data infrastructure into analytics. However, the road map described in this chapter is generalizable to a variety of settings, and throughout the chapter, we emphasize its adaptability to other settings.

The conceptual framework is divided into two parts. The first section provides a typology of the modules that comprise an HRMIS, describing both their content and how they relate to one another. The emphasis is on the distinction between foundational and analytics modules—in particular, how the foundational modules feed into analytical products. Equipped with conceptual clarity about the structure of the information system, we move on to the operational framework for HRMIS reforms. This section describes in detail a framework for HRMIS reforms, outlining a sequential approach to reform (Diamond 2013). The operational framework describes the different stages in HRMIS implementation, their requirements, and best practices for each.

After laying out this conceptual framework, the chapter focuses on a set of case studies to illustrate how it can be applied in practice. Luxembourg showcases the development of a human resources business intelligence competency center (HR BICC), an intricate dashboard that has revolutionized how HR analytics are conducted. The case of Brazil describes how a machine-learning-empowered fraud detection system reduced costs and improved the efficiency of an audit team responsible for overseeing the federal government payroll. Finally, the case of the United States highlights the experience of a team in the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) that developed a dashboard enabling the fast and intuitive use of employee engagement surveys for policy making.

We recognize that these cases are drawn primarily from developed countries (the United States and Luxembourg) and a developing country with a relatively mature HRMIS (Brazil). Practitioners should be aware that while the lessons are generalizable to contexts in which MIS may be less mature, the challenges faced may differ. For instance, in each of these cases, an HRMIS was already in place with foundational modules to use in analytical transformation. This may not be the case in countries where the foundational modules have not been set in place. As a result, while a similar operational framework may be deployed, the types of products (analytical or foundational) may differ substantially. Having said this, we believe these cases illustrate general principles that are useful for all practitioners interested in applying an operational framework for HRMIS reforms. Each case study is described in greater detail in case studies 9.1–9.3 of the *Handbook*.

A set of practical lessons emerge from the conceptual framework and case studies. First, a modular approach to HRMIS reforms enables a precise definition of the reform's scope and how the intervention will affect the broader data ecosystem. Reform teams should consider available resources when deciding which modules to target for reform, as well as how data flow across modules. Second, foundational modules, which focus on payroll, personnel management, and position management, should take precedence over analytical layers, in large part because analytics requires well-structured data records. Nevertheless, analytical layers can be designed for specific modules within an HRMIS if their scopes are sufficiently narrow and well defined. In general, a sequential and flexible approach to data infrastructure reform is recommended. An ex ante assessment of key issues in human resource and wage bill management (both in terms of their likelihood and the severity of their impact on the system as a whole) will enable governments to hedge risks to their initial design.

Finally, the implementation of data infrastructure reforms needs to navigate political-economic issues and ensure leadership commitment and institutional coordination among agencies. To do so, it helps to anchor measurement and analytical outputs in problems the government considers priorities to address. In an HRMIS, these are typically wage bill controls, analytics for personnel decision-making, and workforce planning. Coordination can be facilitated by including key stakeholders and clarifying institutional mandates over data collection and processing. On a more technical note, capacity issues should be considered before and during implementation. Governments often implement large-scale data infrastructure reforms for the first time and may require external assistance from experts who have engaged in similar reforms before.

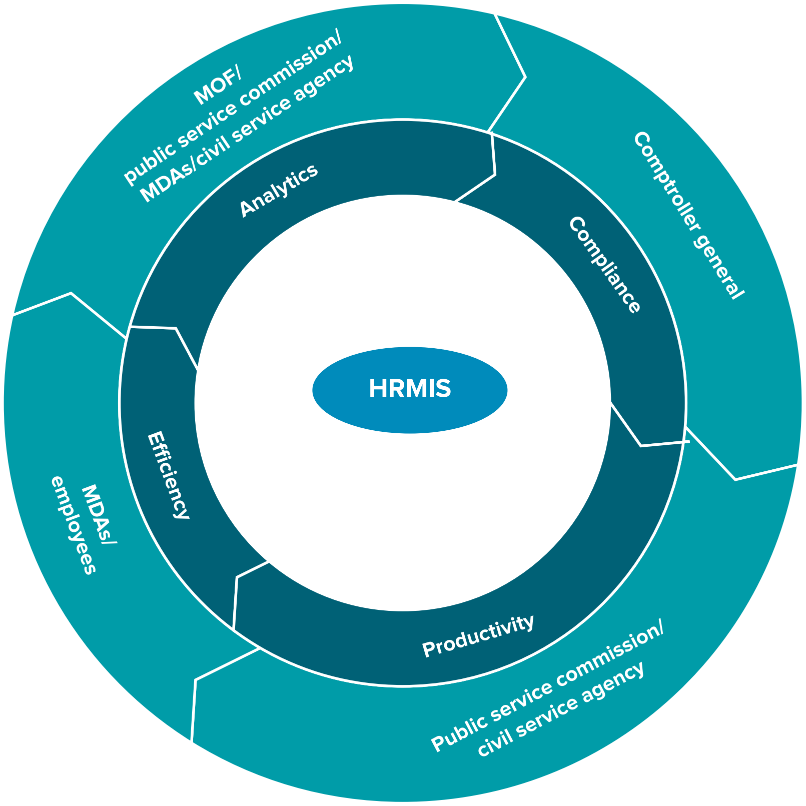
These lessons are generalizable to other data infrastructures. A modular approach to reform can be applied in non-HRMIS settings as well. For instance, reforms for public procurement information systems may focus on the bidding process or on contract implementation. Additionally, analytical insights can be derived from each one of these modules, but only once data quality and completeness are ensured.[3] A sequential and flexible approach to reform is beneficial in non-HRMIS settings as well. Shocks, whether political or technical in nature, can occur regardless of the content of the information system, and reform teams should be ready to address them in both the preparation and implementation phases.

This chapter is structured as follows. Section 2 presents policy objectives and associated issues for an HRMIS. Section 3 presents a typology of HRMIS core modules and how these modules relate to one another. Section 4 presents a framework for HRMIS reform implementation. Section 5 provides an overview of the case studies and applies the conceptual framework to the cases. Finally, section 6 concludes.

HRMIS POLICY OBJECTIVES

In this section, we present four objectives common to HRMIS: compliance for fiscal sustainability, employee productivity, process efficiency, and analytics for decision-making. In so doing, we also highlight issues in their implementation. Each HRMIS module maps onto key policy objectives for an HRMIS: the compliance

FIGURE 9.1 Policy Objectives for Human Resources Management Information Systems and Respective Stakeholders



Source: Original figure for this publication.
Note: HRMIS = human resources management information systems; MDAs = ministries, departments, and agencies; MOF = ministry of finance.

objective is associated with core modules, such as payroll, employee productivity is associated with talent management modules, and analytics for decision-making is associated with analytical modules. Process efficiency is cross-cutting and often directly relates to the way HR records are produced. For example, the appointment of civil servants may require long verification processes, often done manually. Automation of the process could increase HRMIS efficiency. Figure 9.1 highlights these policy objectives and their stakeholders.

Compliance for Fiscal Sustainability

Wage bill compliance with the established budget is necessary for fiscal sustainability. The comptroller general, or another agency responsible for government payroll, manages payroll controls for budgetary compliance. At a more granular level, payroll controls include verifying the authenticity of employment and ensuring accurate payroll calculations and reconciliations. Verifying the authenticity of employment requires identifying and eliminating ghost workers.[4] Ghost workers adversely impact fiscal sustainability and often draw negative attention from the public and policy makers. Another important control is compliance with budgetary ceilings. In many jurisdictions, the approved budget is not directly linked to payroll, leading to overspending. HRMIS regulations should be interpreted correctly and consistently to calculate pay, allowances, and deductions. Employee records should be updated regularly, with bank reconciliations and payroll audits to ensure integrity and compliance.

Employee Productivity

The public service commission and the civil service agency are entities focused on employee productivity and engagement. An HRMIS should give them an accurate overview of employees to help them improve recruitment, develop the performance of the workforce, and enhance their skills. Information on employee qualifications and skills can inform a strategic view of workforce training so that these entities can design training strategies around skills shortages and respond to emerging capacity needs. Recruitment patterns can be analyzed to improve talent pools and reduce potentially discriminatory practices. The performance of the workforce can be monitored using metrics on engagement and attrition rates. In the absence of these measurements, stakeholders are unable to approach employee productivity in an evidence-based and strategic manner.

Process Efficiency

Government agencies, including the ministry of finance, the public service commission, and the civil service agency, are also interested in improving operational efficiency. In some settings, the HR department manually calculates the salaries of employees each month using spreadsheets. This process is not only extremely inefficient but also prone to error. Another example of operational efficiency lies in the hiring process. Hiring departments perform multiple verifications for the first-time appointment of civil servants. These may involve verifying hard copies of key information, such as prosecution history or educational qualifications, from multiple departments and ministries. Manual procedures and hard-copy files delay administrative actions, leading to lower morale, productivity, and efficiency. Process efficiency is therefore another key policy objective for an HRMIS.

Analytics for Decision-Making

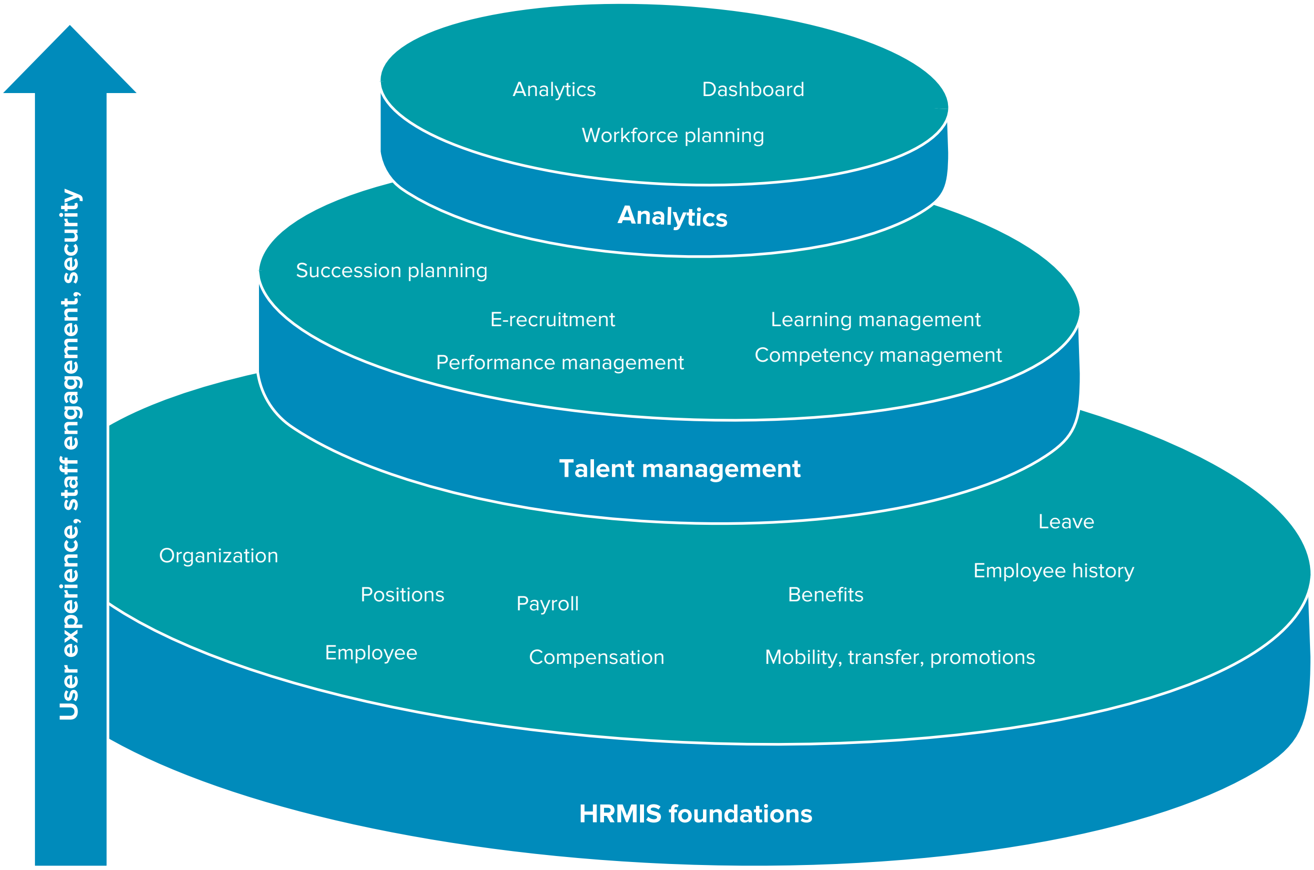
The ministry of finance—together with the civil service agency and the establishment or human resources (HR) department—needs data on HR for making evidence-based, strategic decisions. These decisions concern a range of issues, such as the overall size of the wage bill, salary and pension increases, cuts on allowances, and the financial impact of employee benefits, like medical insurance. Decision-makers need reliable HR reports in a timely manner. In most jurisdictions, these reports are collected manually, through ad hoc measures that take weeks or months, adversely impacting efficient decision-making. In advanced settings, analytics, dashboards, and business intelligence applications are used to enhance effective decision-making.

While these policy objectives are specific to an HRMIS, note that a more general exercise can be done for other information systems. For example, chapter 11 highlights how policy objectives such as budgetary compliance can inform the analytical use of public expenditure data. Ultimately, information systems are designed to assist policy makers in accomplishing their goals. It is only sensible that, depending on the policy area, these goals may differ, but this policy orientation remains the same.

HRMIS CORE MODULES

An HRMIS is an information system designed to enable the management of human resources in the public sector. As such, these technological solutions offer a variety of functionalities, which correspond to modules such as payroll or talent management (figure 9.2). HRMIS reforms require careful consideration of the scope of an intervention—in particular, consideration of which module within the HRMIS will be targeted.

FIGURE 9.2 Human Resources Management Information Systems Data Infrastructure Modules



Source: Original figure for this publication.
Note: HRMIS = human resources management information systems.

Identifying what modules comprise an extant HRMIS enables the consideration of interdependencies across modules, as well as the feasibility of reform.

HRMIS modules comprise four broad categories, ordered from foundational to analytical:

1. **Foundational modules** ensure compliance and control of both the wage bill and personnel. They include payroll (benefits and compensation) management, position and organizational management, career management, and employee profiles, among others.
2. **Talent management modules** include recruitment and performance management, competency, and learning, as well as succession planning. Talent management is used primarily to improve employee productivity.
3. **User experience and employee engagement modules** improve user experience and employee engagement. These modules encompass employee and manager self-service and staff survey systems.
4. **HR analytics** can be developed for workforce planning, as well as strategic and operational decision-making, once the data infrastructure layer has been developed.

These modules provide the basic infrastructure layer for HR data and associated analytical outputs. The foundational modules are responsible for the accurate and reliable storage of any form of HR record. Talent management modules monitor career life cycles, and user experience modules monitor the overall experience of users who interface with the MIS or whose data comprise its case data. Finally, the analytics layer extracts value from the underlying data infrastructure to inform strategic decisions at an operational level.

An analogous structure can be found in other MIS. For instance, when considering customs data (see chapter 14), foundational modules include revenue collection and the release time of goods. These modules comprise their own records ("What was the monetary value of the customs declaration for a particular good?") and potentially their own indicators ("What was the total revenue on exported goods for the month of June?"). Analytical modules provide these indicators to inform policy making. As an example, if customs authorities detected an atypical decrease in tax revenues for a given month, they might send a team of auditors to verify why this occurred. This example highlights how, while the data content of these systems differs, the logic by which they are organized remains largely the same.

Note that each of the modules outlined in figure 9.2 is connected to a set of records and measurements, described in further detail in table 9.1. The table highlights the variety of available HRMIS indicators and

TABLE 9.1 Human Resources Modules and Associated Measures

Module	HR measures
Foundational	
Payroll	Size of wage bill and budget/position compliance; deviation of wage bill expenditures from the budget; sector, administration, and geographical breakdown; percentage of employees in various categories—civil servants, public servants, part-time, wage bill arrears
Position management	Establishment control—employees paid against the approved positions in the budget; average tenure on a position; regional quotas; ratio of public servants, civil servants, political appointees, temporary workforce, and other employee categories
Organization management	Organization architecture reflecting government structure
Employee master data	Tracking of policy effectiveness on gender ratios, regional quotas, and minorities, disabled, and other minority groups; education profiles—degrees and certifications; experience; history of service; ghost workers as a percentage of total workers
Personnel development	Competency improvement measures; promotions; secondments
Benefits management	Benefits and their cost impact
Compensation management	Salaries and allowance structures; compensation equity/disparities in allowances/pay across sectors, administrations, and geographies
Time management	Absentee rate; overtime; staff on various types of leave
Pension	Pension as a percentage of the wage bill; future liabilities; impact of pension increases on budget
Talent management	
Performance management	Top-rated and lower-rated employees disaggregated by ministry, department, and agency; rate of performance reviews completed—ministrywide
E-recruitment	Time to hire; time to fill; applications per job posting; recruitment patterns; applicant profiles; recruitment method—through public service commission, direct contracting, contingent workforce, political appointments, or internal competition; ministry-level appointments
Learning management	Training and skills metrics
Succession planning	Percentage of identified positions that have an identified successor
Workforce planning	Ratios—gradewise and sectorwise; promotions; vacancies
Career development	Promotion rate; average time to promote in a grade per service category
Competency management	Percentage gaps in required competencies

(continues on next page)

TABLE 9.1 Human Resources Modules and Associated Measures *(continued)*

Module	HR measures
User experience and employee engagement	
Employee self-service	Time taken to complete an HR transaction; number of self-service systems available
Manager self-service	Time taken to decide or approve HR transactions; number of manager self-service systems available
Mobile apps	Number of mobile apps available for various HR functions—leave, profile
Employee engagement surveys	Number of employees responding to surveys; satisfaction rate with management and HR policies
HR analytics and reporting	
HR reports, analytics, or dashboards/workforce planning	Levels and distribution of employment; wage bill and its distribution across sectors, administration, and geographies; wage bill and its impact on fiscal sustainability; wage bill as a percentage of revenue; wage bill as a share of GDP; public vs. private employment; sector, administration, and geographic distribution of public employment

Source: Original table for this publication.
Note: HR = human resources.

their corresponding modules, which can be selected and adjusted to practitioners' needs. A payroll module may include different sets of measurements, from the size of the wage bill to a breakdown of contract types for civil servants. This diversity implies that practitioners may select measurements that are relevant to their use case, prioritizing some modules and indicators over others.

OPERATIONAL FRAMEWORK FOR HRMIS REFORMS

HRMIS reforms are designed to address issues and bottlenecks that prevent stakeholders from accomplishing their policy objectives, such as improving compliance and employee productivity. The implementation of HRMIS reforms can be divided into three stages: preparation, implementation, and monitoring. Figure 9.3 outlines the different phases and their respective components, and the following subsections discuss each of the steps outlined.

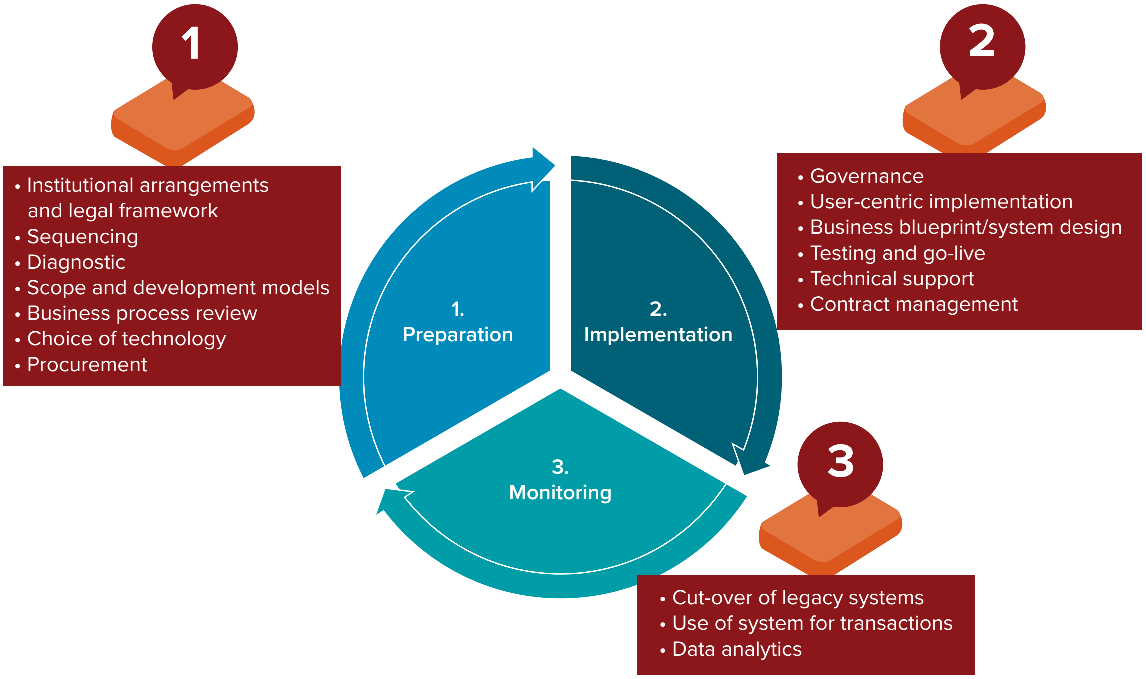
Note that the stages of HRMIS reforms described in the subsequent sections are agnostic with respect to the particular module targeted for reform. Different HRMIS reforms, whether in building an analytics dashboard or improving foundational modules, require a similar approach with regard to the sequence of implementation. Of the multiple elements contained in each of these phases, practitioners are encouraged to begin by assessing which elements are of the greatest importance to the implementation of a particular HRMIS reform project in their setting.

Preparation

The preparation phase lays the groundwork for the implementation of HRMIS reforms. In this phase, key design choices occur, such as defining the scope and the technology to be deployed. Additionally, the preparation phase is an opportunity to engage in a comprehensive diagnostic of the current HRMIS, as well as define the scope of the reform and identify the modules that will be addressed in the intervention.

The preparation phase is an opportunity for the reform team to familiarize themselves with their institutional context as well as the extant data infrastructure, adjusting their strategy in the process. In settings

FIGURE 9.3 Human Resources Management Information Systems Reform Sequence


1
• Institutional arrangements and legal framework
• Sequencing
• Diagnostic
• Scope and development models
• Business process review
• Choice of technology
• Procurement
2
• Governance
• User-centric implementation
• Business blueprint/system design
• Testing and go-live
• Technical support
• Contract management
1. Preparation
2. Implementation
3. Monitoring
3
• Cut-over of legacy systems
• Use of system for transactions
• Data analytics

Source: Original figure for this publication.

where a decentralized HRMIS is in place, the implementation team may require senior management support to promote reforms to different agencies. Much of the effort in the preparation phase should be spent on ironing out institutional coordination issues.

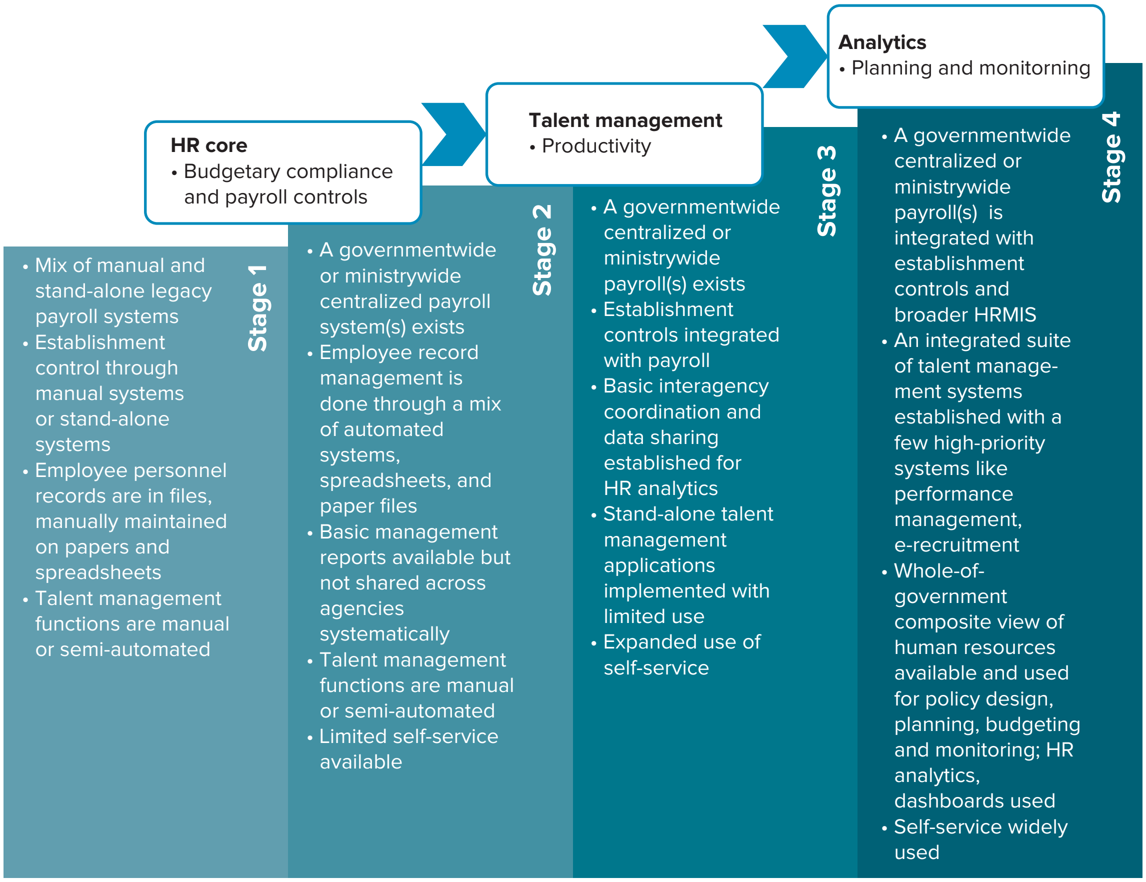
Rollout Strategy and Sequencing

A system rollout strategy should be prepared to guide the implementation. The strategy should address key issues of the scope and sequencing of the rollout. The rollout strategy builds on the mapping of the different modules of the HRMIS. In contrast to the broader HRMIS implementation sequence, the rollout strategy defines how the intervention will achieve the development of the HRMIS. This sequencing accounts for the level of maturity of the current HRMIS, as outlined in figure 9.4.

In terms of sequencing, most countries spend considerable effort on first establishing the basic data infrastructure layer and ensuring compliance and controls (stage 1). Compliance generally centers on the financial compliance of payroll with budget but also on the accurate calculation of payroll to avoid paying ghost workers. Once the data infrastructure layer has been established, it would be prudent to work on data analytics, HR reports, and workforce planning through the implementation of these modules. However, it is worthwhile to note that raw analytical reports can be shared with decision-makers even during the foundational modules, without waiting for the full-blown implementation of analytics modules.

We can generalize stages of maturity in data infrastructure to other cases beyond HRMIS. Budgetary compliance and control in an HRMIS refer to more general principles of data quality and comprehensiveness, which allow for a transition from stage 1 to 2. This is a foundational step for data infrastructure in the public sector: it allows for the reliable collection and retention of data on a variety of HR procedures. The next step is the transition from stage 2 to 3, focusing on productivity: the use of HRMIS data to generate performance or productivity metrics on personnel. This requires access to reliable data, produced by

FIGURE 9.4 Human Resources Management Information Systems Maturity



HR core
- Budgetary compliance and payroll controls

Stage 1
- Mix of manual and stand-alone legacy payroll systems
- Establishment control through manual systems or stand-alone systems
- Employee personnel records are in files, manually maintained on papers and spreadsheets
- Talent management functions are manual or semi-automated

Talent management
- Productivity

Stage 2
- A governmentwide or ministrywide centralized payroll system(s) exists
- Employee record management is done through a mix of automated systems, spreadsheets, and paper files
- Basic management reports available but not shared across agencies systematically
- Talent management functions are manual or semi-automated
- Limited self-service available

Stage 3
- A governmentwide centralized or ministrywide payroll(s) exists
- Establishment controls integrated with payroll
- Basic interagency coordination and data sharing established for HR analytics
- Stand-alone talent management applications implemented with limited use
- Expanded use of self-service

Analytics
- Planning and monitoring

Stage 4
- A governmentwide centralized or ministrywide payroll(s) is integrated with establishment controls and broader HRMIS
- An integrated suite of talent management systems established with a few high-priority systems like performance management, e-recruitment
- Whole-of-government composite view of human resources available and used for policy design, planning, budgeting and monitoring; HR analytics, dashboards used
- Self-service widely used

Source: Original figure for this publication.
Note: HR = human resources; HRMIS = human resources management information system.

the careful implementation of step 1, as well as defining how to measure productivity and which indicators would be necessary to do so. The final step is planning and monitoring, where reliable data produce indicators that inform policy making. This characterizes an HRMIS that has transitioned to stage 4 of data analytics.

Institutional Coordination and Legal Framework

The implementation of a new HRMIS requires institutional coordination due to its complexity and its multiple stakeholders. One of the first steps in institutional coordination is identifying the principal stakeholders in an HRMIS. In most settings, the comptroller general of accounts, under the ministry of finance, is responsible for the payroll and leads the initial implementation to improve budget compliance, payroll calculations, and overall efficiencies. However, in other settings, it is the civil service agency that is responsible for an HRMIS and leads the initial implementation.

An HRMIS may also be decentralized. In this context, the ministry of finance is more focused on ensuring payroll compliance with existing regulations. It requires line ministries and respective HR departments to send payroll payment requests to the comptroller general to ensure budgetary control through an integrated financial management system. This decentralized arrangement could pose additional challenges to reform due to the multiplicity of stakeholders, siloed data, and the need for coordination. In such a case, reform may require additional coordination and buy-in from the implementation team.

Legal authorizations and associated reform strategies should be secured prior to reforms. For instance, the implementation team may require a set of legal documents authorizing the implementation of the reform, which may include terms of reference and procurement requirements to hire a vendor. These documents clearly articulate the scope of the reform to the implementation team and affected agencies. They provide assurance that necessary permissions have been secured for the project, reducing uncertainty and potential negative repercussions for not complying with the existing regulatory framework.

To avoid political resistance to reform, a softer approach can be adopted. Under this approach, line agencies can continue with their previous HRMIS but are asked to provide HR data to a central repository directly from the legacy system. However, this approach does not address inefficiencies associated with duplicative investments and siloed approaches, nor does it ensure compliance once the reform is implemented. Considerable delays and noncompliance can occur, though analytics information for decision-making may become initially available.

Defining the Scope of an HRMIS

Changes to an HRMIS can target one or more of the HRMIS modules outlined in section 2. Note that the choice of scope entails important trade-offs between the breadth of the intervention and the feasibility of the reform project. Three possible scopes for HRMIS reforms are:

- **Increasing the number of indicators and modifying existing indicators for a particular HRMIS module.** For instance, reform stakeholders may be interested in obtaining additional information on the types of appointments of civil servants. This may include further disaggregation of appointment indicators: whether civil servants are tenured, political appointees, or on temporary contracts.

- **Expanding HRMIS coverage by increasing the number of modules.** A key decision may be what institutions will be covered under the HRMIS and how that coverage will be phased in. Certain institutions may differ in important ways from those in the wider service. How will these differences affect the types of indicators available?

- **Ensuring national coverage of the HRMIS, including subnational governments.** Civil servants may be spread across the country in various regions, provinces, or districts. Is the HRMIS meant to reach the entire country or some subset of the country? An analysis of employee coverage in these geographical areas would help define the scope and logistical effort in the implementation of the HRMIS.

By choosing the scope of the HRMIS along each of the above dimensions, the implementation team identifies the key issues about which choices must be made. Key areas to be covered could include legal, policy, and institutional frameworks; HR and payroll regulations (including pay scales and allowances and the extent of their consistent application across ministries, departments, and agencies [MDAs]); budgetary allocation; and key issues of compliance, productivity, efficiency, and analytics.

Choice of Technology

Another key decision point during the preparation phase is the choice of technology (figure 9.3). Two major categories of software technology are available: custom-developed software and commercial off-the-shelf (COTS) software, though some open-source software choices are also available.

Under the custom-developed software approach, the likelihood that users accept the new software and procedures is higher because these implementations can be adapted to user requirements. For example, case study 9.2 outlines how the implementation team in Brazil tailored a solution to detect payroll irregularities using custom-developed software. The solution extracted HRMIS data and presented them in exactly the way payroll analysts required to execute their tasks. This level of customization, however, comes at a cost. Custom-developed systems require higher in-house capacity because all parts of the software have to be coded from the bottom up, rather than relying on a prepackaged solution.

Additionally, maintenance for custom-developed software tends to be higher in the long run because any changes to the underlying data infrastructure require changes to the software itself. If the original implementation team is no longer present—as is often the case—a new implementation team has to start from ground zero.

COTS software often contains prepackaged good practices and tighter integration between different parts of the information system. It also frequently comes with regular software updates, reducing the risk that the technology becomes obsolete. Major COTS packages include SAP, Oracle, and PeopleSoft, though financial management software like FreeBalance also provides public-sector-relevant HRMIS modules. As a result, COTS software applications are more robust and easier to maintain than their customized counterparts. However, this robustness comes at a cost. Adaptation to user needs—such as introducing novel indicators or modules—is, in general, difficult if not impossible to implement within the existing COTS software. Because the software is proprietary, modifications to the underlying software are not available to the implementation team.

Overall, custom-developed software is more suitable for nonfoundational modules, while a COTS solution is better suited to foundational modules because it ensures tighter linkages of these modules with each other. For modules like e-recruitment or performance management, governments can choose any technology platform from the market that meets their requirements and is cost efficient. Integration of these modules with the foundational HRMIS modules will ensure data integrity.

Procurement

In most cases, HRMIS reforms are not fully delivered "in-house" by governments, for a variety of reasons. For instance, COTS solutions require that an external vendor build and deploy an HRMIS for use by a government agency. This includes the introduction of new modules and the training of government officials on how to properly use and manage the software. For customized solutions, the government may lack access to a team of software developers and data engineers to fully develop the solutions. As a result, it may have to rely on external vendors with the required expertise to do so.

As a result, an external vendor must be procured to support reform implementation. The culmination of the preparation phase is preparing a procurement package for the HRMIS implementation partners. The procurement document should cover multiple aspects of implementation: business process review, the deployment and rollout plan, quality assurance and testing of the solution, and the help desk and support strategy, among others. Client and vendor responsibilities, risks, and rights—such as intellectual property—should be protected equitably, in line with industry good practices.

Implementation

The second stage of the implementation of HRMIS reforms, as outlined in figure 9.3, is the actual implementation of the reform plan. The implementation stage includes the management of HRMIS reforms, which first requires considering and defining the governance structure. Additionally, during the implementation phase, it is the responsibility of the implementation team to provide and adapt the business blueprint that guides the project. Iterative testing must take place to ensure that the project scope is being successfully developed. Technical support and a help desk ensure that users are supported throughout the implementation phase. Contract management ensures that expectations are aligned between government clients and external vendors.

The implementation stage of HRMIS reforms thus requires clear authority by the implementation team to make decisions and communicate them clearly to potential external vendors and end users. Flexibility is also required during the implementation stage, as well as proper documentation of any changes in the project design as a result of the implementation stage. Due to this flexibility, it is important to coordinate with external vendors during the rollout of implementation and to collaboratively decide whether changes are indeed feasible under the existing contract or if additional resources—financial or time—may be necessary to successfully roll out the reform. We provide further detail below.

Governance Structure

For effective implementation of HRMIS reforms, it is often necessary to form a steering committee to provide strategic guidance and ensure support from the project sponsor. This steering committee should ensure that key stakeholders are fully represented and consulted. The committee should have the authority to make strategic decisions and resolve strategic-level conflicts. To improve the efficiency of decision-making and the quality of implementation, the steering committee in some settings can also issue basic principles of implementation. These principles can be customized by context and include standardized business processes, user-centric system design, security, and privacy, among others.

The project director should be supported by a project management team, where possible, including procurement and financial management specialists, a project manager, a communications team, and change management teams, among others. A core team of subject matter experts from the ministries should be consulted to ensure they codesign and codevelop the system with the implementation partners. The core team should have a say in decisions carried out by the steering committee, ensuring co-ownership of the solution.

System Design Document

The implementation team should prepare and revise a system design document throughout the implementation of the project. The system design document defines the needs and requirements for the new design of the HRMIS and should be approved by the steering committee before implementation. After launch, any modifications to it should be subject to steering committee approval as well. This living document becomes the final scope document with the technical details of the implemented solution. It also becomes the reference technical design document for future upgrades, and for any new implementation team if the existing vendor changes.

Iterative Testing

Changes to the HRMIS should be developed iteratively. Iterative testing allows for controlled and reversible innovation within the HRMIS reform project, relying on feedback from senior management and staff who will ultimately use the new HRMIS. For instance, an implementation team may be interested in developing an interactive dashboard to measure employee engagement. However, an initial focus on indicators such as employee satisfaction may have to be replaced by employee exit surveys after an initial round of feedback from the steering committee, which is concerned about employee turnover. Iteration preserves flexibility and identifies features that, in the implementation stage, may not be considered relevant. Additionally, it enables adjustment to happen in reversible and controlled stages that do not jeopardize the wider integrity of the project. All changes made during iterative testing should be documented in the system design document.

Technical Support and Help Desk

Technical support allows users to successfully navigate the transition to the reformed HRMIS. Clear documentation on how to use the remodeled HRMIS, as well as a help desk, should be implemented during the project rollout. This ensures users have sufficient information to use the HRMIS during and after the reform process. Failure to do this may result in increased user resistance because users may be confused and unable to operate the new system. Standardized help desk software tools, together with a telephone helpline, should be provided to ensure that user requests are appropriately logged, assigned, resolved, and monitored. Frequently asked questions should be compiled and shared, empowering users to find solutions to their own problems, minimizing help desk calls, and building a knowledge base of solutions.

Contract Management

Contract management is another critical aspect of implementation. Implementation failures are often the result of inadequate contract management. Issues like the scope of the contract and any modifications require that both the steering committee and the vendor align expectations before and during the implementation of HRMIS reforms. Expectations should also be aligned regarding the payment schedule and the responsibilities of the contractor and vendor during the implementation process to avoid confusion and ensure smooth implementation of the project. A collaborative approach in contract management, which considers vendors as partners and not as contractors, is recommended. This collaborative approach creates a mindset of shared responsibility for successful HRMIS reforms.

Monitoring

The third phase shown in figure 9.3 is monitoring the HRMIS once it has been implemented and is in place. The monitoring phase focuses on issues the implementation was meant to address and quantifies the benefits in terms of business results. Often, the implementation team monitors the project in terms of module development, user acceptance, trainings, and so on. While this approach could be useful for internal project management, it has limited utility at the strategic level if the modules have been developed but the business results are not delivered. Therefore, utilization of the system and its coverage should be the key focus of monitoring. If user departments continue to use legacy arrangements while the newly developed HRMIS is only used as a secondary system, the business benefits will be limited.

Transition from Legacy Systems

Even after HRMIS implementation, it is often difficult to fully transition from the legacy system to the redesigned HRMIS. The use of the legacy system as the primary system of records and transaction processing poses a serious challenge. Continued use increases the workload by requiring the constant synchronization of old and new data systems. If the legacy system is still used as the primary system of records after the reform, this reduces the likelihood that the newly developed HRMIS will be used as the primary system. Therefore, during and after the implementation of HRMIS reforms, the legacy system should be gradually shut down to ensure there is a complete switchover to the new system. If required, governmentwide regulations and directives should be issued to ensure the use of the new HRMIS.

Key Performance Indicators

Key performance indicators can help implementation teams gauge the relative success of the implementation process. These indicators should allow the implementation team to monitor how well the reform has performed. For instance, if the implementation team is intervening in a payroll module, it may develop an indicator on the proportion of the wage bill processed through the new HRMIS. Additionally, if one of the goals of the reform is to ensure payroll compliance, indicators can be developed to detect ghost workers. The proportion of employees with verified biometrics is an example of a key performance indicator that enables measurement of this goal.

Monitoring Analytics

The use of monitoring analytics can provide stakeholders with immediate feedback on implementation. An HRMIS should be used to provide analytical information to key ministries involved in strategic decision-making. Initial monitoring should be provided even of foundational modules while the data analytics pipelines and dashboard applications are not fully developed. This will maximize the business value of the data gathered in the HRMIS. It will also provide a political support base for the system when the key decision-making ministries harness the benefits of these investments. These ministries could include the ministry of finance, the public service commission, the civil service agency, and other large MDAs.

CASE STUDIES: HRMIS REFORMS IN PRACTICE

To illustrate the implementation of HRMIS reforms in practice, we provide a set of HRMIS case studies that showcase how government officials and practitioners have employed the techniques outlined above in the reform process. In so doing, we highlight patterns in the development of data infrastructures, common challenges, and the design choices that guided these teams in their development efforts. These cases describe the HRMIS reform process as it was experienced by practitioners. We recognize that these cases represent two developed countries and one developing country with access to a mature HRMIS. As a result, practitioners should tailor lessons in this section to their own context. We highlight how the operational framework for HRMIS reforms is generalizable to other settings as well, from building foundational modules to implementing analytical modules. Subsequent case studies provide a fuller description of the cases, while this section provides a comparative analysis of all three.

Luxembourg

In Luxembourg, the State Centre for Human Resources and Organisation Management (CGPO) is a central government administration, located in the Ministry of the Civil Service. Its mandate spans multiple responsibilities, from managing the life cycle of civil service personnel to strategic workforce planning. In 2016, the CGPO faced growing demands and follow-up needs from HR specialists and decision-makers in the federal government of Luxembourg. As the volume of these requests increased, it became clear to the CGPO that its HRMIS had to change.

In 2017, the CGPO developed and deployed a comprehensive HRMIS reform, which enabled the CGPO to build a comprehensive HR data infrastructure and framework to plan and monitor HR in the government of Luxembourg. The solution developed was large in scale, involving multiple data sources and HR specialists. This analytics center, the HR BICC, was developed over the course of a year and had important transformational consequences for the way HR was conducted. An illustration of the novel dashboard is presented in figure 9.5. It integrates both HRMIS data and strategic planning documents in a comprehensive dashboard portal (in orange).

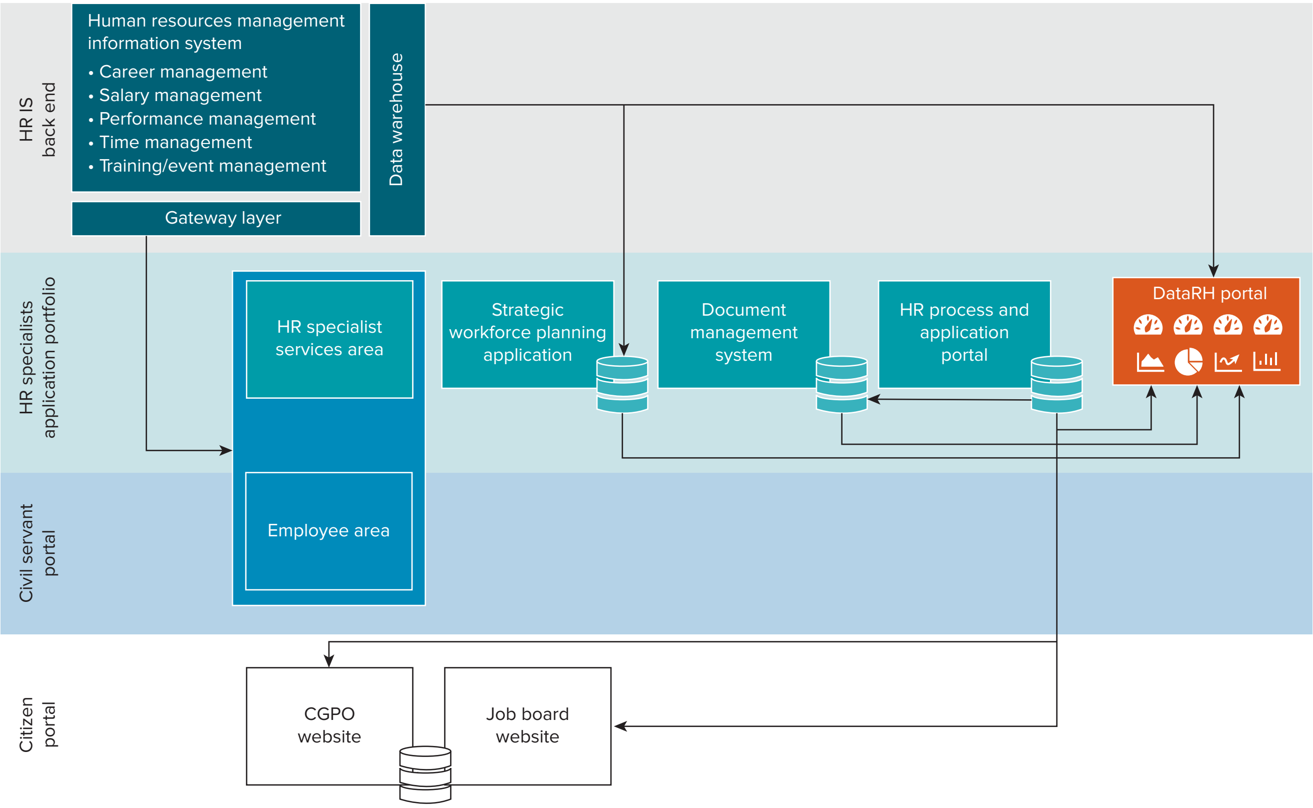
The Luxembourg case presents a fully integrated HRMIS pulling together all the major databases that are typically the focus of such exercises. As such, this case is the most comprehensive example of the implementation of a full HRMIS analytics module, as outlined in figure 9.2.

Brazil

In Brazil's federal government, payroll quality control is the responsibility of the Department of Compensation and Benefits (DEREB). DEREB flags paycheck inconsistencies before disbursement, which are then forwarded to federal agencies' HR departments for correction. The task is challenging. The case volume is large, with tens of thousands of individual paychecks processed daily. Additionally, a complex set of regulations governs how payments should be disbursed. To enforce these rules and detect inconsistencies, a team of payroll analysts individually verifies each paycheck. The implementation team sought to improve this process.

In 2019, a partnership between DEREB and a private data science consulting firm (EloGroup) resulted in the development of a machine-learning-empowered fraud detection system. To generate the necessary data

FIGURE 9.5 Human Resources Management Information System, Luxembourg


HR IS back end
Human resources management information system
• Career management
• Salary management
• Performance management
• Time management
• Training/event management
Gateway layer
Data warehouse
HR specialists application portfolio
HR specialist services area
Strategic workforce planning application
Document management system
HR process and application portal
DataRH portal
Civil servant portal
Employee area
Citizen portal
CGPO website
Job board website

Source: Adapted from CGPO.
Note: CGPO = State Centre for Human Resources and Organisation Management; HR = human resources; IS = information system.

to train this algorithm, a thorough restructuring and integration of the extant data infrastructure on payroll, compensation rules, and HR were developed. Through the development of new extraction, transformation, and loading (ETL) processes, this solution enabled auditors to better detect irregular payroll entries, increasing savings and improving efficiency.

The Brazil case illustrates that, although some HRMIS reforms may be relatively narrow and aimed at a particular outcome—in this context, fraud detection—many of the themes outlined in earlier sections are still of relevance to their implementation. Many of the steps taken in the development of the fraud detection system are foundation stones for wider HRMIS reforms, highlighting how the same methodology can be applied even in smaller contexts.

United States

Every year, the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) is administered to over 1 million federal civil servants in the United States.[5] The FEVS measures employees' engagement through a variety of survey questions and provides valuable information to government agencies. In theory, it presents data on agency strengths and opportunities for improvement in employee engagement. However, extracting insights from the FEVS is challenging. Once given access to the survey, government agencies spend weeks analyzing the data to operationalize their findings. This effort is labor intensive and costly. An HRMIS reform team sought to accelerate this process.

In 2015, the NIDDK, within the National Institutes of Health (NIH), developed the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) to extract rapid and actionable insights from the FEVS. While not generating a new data infrastructure, EVS ART relied on the creative use of Microsoft Excel to extract and transform data to produce dashboards automatically from a single data file. Effectively, the Excel worksheet developed by the NIDDK team integrated a data infrastructure and a dashboard into a single platform.[6]

The US case illustrates how a grassroots initiative, undertaken within the public service rather than as a centralized effort, faced some of the key issues outlined above in its HRMIS reform process. While limited in scope, the implementation team found creative solutions to derive strategic value from the FEVS. It adapted the solution to its needs and was able to effectively improve how survey evidence could be operationalized into improvements to employee engagement.

What Modules Were Targeted for HRMIS Reform?

All the cases we have presented directly relate to HRMIS and can be mapped directly onto the HRMIS core modules in table 9.1. In the case of Luxembourg, a new HR analytics module was developed, with dashboards and reports, to enable HR management by the CGPO. In Brazil, a machine-learning algorithm was deployed to ensure that payments followed established regulations, in a clear example of the compensation management module. Finally, EVS ART is an example of an employee engagement module: a federal engagement survey to guide strategic planning in the NHS.

Note that in the cases of Luxembourg and Brazil, although the end product targeted a single module, the solutions required the deployment of multiple modules. For instance, in Brazil, both the compensation module as well as the employee and organizational modules were combined to provide data for the machine-learning algorithm. In the case of Luxembourg, the analytics dashboards were supplied with data from various core modules in the HRMIS, such as compensation, organizational management, and performance management. For the United States, since EVS ART was based on a single employee engagement survey (the FEVS), no additional HRMIS modules were integrated.

Preparation: Laying the Groundwork for the Intervention

This section outlines general principles involved in the two initial stages of the development of data analytics: preparation, where practitioners lay down the groundwork for the intervention, and implementation, where a decision-making process that is collaborative and adaptable plays a crucial role. We highlight what is generalizable and particular about each phase for the specific cases analyzed in this chapter. The accounts are not designed to be exhaustive: rather, they illustrate key concepts and sequential logics that may apply to the practitioner.

Institutional Coordination

A key factor in the preparation phase is obtaining the necessary support from senior leadership. This support is what confers on the reform team the authority to make executive decisions and secure collaboration for the intervention. In general, a centralized authority with a mandate over a policy area makes reform easier. In Luxembourg, the implementation team was commissioned by the CGPO. The CGPO enjoyed a broad mandate that focused specifically on HR, from the management of the life cycle of personnel to strategic workforce planning. This broad mandate meant that once the decision to develop a new dashboard was made, no additional permissions were necessary.

In the United States and Brazil, leadership support was granted by senior management within the respective agencies. In the United States, the implementation team was based in the NIDDK, situated within the NIH. The NIDDK's senior leadership understood the importance of the effort and supported the team's effort—granting time, flexibility, and necessary resources. In Brazil, the senior leadership of the Department of Personnel Management and Performance (SGP), which oversees DEREB, gave full support to the project.

Respecting the Legal Framework

The development of innovative technologies, such as data analytics, requires careful consideration of the existing legal framework, particularly in the public sector. It is necessary to assess whether there are rules and regulations in place that may limit the scope of the intervention and to ensure that the proper permissions are obtained. Depending on the mandate of the agency, as well as the regulatory environment, different legal permissions may be necessary. For instance, in Luxembourg, due to the CGPO's broad legal mandate to generate analytical insights on HR, it was not necessary to request additional permissions to implement the analytics pipeline. In the US, likewise, due to the limited scope of the intervention, no extensive legal framework was needed.

In Brazil, however, where regulations and norms govern how projects are implemented, extensive legal consultations were necessary. The agency partnered with the consulting firm, as well as with another agency familiar with technological innovation projects, to draft the project proposal and obtain the necessary permissions and legal documents. These cases highlight how interventions operate within the boundaries of existing legal frameworks and need to abide by laws and regulations to ensure their legality and feasibility.

Choice of Technology

As outlined above, COTS solutions strengthen sustainability in the long run because they offer the technical assistance of a dedicated enterprise and tightly integrated tools. On the other hand, COTS solutions often lack the precision of custom-developed solutions, which are tailored to the specific needs of clients. COTS solutions may also cost more due to the high cost of licenses and upkeep. Custom-developed solutions, while more adaptable and flexible, require costly investment in a team of skilled developers to create as well as a long period of iterative maturation. Additionally, upkeep may be expensive if proper code documentation and dedicated maintenance staff are not set in place.

Our cases illustrate these trade-offs. Luxembourg opted for a COTS solution—in particular, a dashboard tool that had already been deployed by the implementation team in another, non-HRM context. The team opted to repurpose that tool for their needs, capitalizing on accumulated experience from a previous project, with a relatively short maturation period. The United States also opted for a COTS solution, Microsoft Excel, which was heavily customized for the requirements of EVS ART. The tool allowed the team to generate indicators and dashboards through the development of scripts that automatically converted data input from the FEVS into dashboard outputs.

Brazil opted for custom-developed, open-source software, developing its solution using Python and open-source machine-learning packages. The solution was deployed in a computing cluster on the cloud, where both a data pipeline and a fraud detection statistical model were hosted. The solution was tailored to the specific requirements of the auditing team, capturing both business process regulations and anomaly detection algorithms with the available HR and payroll data. Due to the technical nature of the project, its implementation was outsourced to a consulting firm.

Scope and Deployment Models

There are clear trade-offs embedded in the choice of the scope of a project. Narrow scopes allow for quicker implementation and greater ease of use. However, they make it more difficult to scale across agencies due to their highly specialized nature. Broad solutions require intensive training and adaptation by users, as well as additional resources for the building of complex tools.

Luxembourg's CGPO opted for a broad scope, commensurate with its broad HRM mandate. The dashboard ecosystem was expansive and provided a wide array of insights, ultimately producing over 157 applications (HR dashboards) and over 2,600 sheets. This complexity required extensive data-quality assurance processes, as well as the training of HR specialists to learn how to use these different tools. A dedicated helpline provided additional assistance.

In contrast, Brazil and the United States had a narrower scope for their solutions. Brazil's solution focused specifically on fraud detection in the federal payroll for the subset of manually imputed payments only. This tailored approach was limited in use to a specific agency and was not amenable to scaling. The NIDDK in the United States focused exclusively on generating insights from the FEVS to guide the agency's decision. The focus was on employee engagement and methods to improve the agency's responsiveness. Due to the broad coverage of the survey itself, however, other agencies expressed interest in deploying the dashboard, proving that it was, in fact, generalizable.

Implementation: An Adaptive Journey

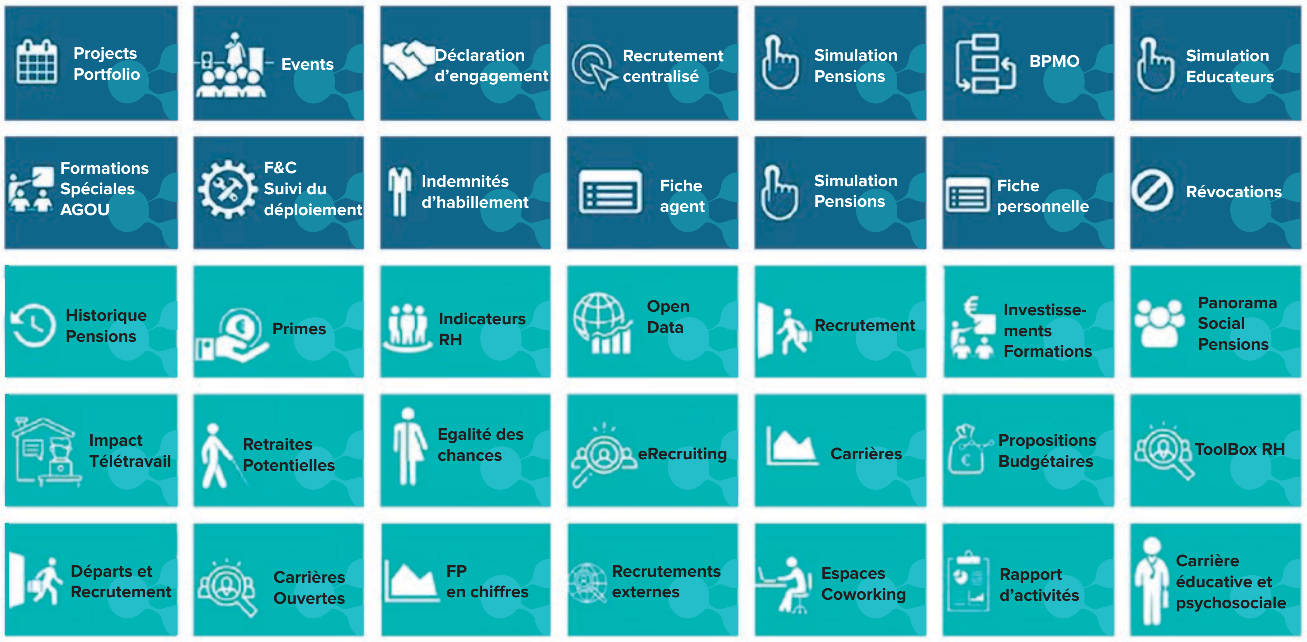
User-Centric Implementation

In user-centric implementation, the data infrastructure and solution requirements are defined by how users will use information. Data analytics pipelines are designed to answer user queries and provide answers to a well-defined set of problems, which then inform the required data infrastructure to provide these input data.

For Luxembourg, the mandate for the solution was broad, and the user base varied. The final design of the dashboard attended to multiple user bases, from citizens to HR specialists within the government. Mapping out each user to their use case and ensuring that the dashboards could attend to those needs separately but simultaneously was a key design choice by the implementation team. Multiple data pipelines and dashboards were designed, each for particular areas and users, and within each of these dashboards, multiple data visualizations were available. Figure 9.6 outlines the multiple modules contained in the dashboard solution, including information on pensions and recruitment processes.

For Brazil, extensive consultation occurred among frontline providers (auditors) who were going to use the solution. Feedback regarding the necessary data structure and how it would feed into their auditing decisions was crucial. The team opted for a simple risk score associated with each payment, along with flag indicators for the type of rule violated. In the United States, the users were primarily the management making strategic planning decisions for the agency. As such, the indicators were actionable, such as worker

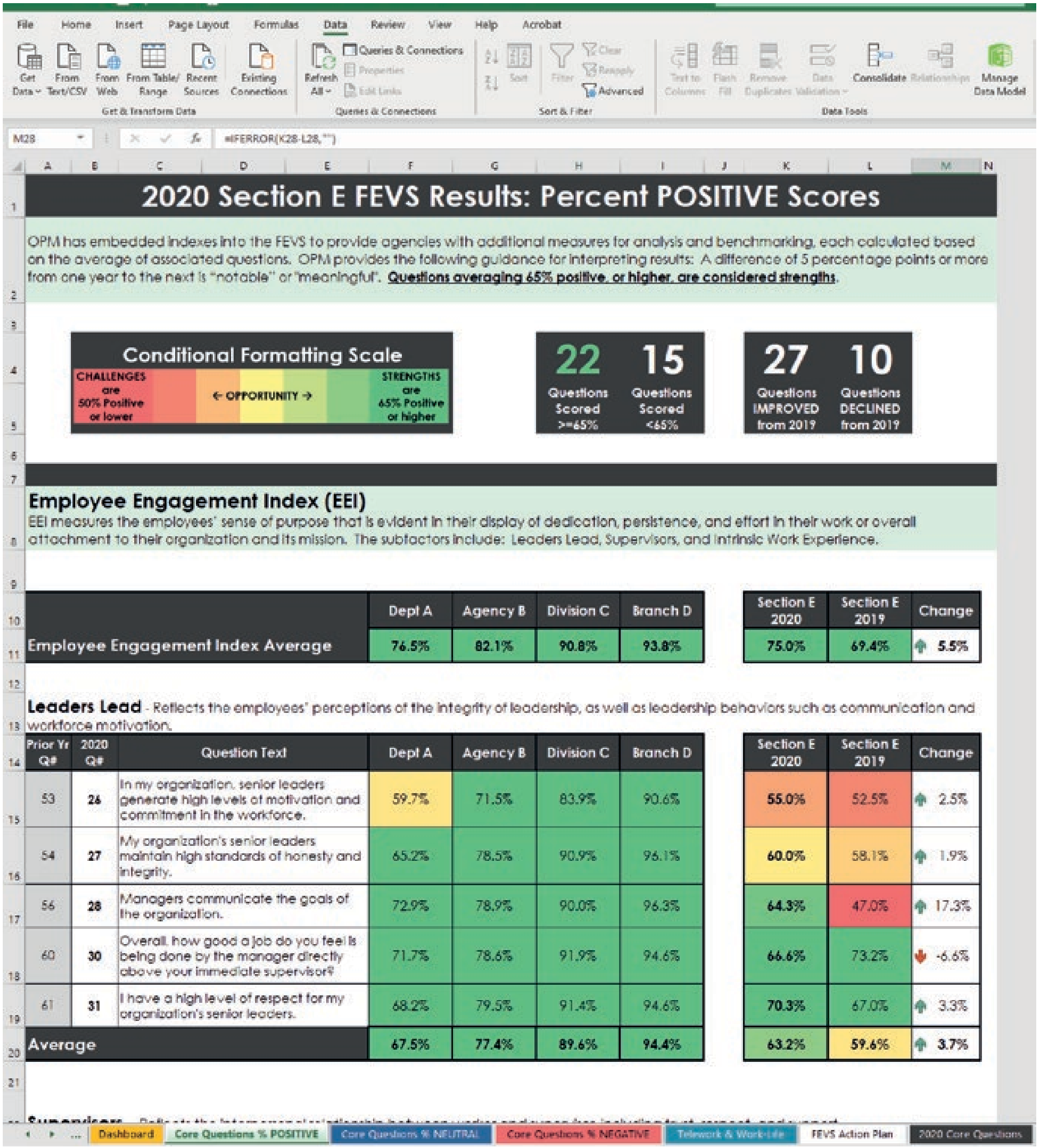
FIGURE 9.6 Luxembourg's Dashboard Modules



Source: State Centre for Human Resources and Organisation Management (CGPO).

engagement and performance metrics. Organization-level indicators, with positive, neutral, and negative scores, provided ready-access insights into the relative performance of the agency compared to the previous year (figure 9.7). EVS ART also provided an action-planning tab to facilitate strategic planning.

FIGURE 9.7 Percentage of Positive Employee Engagement Scores from the Federal Employee Viewpoint Survey



M28 =IFERROR(K28-L28,"")

2020 Section E FEVS Results: Percent POSITIVE Scores

OPM has embedded indexes into the FEVS to provide agencies with additional measures for analysis and benchmarking, each calculated based on the average of associated questions. OPM provides the following guidance for interpreting results: A difference of 5 percentage points or more from one year to the next is "notable" or "meaningful". **Questions averaging 65% positive, or higher, are considered strengths.**

Conditional Formatting Scale

CHALLENGES are 50% Positive or lower	← OPPORTUNITY →	STRENGTHS are 65% Positive or higher

22	15	27	10
Questions Scored >=65%	Questions Scored <65%	Questions IMPROVED from 2019	Questions DECLINED from 2019

Employee Engagement Index (EEI)

EEI measures the employees' sense of purpose that is evident in their display of dedication, persistence, and effort in their work or overall attachment to their organization and its mission. The subfactors include: Leaders Lead, Supervisors, and Intrinsic Work Experience.

	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
Employee Engagement Index Average	76.5%	82.1%	90.8%	93.8%	75.0%	69.4%	↑ 5.5%

Leaders Lead - Reflects the employees' perceptions of the integrity of leadership, as well as leadership behaviors such as communication and workforce motivation.

Prior Yr Q#	2020 Q#	Question Text	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
53	26	In my organization, senior leaders generate high levels of motivation and commitment in the workforce.	59.7%	71.5%	83.9%	90.6%	55.0%	52.5%	↑ 2.5%
54	27	My organization's senior leaders maintain high standards of honesty and integrity.	65.2%	78.5%	90.9%	96.1%	60.0%	58.1%	↑ 1.9%
56	28	Managers communicate the goals of the organization.	72.9%	78.9%	90.0%	96.3%	64.3%	47.0%	↑ 17.3%
60	30	Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor?	71.7%	78.6%	91.9%	94.6%	66.6%	73.2%	↓ -6.6%
61	31	I have a high level of respect for my organization's senior leaders.	68.2%	79.5%	91.4%	94.6%	70.3%	67.0%	↑ 3.3%
Average			67.5%	77.4%	89.6%	94.4%	63.2%	59.6%	↑ 3.7%

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

Iterative Testing

The development of each data infrastructure and analytics pipeline was gradual and flexible. All implementation teams demonstrated a willingness to test, adapt, and redeploy their analytics solution at each stage of the implementation process. For instance, the Luxembourg CGPO first developed a set of dashboards on legacy career and payroll data. Once the initial dashboards were complete, the team realized that quality issues compromised the integrity of the data analysis. As a result, additional quality controls were set in place to ensure that the dashboards were generated as expected. User feedback and demands gradually expanded the scope of the dashboards, and the implementation team worked on iteratively expanding the scope of the HR BICC.

In the United States, the NIDDK had its own learning curve. The team had to work through a process of backward induction, starting from their conceptualization of the final product while researching and learning how to accomplish each step along the way. From the visual appearance of graphs to the data pipeline required to feed them, each task was iteratively resolved and incorporated into the final product. In Brazil, the implementation team tested alternative machine-learning algorithms to improve fraud detection. Repeated consultation with the auditor team generated new ideas, such as incorporating business-rules flags.

Technical Support

Technical support is crucial to help users navigate the complexity of novel data analytics pipelines—and to be able to build on them. These support systems reduce confusion and facilitate the adoption and diffusion of the technology by new users. In Luxembourg, the CGPO created a helpline to assist HR specialists in the use of the newly developed tools, increasing uptake and facilitating the transition from the legacy system to the new one. The NIDDK team in the United States organized training workshops with different teams and agencies to explain how to use EVS ART as a planning tool. The implementation team created an instruction manual that allows users to navigate the dashboard easily, along with clear and accessible explanations for key indicators and metrics.

In contrast, the Brazil payroll team developed its solution through outsourcing and did not provide a robust system to assist users. The consulting firm, while communicating about the development of the tool, did not create formal channels to address questions and bugs. Rather, support was provided in an ad hoc fashion, depending on user feedback, to address bugs in the code or deployment. While the intense coordination between the consulting firm and the agency reduced user confusion, the lack of a dedicated support team, particularly after the completion of the project, raises concerns regarding the future correction of unexpected changes in the data infrastructure.

LOOKING BEYOND: FEASIBILITY AND SUSTAINABILITY

The cases presented in this chapter illustrate key design choices and their feasibility constraints. Luxembourg's HR BICC has high entry barriers: its implementation relied on an in-house team that had previously deployed a similar business intelligence solution, as well as on hiring a team of in-house data scientists and IT staff to maintain and develop the solution. These investments are costly and rely on institutional changes that may be prohibitively difficult in other contexts. However, these investments facilitate the sustainability of the solution and its continued development.

Brazil's solution was agile and less costly but in many respects brittle. In less than a year, the implementation team was able to produce a machine-learning-based fraud detection system, but technical and political-economic issues raise concerns regarding its sustainability. Reliance on an external development team meant that in-house capacities were not developed. The sustained development and maintenance of the solution are at risk. Additionally, changes in the management of the agency mean that accumulated expertise during the implementation phase can be lost through succession cycles.

In EVS ART, a narrow scope and sustained implementation—over the course of two years—meant that the solution was widely disseminated and consolidated within the NIDDK. Additionally, it was designed proactively for planning and monitoring within the agency, resulting in tight integration with the analytics component. In many respects, the project is replicable with low feasibility constraints, given the ubiquity of Microsoft Excel in the public sector. At the same time, the highly specialized scope of the solution means that it is not easily modularizable and portable to other domains. Excel spreadsheets are, in general, not amenable to scaling. Furthermore, manual data imputation and modification make developing an automated analytics pipeline challenging.

Note, additionally, that monitoring of the solution was an important component in some but not all of these cases. For Luxembourg, the cutover of legacy systems was implemented in tandem with technical support and the training of HR specialists. The deployment of an extensive analytics dashboard gave administrators live feedback and an overview of the new HRMIS. In the United States, EVS ART replaced the manual approach to obtaining insights from the FEVS survey. In Brazil, a new data pipeline was built on top of an existing legacy system but did not seek to replace it.

These concerns are generalizable to non-HRMIS settings. If the implementation team lacks the financial resources and capacity to engage in a large overhaul of the information system, the scope of the project should be limited to a single module or two. At the same time, the team should consider whether a smaller intervention could have potential linkages to other data modules. Additionally, when engaging with external actors, implementation teams should consider how to ensure the sustained development of the solution after implementation concludes. To reiterate, these lessons are not restricted to HRMIS and can be applied in other administrative contexts, such as public procurement and customs data information systems.

CONCLUSION

This chapter has outlined the practical issues and challenges in developing data infrastructure for improved government analytics. It has focused on MIS targeted at HR data, but the lessons presented in the chapter apply to public sector data systems more generally.

The chapter has presented the key stages through which HRMIS implementation or reform typically occurs, structured around an operational framework for HRMIS reforms. It has grounded this conceptual discussion by illustrating these stages using case studies from Brazil, Luxembourg, and the United States. The discussion is based on the World Bank's experience implementing data systems in government agencies across the world, as well as on the experiences outlined in the case studies.

There are trade-offs involved in each of the design choices presented. Without robust quality assurance processes in place, the validity of analytical insights is fragile. But expansive quality assurance may be prohibitively costly and is not feasible for all contexts. Deciding the optimal, feasible level of data quality for an analytical pipeline is a design choice, which highlights how the pipeline of data analytics is highly adaptable. Some countries opted for COTS solutions, while others opted for more customized approaches. Agile development, outsourced to external companies, may provide quick results, but it raises sustainability concerns.

The case studies presented in this chapter demonstrate the complexity and diversity of HRMIS implementation. While defying a one-size-fits-all approach, the cases illustrate how a set of different tools, when applied by a dedicated implementation team, can carve out the space for a more analytically driven HRMIS and data infrastructure more generally. Developing systems that both store and extract analytical insights from public data requires widely applicable methodologies. While the specific applications of data systems may vary, the methodology outlined in this chapter and illustrated here in practice provides a conceptual framework with which to approach this challenge. More detailed expositions of the chosen case studies now follow for those readers who want to better understand the individual HRMIS solutions described in summary here.

Beyond the examples presented in this chapter, we highlight the innovative uses of an HRMIS beyond payroll and HR. The underlying theme for this innovation is the use of disruptive technologies like data lakes and artificial intelligence (AI) to cross-reference HRMIS data with multiple other data sources in order to accomplish a policy objective. For example, HR data can be used to analyze procurement and economic activity data in order to identify corruption. In Brazil, HR data on civil servants were cross-referenced with public procurement contracts through the use of big data and AI. The AI tool identified more than 500 firms owned by public servants working at the same government agency that executed a public contract.[7] HR data can also be used to cross-reference budget data in order to improve performance by identifying which civil servants lead particular budgetary programs.

In sum, HRMIS—and MIS more generally—can play a crucial role in the innovative use of data to further policy objectives such as reducing corruption and improving the overall performance of the public sector. The conceptual framework presented here extends beyond HRMIS: the identification of data infrastructure modules and an operational framework for reforms can be applied in a variety of policy settings, as highlighted in other chapters of this book. Ultimately, extracting value from data—transforming them into public intent data—means anchoring them to clearly articulated policy objectives. Articulating what these policy objectives are, and what data are required to measure the achievement of these goals, is the first step toward creating data infrastructures for government analytics.

CASE STUDY 9.1 HRMIS CASE STUDY: HUMAN RESOURCES BUSINESS INTELLIGENCE COMPETENCY CENTER (LUXEMBOURG)

Ludwig Balmer, Marc Blau, and Danielle Bossaert

SUMMARY

In 2017, the State Centre for Human Resources and Organisation Management (CGPO) developed and deployed a human resources business intelligence competency center (HR BICC), which enabled it to build a comprehensive HR data infrastructure and framework to plan and monitor HR in the government of Luxembourg. The solution developed was large in scale, involving multiple data sources and HR specialists. This analytics center, developed over the course of a year, had important transformational consequences for the way HR was conducted.

INTRODUCTION

A seemingly narrow question—how much does the government spend on personnel?—requires integrating human resources (HR) data from multiple modules. Which employees (position), types of payment (payroll), and government agencies (organization module) should be included in the wage bill? Policy makers require immediate answers to these questions to make informed personnel decisions. However, a human

Ludwig Balmer is the head of information technology for the Centre for Human Resources and Organisation Management (CGPO). Marc Blau is the director of the CGPO. Danielle Bossaert is the head of the Observatory of the Civil Service (Ministry of the Civil Service).

resources management information system (HRMIS) often reacts to ad hoc queries rather than proactively offering a system of answers. This project sought to change that.

In Luxembourg, the State Centre for Human Resources and Organisation Management (CGPO) developed a human resources business intelligence competency center (HR BICC) to provide an integrated overview of HRMIS data in accessible dashboards. This case study shows how this complex technology was developed. The comprehensive scope of the project meant that it integrated a variety of modules, from payroll to talent management. This contrasts with the more tailored approaches of Brazil (case study 9.2) and the United States (case study 9.3). It also provides the clearest example of what chapter 9 describes as an analytics module, the use of HRMIS for strategic and operational decision-making.

A few key lessons emerge from this project. First, quality assurance is paramount to the integrity of the analytics module. The team iteratively cleaned the data and established control protocols to protect its integrity. Second, it is important to reduce the burden of visualization on users. Ensuring visual coherence across dashboards and providing different choices of visualization reduces confusion and increases accessibility. Finally, it is important to provide users with additional support outside the dashboard itself. A helpline can guide users in the proper use of the dashboard as well as generate feedback on whether it is functioning as intended.

This case study is structured as follows. Section 1 provides institutional context on the HRMIS and its management. Section 2 describes the initial challenge and gives an overview of the solution itself. Section 3 explains the project's rollout strategy and reform sequence. Section 4 outlines the lessons learned in the project. Section 5 outlines the impact of the solution. Finally, we conclude.

INSTITUTIONAL CONTEXT

The CGPO is a central government administration in Luxembourg, located in the Ministry of the Civil Service. Its mandate spans multiple responsibilities, including:

- Management of the entire life cycle of personnel, including candidate selection, onboarding, and professional development
- Calculation and management of remuneration and the careers of active state officials
- Management of retired state officials and pension beneficiaries
- Strategic workforce planning management, as well as HR data dashboard publication.

Alongside these responsibilities, the CGPO also provides consulting services. These include business process management and optimization, organizational development, digitalization, and project management. To manage HR data, the CGPO uses an integrated HRMIS, customized to suit its needs. Before the deployment of this solution, the system included information on the careers and salaries of civil servants in Luxembourg. HRMIS data were already centrally managed and stored. Regular and ad hoc extraction routines were executed to provide data insights to CGPO users as well as other public institutions.

INITIAL CHALLENGE AND PROPOSED SOLUTION

In 2016, the CGPO faced growing demand and daily follow-up needs from internal HR specialists and decision-makers in the government. As the volume of demands increased, the CGPO decided to design and deploy an HR BICC. The purpose of the center was to facilitate a comprehensive overview of HRMIS data through the development of dashboards. This would reduce the burden on the CGPO to respond reactively to demands and would empower consumers of HRMIS data to formulate questions and search for answers within each dashboard.

User orientation was an important principle in the project and was reflected in the development of interactive dashboards (an example is given in figure 9.8). The dashboard included two components: a more general data analysis perspective and an operational HR perspective including key indicators for HR specialists to track. In contrast to the previous reactive approach, the project generated a set of readily available visualizations to inform policy making by HR specialists and other agencies in Luxembourg's government.

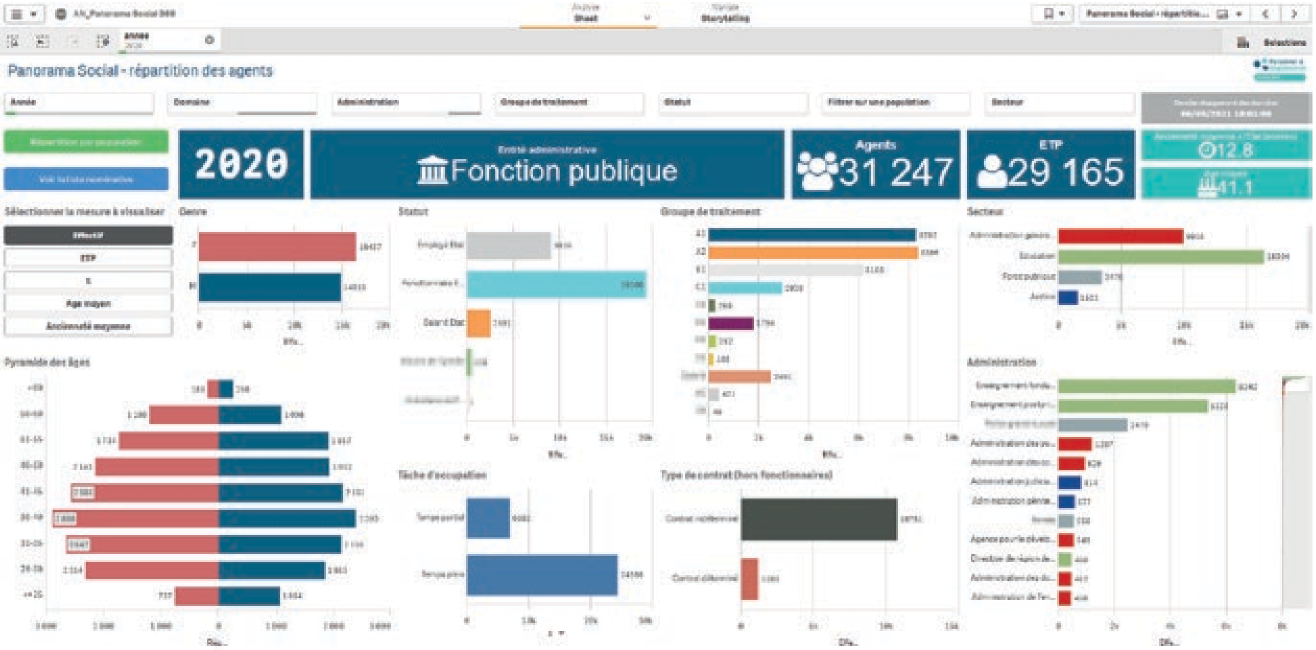
Before the project's implementation, the legacy HRMIS only considered the management of careers and salary computation. The project expanded the set of modules in the HRMIS, including performance and training modules. The simplified diagram presented earlier in figure 9.5 shows the main applications and the workflow of the solution. The HR BICC integrates multiple databases and dashboard applications, each tailored for different use cases, including HR specialists, employees, and citizens.

Note that in figure 9.5, the DataRH portal (in orange) is fed by multiple databases beyond the HRMIS itself. Its data pipeline includes more strategically oriented databases, such as the strategic workforce planning application. This tight integration between databases designed for strategic workforce planning and the HRMIS data promotes a strategic orientation for the HR BICC.

ROLLOUT STRATEGY AND REFORM SEQUENCE

In mid-2016, the initial decision was made to develop the HR BICC (table 9.2). In October of the same year, the project was formally launched. The first step was procurement and the launch of data warehouse deployment. The CGPO identified a business intelligence (BI) team that would be responsible for the implementation of the dashboard. After completing the selection process, the CGPO opted to hire an in-house team that had developed a similar solution in another, non-HR area within the government. It therefore opted against procuring the BI tool externally, in contrast to the Brazil HRMIS case study.

FIGURE 9.8 Sample Dashboard from Luxembourg's HR BICC



Source: Screenshot of HR BICC dashboard, CGPO.
Note: CGPO = State Centre for Human Resources and Organisation Management; HR BICC = human resources business intelligence competency center.

TABLE 9.2 Project Timeline for Luxembourg's HR BICC

Period	Main steps
Mid-2016	Decision to put an HR BICC in place
October 2016	HR BICC project kickoff ● Launch of BI tools procurement process (call for proposals) ● Launch of data warehouse deployment project
February 2017	BI tool selection and deployment, start of governance process and documentation
March 2017	Setup of data warehouse architecture
March 2017	Start of dashboard production

Source: Original table for this publication.
Note: BI = business intelligence; HR BICC = human resources business intelligence competency center.

The main consideration was that the solution that had previously been deployed by the BI team would not only fit the CGPO's initial needs but would also be scalable in the future. The skills developed by the in-house team were transferrable: they had already developed data infrastructure and a previous version of the dashboard tool in another area. This procurement strategy allowed the CGPO to capitalize on previous experience and substantially accelerate the deployment of the solution. As a result of this decision, dashboard production was initiated shortly after the BI tool was selected, in March 2017. In the same month, the redesign of the data warehouse architecture for the HRMIS commenced.

The legal framework was an important consideration for the project. The General Data Protection Regulation (GDPR) impacted both the source side of the data export routines as well as user access management. Monitoring technologies were built into the BI tool to address security concerns. Plug-in tools tracked user activity, tracing how apps, sheets, and data were used or visited by users. This allowed the CGPO both to understand how the HR BICC was used and to ensure that user access was carefully monitored.

The implementation team faced several challenges during the rollout of the project. The first was ensuring quality control of HRMIS data. Because the HRMIS was initially built to perform specific operations, such as salary computation and career management, HRMIS data were not always complete or consistent. As a result, in the initial stages of statistical analysis and dashboard preparation, the team identified missing data series and inconsistent results. To overcome this issue, the team designed a data relevance and quality review process while, in parallel, training civil servants on how to respect it. This quality review process is now part of the CGPO's daily routines.

The second main challenge was providing technical support and a help desk for CGPO staff. The dashboard introduced a new way of working for HR internal specialists. Due to the novelty of the dashboard, internal teams had to adapt their business activities and processes to benefit from the new sources of information and ways of interacting with it. The implementation team also had to respond to new requests by users. Their responses ranged from converting legacy worksheets to operational dashboards to improving existing dashboards in response to user needs.

LESSONS LEARNED

Valuable lessons were learned in the implementation of the project. The implementation team faced data infrastructure constraints as well as pressure to deliver quick results. To address this, the team opted for a pragmatic and flexible approach to exporting data from the HRMIS data warehouse. This meant simplifying extraction to a few data pipelines that would clean and load the HRMIS data to the HR BICC itself.

Another lesson was the importance of data quality and how to establish processes to protect it. The team defined a data glossary to establish a common understanding of expectations regarding data structure and shared this glossary with users. It also established data governance practices and quality checks to ensure the integrity of data fed into the HR BICC. The team implemented automated controls and routines for data entered and managed by HR departments and also conducted regular trainings and communication to increase awareness of data quality concerns.

The team also learned that standards and development guidelines improve user experience and accessibility. It designed uniform layouts, chart types, navigation practices, and colors, while documenting dashboard-development requirements. However, it also learned that end users should not be tasked with developing dashboards. Even with proper documentation, developing a dashboard is a complex task. Although BI tools can convey and promote a self-service approach, end users rarely master dashboard development without proper training. Different users may not follow the guidelines for building dashboards, resulting in heterogeneous dashboards.

A final lesson was that, while limiting the scope for end users, the dashboard development team has to remain flexible and respond to user needs. Responsibilities for the implementation team include developing new dashboards, modifying existing analyses, and generating reports. The team should consult with clients until dashboards meet end users' expectations. Finally, support systems for users are strongly recommended. A helpline proved particularly useful, with a service-desk phone number and an online form to receive and answer user questions and requests.

IMPACT OF THE SOLUTION

As a result of the project, the HR BICC provides a comprehensive and detailed view of HRMIS data across the government (ministries and administrations/agencies) of Luxembourg. It includes multiple dashboards to visualize HRMIS modules, such as career management and pensions (see figure 9.6). This dashboard ecosystem keeps growing. As of today, the HR BICC maintains over 56 streams containing 157 HR dashboards with over 2,600 sheets.[8] In addition, it hosts 320 active users with more than 20,000 connections per year.

The HR BICC accommodates a variety of use cases. Active users are, on the one hand, internal HR specialists for whom dashboards provide a new tool to monitor and verify HRMIS data. Other users include HR managers and members of HR teams within ministries and agencies. For these users, the dashboards offer a better overview of their own HR, better control over the key dates in their HR processes, and better follow-up on their personnel.

The overall benefits of such an approach are, for all users, a gain in the quality of HRMIS data and a clear and guided HR data journey. This journey ranges from a broad overview of the HRMIS to deep dives into a particular topic, such as compensation. One example of a key daily benefit is the use of aggregated trend data to project new HR initiatives, orientations, decision-making, and negotiation arguments at the ministry level. Additionally, the HR BICC provides users with accurate and fast information, accelerating business processes and decision-making. Because some of the dashboards are shared with decision-makers at the ministry level, it helps build, improve, and adapt laws and regulations. Overall, this also increases the data literacy of government organizations.

CONCLUSION

This case study has described how Luxembourg's CGPO developed an integrated dashboard system to inform policy making. The project included both the development of a dashboard ecosystem and the necessary data infrastructure to maintain it. The solution has grown considerably since its launch, hosting over

150 applications, each with its own set of dashboards. Together, these applications cover a variety of topics, from career management and pensions to recruitment.

The dashboard ecosystem has had a considerable impact on the way HRMIS data are consumed and analyzed. It provides immediate access to information on HR that allows policy makers to make better-informed decisions. It establishes quality controls and trains civil servants to better use the platform. Dashboards also increase data literacy within ministries and among HR specialists. However, it is important to note that the CGPO relies on an in-house team with experience in developing and deploying dashboards. This means that the project rollout and implementation were both fast and sustained over time. This experience contrasts with other cases, such as Brazil, more common in developing contexts, where solution maintenance and improvement were constrained by dependency on external actors.

The case study highlights the benefits of a systematic approach to HRMIS analytics, supported by a civil service with the capacity to implement and maintain it. Not all governments have access to these human capital resources. As a result, their dashboard ecosystems may require a more limited approach. Yet beyond the technical expertise, a valuable lesson can be learned from CGPO's methodical approach. The CGPO carefully developed a systematic array of protocols and documentation to protect the integrity of HRMIS data and dashboard visualizations. This requires not a group of IT experts but a careful consideration of the bureaucratic protocols necessary to both maintain and grow the solution. This approach could certainly be replicated in government agencies elsewhere.

CASE STUDY 9.2 HRMIS CASE STUDY: FEDERAL PAYROLL CONTROL AND COMPLIANCE (BRAZIL)

Luciana Andrade, Galileu Kim, and Matheus Soldi Hardt

SUMMARY

In 2019, a public-private partnership between a federal payroll auditing team and a consulting firm resulted in the development of a novel payroll irregularity detection system. The solution included an integrated data pipeline to train a statistical model to detect irregularities as well as automated identification of violations of payroll regulations. The fraud detection system was used to assist payroll auditors in their daily work. This complementary approach enabled auditors to better detect irregular payroll entries, increasing savings and improving efficiency.

INTRODUCTION

Governments are responsible for the accurate and timely disbursement of payroll to civil servants. As the volume and complexity of payroll increase, manual approaches to quality control are not sustainable.
In 2019, the Department of Compensation and Benefits (DEREB), a federal agency in Brazil, was responsible

Luciana Andrade is a senior regulatory agent with Anvisa. Galileu Kim is a research analyst in the World Bank's Development Impact Evaluation (DIME) Department. Matheus Soldi Hardt is a partner at EloGroup.

for overseeing over 80 million paychecks annually. To improve the process, DEREB introduced a new technology to support payroll analysts in their quality checks, which combined machine learning and automation. The Federal Payroll Digital Transformation project ultimately increased recovery rates on inconsistent paychecks and is used daily by payroll analysts in Brazil's federal government.

This case study describes how the project improved the workflow for control and compliance in payroll, a foundational module in a human resources management information system (HRMIS). Although the project had a narrow focus compared to the case of Luxembourg (case study 9.1), this limited scope enabled the development of a highly specialized solution to payroll management, analogous to the case of the United States (case study 9.3). This specialization allowed for the relatively quick and low-cost deployment of the solution. However, it also meant that the project was context specific and not necessarily scalable to other modules in the HRMIS.

Here are the key lessons from the case. First, the foundational steps of problem definition and scope were conducted through extensive dialogue with end users. Payroll analysts who would ultimately use the technology were consulted and offered input to the solution itself. Second, an iterative approach reduced risk aversion and secured buy-in from leadership in public administration. Because the payroll system was complex and the analysts themselves did not have complete knowledge of it, the team opted for gradual refinement of the solution. Finally, reliance on external actors allowed for rapid implementation, but due to this external reliance, the solution was not further developed once the intervention was finalized. In-house technical capacity was never built.

The case study is structured as follows. First, we provide institutional context about the federal payroll system. Section 2 outlines the solution. Section 3 highlights the rollout strategy for the solution. Section 4 describes risk aversion in bureaucratic organizations and how iterative disruption overcame it. Section 5 outlines the impact of the solution. Section 6 draws some lessons and cautionary observations about the external implementation of digital solutions. Finally, we conclude.

INSTITUTIONAL CONTEXT OF THE FEDERAL PAYROLL SYSTEM

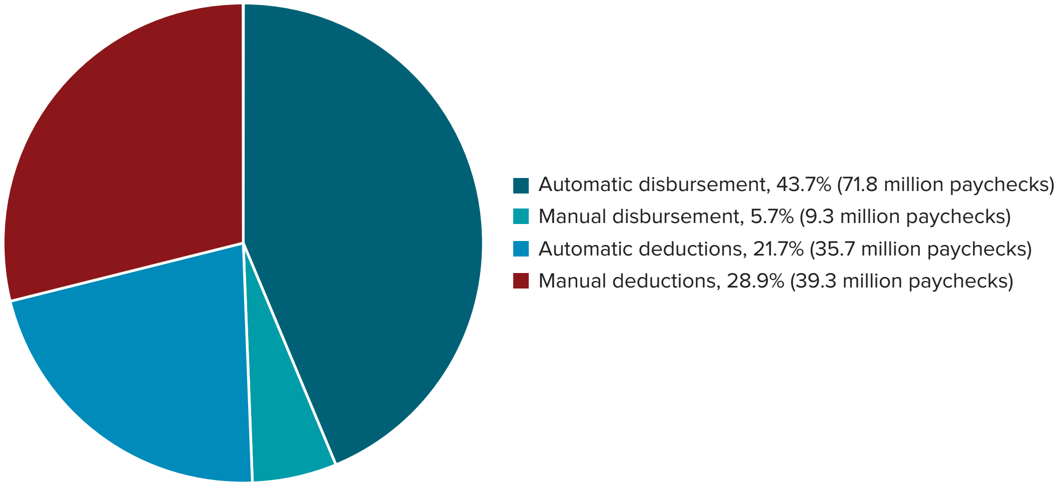
Brazil's federal government disburses over R$150 billion (US$30 billion) in the federal payroll every year, accounting for 1.4 percent of the national GDP in 2019. Of the total paychecks issued, over 43 percent are fully automated, meaning that payments are automatically disbursed according to pre-established rules and procedures (figure 9.9). However, 5.7 percent are still manually submitted entries, amounting to 9.3 million manual entries in 2018. While payroll data are centrally stored and managed by the Ministry of Finance, disbursement and deductions are submitted through claims by human resource (HR) departments in different federal agencies.

As noted in chapter 9, one of the foundational modules in an HRMIS is payroll compliance and control. In Brazil's federal government, payroll quality control is the responsibility of DEREB, which is overseen by the Department of Personnel Management and Performance (SGP). While it does not have the mandate to punish infractions, DEREB flags paycheck inconsistencies prior to disbursement, which must be addressed by HR departments in federal agencies.

The task is challenging. The case volume is large, with tens of thousands of individual disbursements transacted daily. Additionally, a complex set of regulations governs how payments should be disbursed. To enforce these rules and detect inconsistencies, a team of payroll analysts individually verify each paycheck. Over the course of a day, analysts check hundreds of entries to verify whether the values are in accordance with the existing rules, whether the amount issued is too high, and whether the public servant that would receive the value has the actual benefit, among other inconsistencies.

Before project implementation in 2019, payroll monitoring was done through a combination of selecting the highest-value paychecks and random sampling. At this stage, DEREB first determined the

FIGURE 9.9 Brazil's Federal Payroll, 2018



Source: Original figure for this publication.
Note: Payroll excludes the municipal government of Brasília (GDF) and state-owned enterprises.

number of manual entries to be verified based on the productivity of each payroll analyst multiplied by the number of payroll analysts working that day. DEREB would then select payroll entries according to the following rules: 90 percent of the sample was selected from the highest-value entries and the remaining 10 percent was randomly selected. This approach was designed to reduce workload and maximize fund recovery since large entries were overrepresented in the sample.

Although this legacy approach represented an initial attempt to automate the sampling of entries for monitoring, it identified few inconsistencies. In total, only 2 percent of entries were notified for corrections, and of those, 40 percent were corrected. In total, inconsistencies that represented less than R$10 million per year were corrected, less than 0.1 percent of the total amount disbursed by the federal payroll. Management at DEREB wanted to improve this process and opted for an HRMIS reform project in collaboration with a consulting firm.

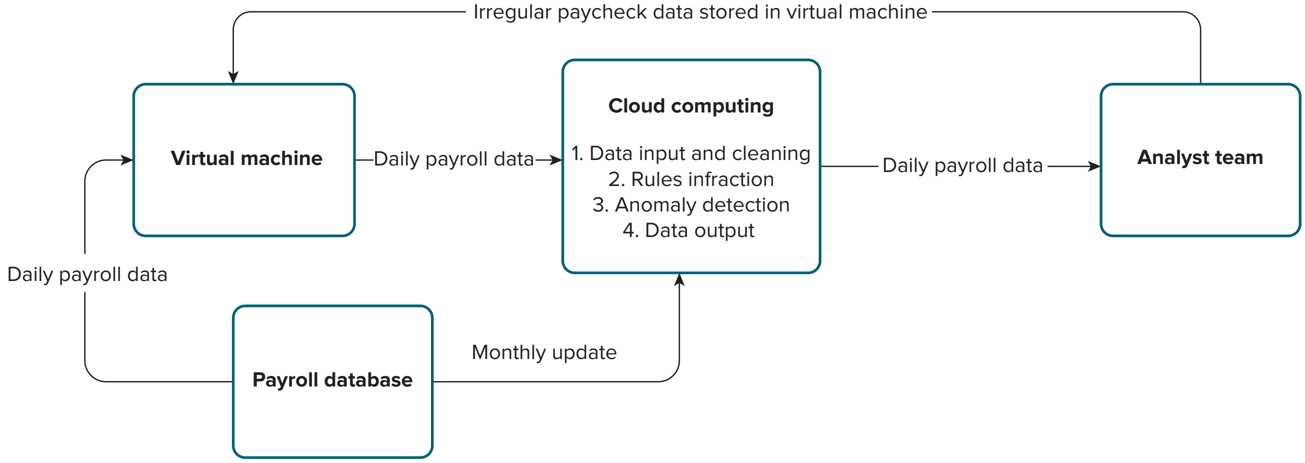
THE SOLUTION: FEDERAL PAYROLL DIGITAL TRANSFORMATION

The Federal Payroll Digital Transformation project changed the workflow for payroll quality control through the implementation of new technologies. The project was a public-private partnership between DEREB and the consulting firm EloGroup. At its core, the solution generated flags and rankings for federal payroll analysts in their effort to detect and notify agencies of potential inconsistencies in their payrolls. The solution was open source and deployed through cloud technology. The development cycle took approximately eight months to complete.

The solution relies on two complementary approaches: qualitative flagging of regulations governing payroll and quantitative analysis through anomaly-detection statistics. The development of the business-rules module relied on translating regulations governing payroll into automated flags indicating whether an infraction has occurred. The quantitative approach adopts statistical techniques developed by credit card companies to detect anomalies in payments. Payroll values that are far off from a predicted value are assigned a greater risk score and prioritized for payroll analysts.

The solution is executed daily. The first step in the pipeline is the extraction of data on paychecks created in the previous working day, reduced to the subset of manually imputed disbursements (figure 9.10). The data

FIGURE 9.10 Brazil's Solution Workflow



Source: Original figure for this publication.

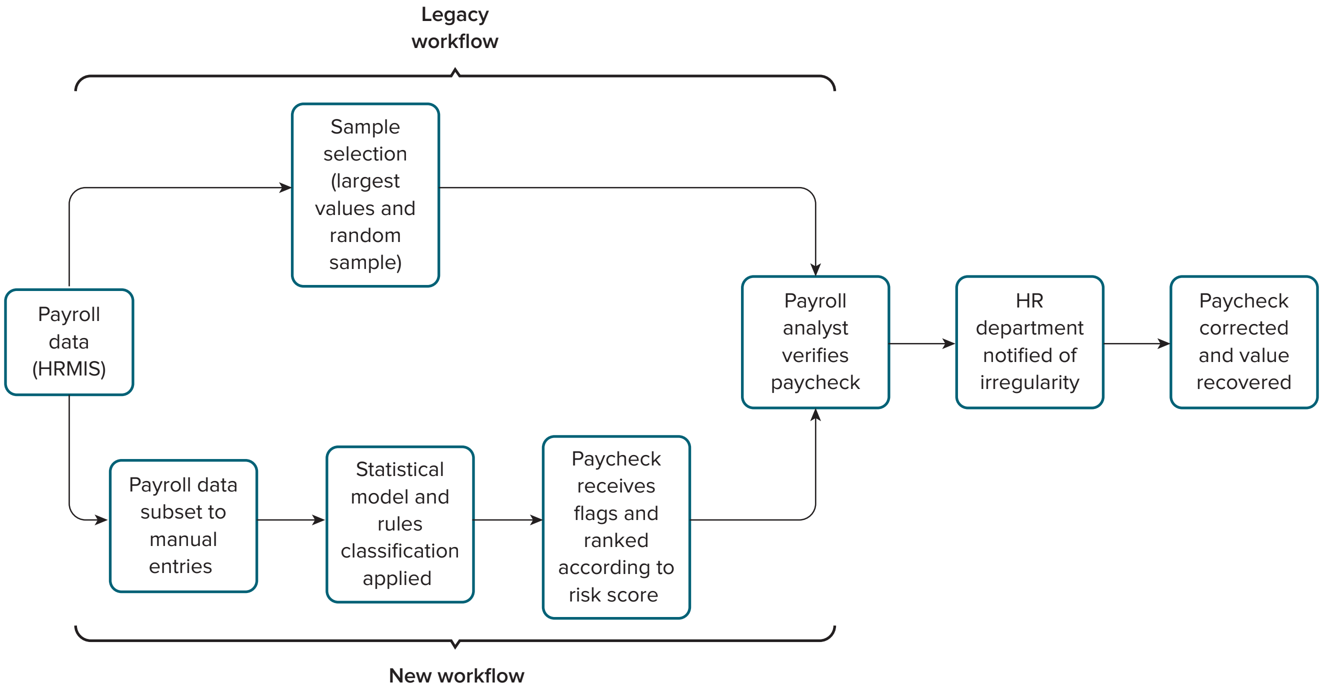
are fed directly from the payroll database into a virtual machine (VM), which receives and stores the daily payroll data. The data are then transferred to a computing cluster in the cloud, where a set of tasks is performed. The data are first cleaned and then go through a rules-infraction module, where they are flagged for potential violations. For example, one rule may be that civil servants are not allowed to claim over 1,000 reais in reimbursement for travel expenses. If the rules-infraction module detects claims that exceed that threshold, it would flag that paycheck and send it directly to the payroll analyst team, indicating that this rule has been violated. If no rule infractions are detected, the paycheck is fed into a machine-learning model that classifies paychecks as anomalous, attributing to them a risk score.

Once the business rules and the statistical model classification are applied, paychecks that are considered most likely to be inconsistent are ranked first and sent to the analyst team. The format in which the data are exported is a simple worksheet, with predetermined labels identifying the risk score and the rule-infraction flags, as well as usual paycheck fields, such as issuing agency and beneficiary. Payroll analysts have discretion over which paychecks to verify and can rank paychecks according to priority, regardless of the classification exercise. It is only at this stage that paychecks are verified and flagged for additional verification by the issuing agencies. Note that the decision to issue a flag remains under the jurisdiction of the analyst.

As a result, the workflow from the analyst's perspective has not changed significantly. The value added is curated information for the analyst, through automated rule-compliance flags and risk scores to facilitate the analyst's decision-making process. Each step in the solution workflow outlined in figure 9.10 is an additional layer of verification, which transparently encodes how the data are cleaned and classified before reaching the analyst's visual dashboard. This choice of design was agreed upon by the monitoring team and the data science team, who opted to make insights from the solution accessible and easy to use. Figure 9.11 compares the new approach with the legacy one.

The machine-learning model and the rules classification do not replace the monitoring team—rather, they enhance its workflow by automating procedures before the data even reach the individual analyst. This complementarity between analog and digital processes is what enabled the new workflow to be well received and adopted by analysts, in contrast to other experiences of technological innovation in which human decisions are eliminated. This hybrid solution provides a more gradual approach toward the goal of digital transformation, accommodating the need for preserving human autonomy while increasing humans' productivity through the use of technology.

FIGURE 9.11 Comparison between Brazil's Legacy and New Payroll Workflows


Legacy workflow
Sample selection (largest values and random sample)
Payroll data (HRMIS)
Payroll analyst verifies paycheck
HR department notified of irregularity
Paycheck corrected and value recovered
Payroll data subset to manual entries
Statistical model and rules classification applied
Paycheck receives flags and ranked according to risk score
New workflow

Source: Original figure for this publication.
Note: HR = human resources; HRMIS = human resources management information system.

ROLLOUT STRATEGY AND SEQUENCING

The director of DEREB decided to improve the existing monitoring system by leveraging the use of digital technologies. Given the agency's capacity constraints and lack of familiarity with technological innovation, the director outsourced the implementation and rollout strategy for the solution to an external consulting firm. The initial legal groundwork was crucial. The director of the consulting firm EloGroup leveraged its experience in the development of digital technologies for other government agencies and guided the drafting of the proposal. The General Coordinator for Special Projects of the Secretariat of Public Policies for Employment was familiar with the regulatory process and provided guidance on obtaining legal approval and initial funding for the solution.

The political environment was favorable for the project. Senior leadership was committed to fighting inefficiency and possible cases of corruption, and the federal payroll was under scrutiny due to its large size and perceived inefficiency. The SGP leadership team gave wide discretion to DEREB regarding the HRMIS reform to be enacted. This autonomy allowed the director of DEREB to make difficult decisions regarding personnel, who initially resisted modifying the existing monitoring process. To obtain funding for the project, the team submitted a project proposal to a technology company that provided seed funding for the project.

The monitoring system was developed by a small but agile team of technology consultants at the consulting firm EloGroup. The initial goal was to design a prototype of the workflow outlined in figure 9.10 to detect inconsistencies that would validate the approach. An intensive consultation process preceded the implementation of the technical solution. Workshops and open discussions with federal agencies highlighted what data would be available to develop the prototype, what unique identifiers there were for merging the data, and what kinds of variables would be available to the machine-learning algorithm. An initial workshop covered problem definition and project scoping, defining how the solution would be embedded into the monitoring tasks performed by the auditors.

Once the project was launched, it faced resistance from staff. Personnel within the monitoring team at DEREB expressed concern regarding the proposed solution because they feared displacement and the disruption of existing procedures. Staff also worried that the digital update would lead to a technological dead end, as had occurred in previous collaborations with external consulting firms. Anecdotally, there was a perception among participating Brazilian public servants that private initiatives introduced off-the-shelf solutions without considering the needs or opinions of public servants who had worked for years in the area.

A collaborative design aimed to assuage these concerns. During the kickoff workshop with multiple federal agencies, staff from different areas within DEREB were able to express their views on the flaws and strengths of the payroll system. On more than one occasion, a public servant in one area identified that his challenge was shared across departments. These open conversations made even the most reluctant employees of the project express interest, or at least not boycott the initiative. In making these concerns transparent and sharing them in an open forum, the team included payroll analysts in the development of the project. Obtaining buy-in within and across departments proved crucial to the success and sustainability of the solution.

Buy-in was necessary not only for personnel but for upper management as well. Due to budget constraints, Brazil's federal bureaucracy had only limited access to cloud resources, for which agencies needed to petition. As a result, after the initial seed funding was spent, it was necessary to secure access to cloud computing through a formal project proposal. To do this, the team presented the results of the initial stage of the solution, highlighting the benefits of the approach and how it could assist the government in saving money. This effort was ultimately successful, securing additional funding to complete the solution.

RISK AVERSION AND ITERATIVE DISRUPTION

Bureaucratic agencies are risk averse, and with good reason: they perform key roles in government and, while doing so, comply with rules and regulations. A task executed improperly or failure to abide by existing norms can have severe consequences, both for the general functioning of the state apparatus and for the individual careers of civil servants. The solution for this project was not to revamp the regulatory framework or standard operations. Instead, the reform team identified small opportunities to improve the workflow of the analyst team through multiple cycles of disruption.

Coordination was key to this approach. The consulting team was responsible for implementing the solution in terms of software and data engineering. Meanwhile, the payroll analysts and the management team at DEREB provided feedback and prototyped beta versions of the solution. To strengthen this partnership, communication channels between both teams were reinforced. The method deployed for the development of the solution was short but agile.

One of the main challenges in implementing the solution was a mutual lack of knowledge between DEREB and EloGroup regarding the other's area of expertise. For the consulting team, the payroll data and governance structures of Brazil's federal bureaucracy were so complex that most of their initial effort focused on learning how the payroll system operated. To address this, the consulting team had to communicate extensively with the monitoring team at DEREB to ensure that relevant data were extracted and that regulations were incorporated into the automated rules and statistical model.

On the other hand, the monitoring team at DEREB had limited exposure to statistics and software development and therefore needed to be introduced to novel techniques without prior knowledge. Conversations revolved around how to formalize the substantive knowledge of analysts in software, but ultimately, analysts had to rely on the consulting team to implement the solution. Lack of familiarity with software development and the platform meant that when bugs in the operations were identified, the consulting team had to address them, and workflow was interrupted.

With the initial data pipeline designed, the business rules and the statistical model were put into production. Anomalous paychecks were sent directly to the monitoring team for validation. The initial results were positive, with the algorithm-empowered monitoring consistently outperforming the previous approach, based on the size of paychecks. As additional resources were necessary to expand the project, the director of DEREB presented the results to government leadership as promising evidence that the approach was correct. This initial buy-in proved key: having an actual solution in production and demonstrating results reduced uncertainty in higher levels of management.

The deployed solution combines two key insights: first, it formalizes existing laws and regulations governing payments in an automated pipeline. This means that the analyst no longer has to verify whether a paycheck complies with regulations; the business-rules module does this automatically. Second, the anomaly-detection algorithm relies on statistical modeling to leverage information about public servants, their departments, and their payment histories. This process fully leverages the information methodically collected by the Brazilian government on its payroll and public servants without imposing additional burdens on the analyst team.

Additionally, the current algorithm is designed to reduce workload and help analysts prioritize paychecks with higher risk. This complementary approach to improving payroll analysts' workflow is key: after initial resistance regarding these changes, the monitoring team realized the benefits of the new digital approach over previous approaches. This hybrid model, incorporating both analog and digital processes, can provide a template for public sector technological innovations.

IMPACT OF THE SOLUTION

The clearest gains from the solution were in efficiency: despite the reduction in personnel, performance increased. Due to staff attrition unrelated to the project, the team of payroll analysts had been reduced in size. Despite this reduction, the reduced analyst team could flag the same amount of resources as inconsistent compared to a larger team, while dedicating less time to each task. This reduction in the cost and maintenance of performance was an important selling point to other departments within the federal bureaucracy, highlighting the gains in efficiency from technological innovation.

An unintended consequence of the project was an increase in data literacy and a change in mindset. Users of the dashboard displayed greater interest in learning how the solution was implemented, with analysts expressing willingness to learn how to code to better understand the data. This growth in data literacy resulted from initial exposure to a set of techniques that had not been available before. Additionally, because of data integration, new linkages were formed between DEREB and other departments in the bureaucracy. Because the solution relied on data generated in other departments, there was a need for communication and transparency to make it work.

Finally, there was a shift in mindset regarding how to monitor payrolls. While previously, analysts had relied on their accumulated experience and intuition, the solution complemented this approach by emphasizing the use of data and regulatory infractions. The analytical framework of the solution provided a new template that analysts could use to assess whether a paycheck was indeed inconsistent. In a sense, the new technology changed the way payroll analysts approached their task.

SUSTAINABILITY OF EXTERNAL IMPLEMENTATION

External solutions are brittle. They introduce dependency on the technical know-how of external actors, and once the engagement is finalized, the beneficiary is no longer able to maintain or improve on the external solution. In this case, technical know-how—including software and data engineering—for the implementation of the project remained with the consulting team once it left. The analyst team at DEREB did not acquire the necessary skills or capacity to develop the solution further, even though it was open source. Although data literacy in the monitoring team increased, the analyst team was not formally trained to modify or further develop the software.

Additionally, changes in the management structure of DEREB after the implementation of the technical solution put the sustainability and continued development of the project at risk. While the previous director locked in the current version of the solution, it has not evolved since. Turnover in management and a contract-based approach meant that desirable additions to the solution—such as the extension of automation to all HR departments across federal agencies—were never implemented. The loss of institutional leadership and the lack of in-house capacity meant that while the product survived, it did not continue evolving.

CONCLUSION

Technological innovation is disruptive, but the costs and uncertainty associated with it can be reduced by adopting a gradual approach. Risk aversion—an important feature of bureaucracies—can be overcome through communication and small modifications to existing workflows. The Federal Payroll Digital Transformation project outlined in this case study showcases this approach. Instead of a complete transformation of the payroll monitoring process, the technology focused on complementing existing workflows by payroll analysts.

A collaborative approach helped build trust in the relevance of the solution and its applicability to daily operations by end users. Iterative cycles of feedback and adaptation ensured that the algorithm proposed was appropriate to the use case and understood by payroll analysts. In addition, this reduced resistance to the final adoption of the solution. Technological disruption can thus be managed and incorporated into existing procedures, giving rise to hybrid solutions that provide a stepping stone for more extensive and intensive solutions.

While the current version of the solution has been finalized, its future development is uncertain. Due to the project's outsourcing, the necessary expertise to implement and develop the solution was not developed in-house. Technological innovation through a public-private partnership therefore comes with associated costs and benefits. There is a trade-off between the agility and rapid gains from outsourcing to external agents and the lack of development of in-house expertise to continue growing solutions. External solutions therefore generate dependency on external actors for developing solutions, lowering the likelihood of maintenance and expansion in the long run.

Finally, the implementation team has emphasized the need for spaces within public administration to incubate technological innovation. These spaces would allow for calculated risks—and mistakes—within the public sector. While the team identified and opened spaces within which the solution could grow, it is important to ensure that those spaces are already set in place. This would incentivize not only managers willing to lead innovations but also staff members, who would prove more willing to engage in changes without fear of reprisal. It would also create incentives for agencies to develop the in-house capacity for technological innovation and reduce dependence on external actors.

CASE STUDY 9.3 HRMIS CASE STUDY: EMPLOYEE VIEWPOINT SURVEY ANALYSIS AND RESULTS TOOL (UNITED STATES)

Camille Hoover and Robin Klevins

SUMMARY

In 2015, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), within the National Institutes of Health (NIH), developed the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) to extract insights from the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS). The solution relied on the creative use of worksheet software to extract and transform data to produce dashboards automatically from a single data file. Effectively, the worksheet developed by the NIDDK team integrated a data infrastructure and a dashboard into a single platform, reducing implementation costs. The tool provides valuable information for senior leadership to promote employee engagement and guide policy making.

INTRODUCTION

It is a leader's responsibility to care for the people in an organization and to create and sustain a culture where employees can flourish—one in which performance is central and employee engagement is maintained. To be successful, these values must be integrated into the function and mission of the organization, not treated as distinct or separate. To create this type of culture, leadership must secure buy-in from staff at all levels. Staff must embrace the organization's vision and emulate its core values.

It is important that the core values not just be lofty or aspirational goals but translate into action on the frontlines, where the people of the organization are doing the work. Values can and should be measured through employee engagement surveys. This measurement allows leaders to keep a finger on the organization's pulse. It is important to combine data analytics with the voices of employees to inform strategies and resource allocation and to verify whether actions are paying off. Employee feedback must inform and orient action, whether in the form of focus groups, town halls, stay or exit interviews, or crowdsourcing.

This case study describes how the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) developed an analytics dashboard to measure and promote employee engagement. The project was named the Employee Viewpoint Survey Analysis and Results Tool (EVS ART). EVS ART provided NIDDK leadership with immediate and informative data analytics on their employees' perceptions of whether, and to what extent, conditions characterizing a successful organization were present in their agencies. Using EVS ART, the NIDDK was able to transform the enormous amount of data provided by the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) into a user-friendly format in mere minutes. The survey topics, in response to which employees candidly shared their perceptions about their work experience, organization, and leaders, covered employee engagement, employee satisfaction,

Camille Hoover is an executive officer at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Robin Klevins is a senior management analyst at the NIDDK.

and several submeasures, including policies and practices, rewards and recognition, opportunities for professional development, and diversity and inclusion—all of which were used to inform decision-making.

EVS ART is an example of human resources management information system (HRMIS) analytics, similar in purpose to the case study of Luxembourg (case study 9.1). However, in contrast to Luxembourg, which generated analytical insights on the entirety of its HRMIS, this case focuses on the employee engagement module within an HRMIS. This module is diverse and expansive as a result of the rich data provided by the FEVS. The FEVS measures employees' perceptions of whether, and to what extent, conditions characteristic of successful organizations are present in their agencies. It is a survey in which employees can candidly share their perceptions about their work experiences, organizations, and leaders. EVS ART therefore includes indicators on employee satisfaction, global satisfaction, compensation, and organization, as well as more customized questions about remote work and COVID-19. The focus on improving a particular module of an HRMIS makes this case similar to the approach in Brazil (case study 9.2), which reformed how the payroll module operated.

The project provided a set of lessons that may be helpful for practitioners. First, keep the solution simple. While the inner workings of a tool itself may venture over to the complex side, do not make the act of engaging with the analysis complex for the user. Second, make the solution accessible to all types of users. This means two things. One, ensure that the tool is accessible to those with disabilities, and two, make the tool available to the broadest audience possible. If people do not know about the tool, they will continue to spend unnecessary time recreating analyses and will not obtain insights from the data. Finally, remember that transparency ensures that data analytics can be trusted by those it benefits**.** When working with data, leadership should not shy away from difficult conversations, because survey takers already know whether something is working well or not. It is incumbent on leadership to be honest, dig deeper, and let staff know that their input will drive organizational change.

This case study is structured as follows. We first describe the institutional context, with particular attention to the FEVS, the largest civil servant engagement survey in the United States. Section 2 explains the initial rollout of the solution. Section 3 provides a detailed overview of the solution. Section 4 outlines the lessons learned during the implementation of the project. Section 5 describes the impact of the solution. Section 6 reflects critically on the importance of looking beyond analytics and effectively promoting change. Section 7 reviews challenges faced and future improvements to EVS ART. Finally, we conclude.

INSTITUTIONAL CONTEXT: THE FEVS

Each year, the OPM administers the FEVS to over 1.4 million full- and part-time permanent, nonseasonal employees governmentwide.[9] The FEVS measures employee engagement, including employees' perceptions of whether, and to what extent, conditions characterizing successful organizations are present in their agencies. It therefore provides valuable insight into agencies' strengths and opportunities for improvement. In 2020, 44.3 percent (624,800) of those receiving the FEVS completed it—each spending, on average, 25 minutes to do so (OPM 2021). This translates to over 260,000 federal hours and equates to over US$10 million worth of staff time taking the survey.[10]

The FEVS provides valuable information because the OPM proactively designed the FEVS to include multiple index measures and key categories, such as employee engagement and satisfaction, to help agencies identify important patterns and themes.[11] Each index is valuable, aggregating multiple answers.[12] While much can be learned from the index measures and key categories, on average, there is a three- to four-month period during which the OPM processes the raw data before distributing it to agencies.

The FEVS allows agencies to obtain valuable feedback from all levels of the organization. Subgroups within an agency that have 10 or more survey participants can receive their own area-specific results, and those with fewer than 10 participants roll up to the next level of report to ensure at least 10 responses. This protects the confidentiality of the survey respondent, which is crucial when the goal is to obtain honest feedback (NIH 2018). In 2018, over 28,000 organizations within the federal government had 10 or more survey participants, for a total of over 280,000 survey respondents—and the number continues to grow (Kamensky 2019).

The FEVS's granular and large-scale data allow organizational leaders within the federal government to tap into the perspective of those on the frontlines and learn from the voices of employees. In turn, the same information can be used to design employee-informed programs and initiatives. It is important for staff to be made aware of changes informed by their feedback. Informed change is noticed, creates ownership, and leads to increased engagement—and engagement is the foundation on which successful missions are built.

Despite this valuable information, extracting insights from the FEVS and putting them into action is challenging. Once given access to the survey, government agencies spend weeks culling large amounts of data to operationalize the survey's feedback. This effort is extremely labor intensive, time-consuming, and costly. Some agencies spend thousands of dollars on manpower or on procuring outside support to analyze the data. In addition, by the time the results are received and the analysis completed, agencies are often on the heels of the next survey—with little time to act on the feedback provided. It is difficult to launch meaningful initiatives with old data, and the lack of timely action, or perceived inaction, often leaves employees wondering whether taking the survey is of value.

INITIAL ROLLOUT

A small team at the NIDDK, within the National Institutes of Health (NIH), took it upon themselves to work with the data and create a framework to deliver results quickly, accurately, and intuitively. The NIDDK's senior leaders appreciated the importance of these data and made it the highest priority to construct a way to translate them. They fully supported the NIDDK team's efforts—giving them time, flexibility, and necessary resources.

The NIDDK team set out to design a tool that brought to life the voice of the people, one that was unlike other tools. As analysts, they wanted to ensure that users could arrive at actionable data quickly. However, they approached it differently from a traditional report. It was important that the tool was easy to look at, that the flow of information made sense, and that it told a story. They also wanted to ensure that actionable target areas—and themes—jumped out at the user. It was of great importance that the tool be both easy to use and accessible to all federal employees.

The team worked for two years to create a tool that would enable leaders to drill down and compare data, have a better pulse on engagement levels, and view FEVS scores in an actionable and targeted way. They began by utilizing a resource that they already had at their fingertips, a common program used across the federal government: Microsoft Excel. The team worked to design an easy-to-use template that provided a report with an easy-to-understand flow, and they ensured that the templates were password protected so that links could not be broken and results would not be compromised. The team also worked to ensure that the tools and associated resources followed the guidelines of Section 508 of the Rehabilitation Act.[13]

OVERVIEW OF THE SOLUTION

The team created the EVS ART—an Excel-based tool that allows users simply to copy data provided by the OPM and paste them into a similarly formatted template. Upon clicking "Refresh," users can review conditionally formatted results, thoroughly compare prior years' data, and conduct a deeper-dive analysis of their outcomes.

EVS ART is different from other tools available to analyze FEVS data because users can arrive at actionable data quickly: the tool and output are easy to look at, the flow is intuitive, and the tool tells a story in a way that allows actionable target areas—and themes—to jump out. It is designed to be easy to use: it requires only basic Excel knowledge, it generates a user-friendly dashboard, and it captures and displays all OPM index measures and key categories.

The tool's utility lies in its simplicity of use but power in transforming massive amounts of information, allowing leaders to home in on important themes and compare prior years' data. EVS ART was designed so this can all be done in a few steps and as little as five minutes. EVS ART pulls data points from each of the main themes in the FEVS, such as employee engagement and global satisfaction. The tool organizes the survey results based on those themes by agency, subcomponent, and office, and it shows the change in responses for a specific item from year to year. This allows NIDDK senior leaders to monitor progress and evaluate the impact of strategies and interventions.

Instructions Tab

The first tab in EVS ART is the instructions tab (figure 9.12). Users enter the organization acronyms for the areas they wish to analyze and the year(s) of the results they wish to use. This information will automatically populate the headers and table titles on tabs throughout the Excel workbook.

Using FEVS data provided by the OPM, users copy and paste the information from their original FEVS data report into the corresponding EVS ART tab. No reformatting is required. This is done for each organization being compared. If prior year data are available, this step is repeated by pasting the data into the appropriate prior year tab(s). When this is completed, the user refreshes the data and EVS ART automatically populates the dashboard itself.

Dashboard Design

Upon feeding the data to EVS ART, users gain access to a dashboard that provides an overarching view of the organization's results. The dashboard delivers top-scoring questions for "positive," "neutral," and "negative" results, as well as the largest positive and negative shifts from one year to the next (figure 9.13). Below the charts, users are provided with a heat map that shows the average scores for each of the index measures and key categories, as well as their subcategories. This is helpful because it provides a clear visual at a high level and allows users to easily compare one organization to another.

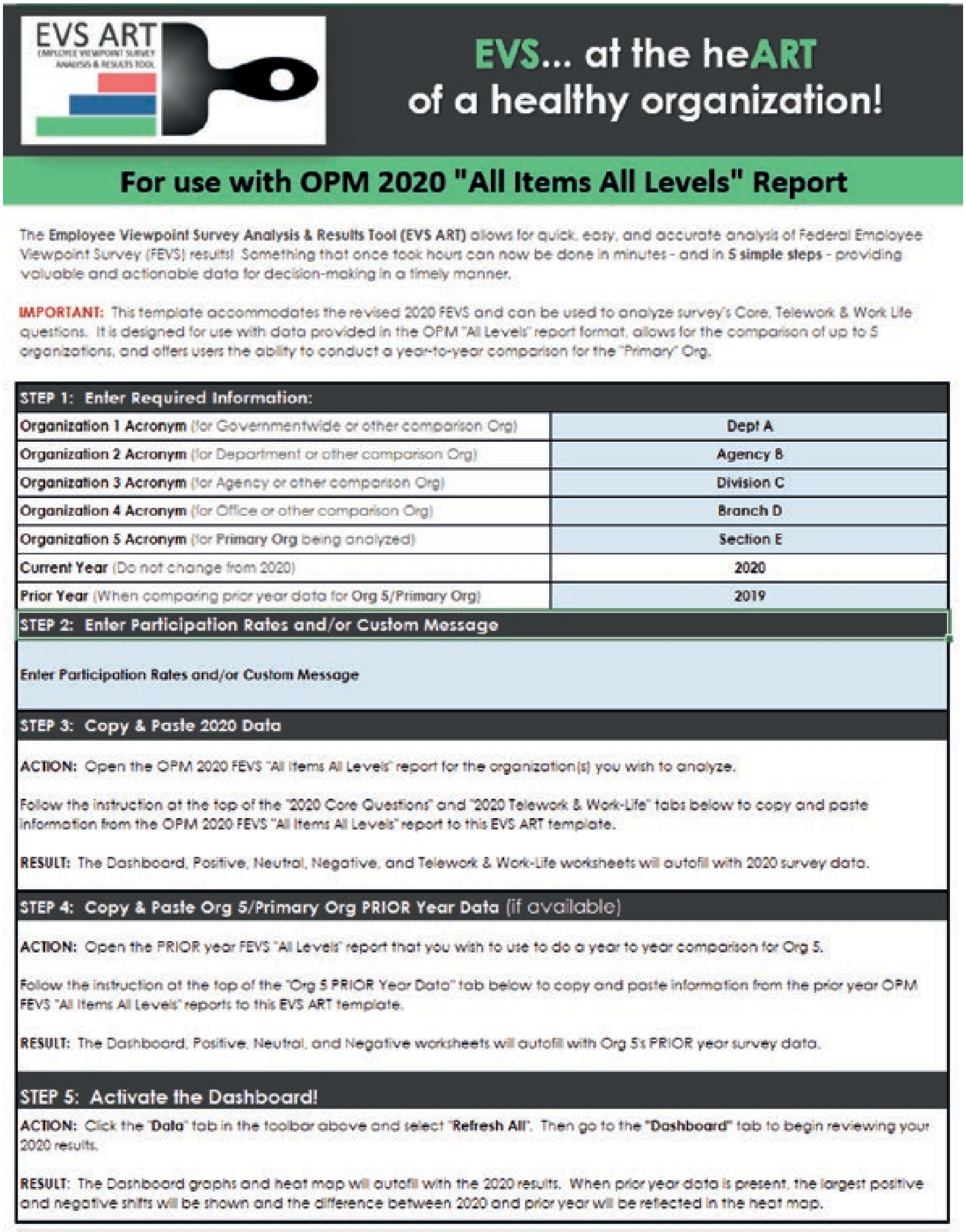
The dashboard also provides a side-by-side visual comparison of FEVS results (figure 9.14). This helps users to determine areas of focus across the organization and identify areas that need more targeted intervention. The conditionally formatted heat-map feature uses color to show managers their highest and lowest scores and identifies areas that might be strengths or challenges for the agency or a specific office. While the dashboard shows where to start looking, the information behind it—in the remainder of the report—provides a path that intuitively narrows the broader topics down to specific focus areas.

Analysis Tabs

While the dashboard is a great place to start, the deeper-dive portion of the report takes the user from a general overview to more specific focus areas, where the organization's scores begin to tell a story. Figure 9.15 shows an example of an organization's percent-positive employee engagement index scores. At the top of the tab is the OPM's guidance for interpreting the results. In the case of the FEVS,

- Questions averaging 65 percent positive or higher are considered "strengths,"
- Questions averaging 50 percent neutral or higher may indicate "opportunities" for improved communication, and
- Questions averaging lower than 50 percent are considered "challenges."

FIGURE 9.12 Instructions Tab in the EVS ART



EVS ART

EMPLOYEE VIEWPOINT SURVEY ANALYSIS & RESULTS TOOL

EVS... at the heART of a healthy organization!

For use with OPM 2020 "All Items All Levels" Report

The **Employee Viewpoint Survey Analysis & Results Tool (EVS ART)** allows for quick, easy, and accurate analysis of Federal Employee Viewpoint Survey (FEVS) results! Something that once took hours can now be done in minutes - and in **5 simple steps** - providing valuable and actionable data for decision-making in a timely manner.

IMPORTANT: This template accommodates the revised 2020 FEVS and can be used to analyze survey's Core, Telework & Work Life questions. It is designed for use with data provided in the OPM "All Levels" report format, allows for the comparison of up to 5 organizations, and offers users the ability to conduct a year-to-year comparison for the "Primary" Org.

STEP 1: Enter Required Information:	
Organization 1 Acronym (for Governmentwide or other comparison Org)	Dept A
Organization 2 Acronym (for Department or other comparison Org)	Agency B
Organization 3 Acronym (for Agency or other comparison Org)	Division C
Organization 4 Acronym (for Office or other comparison Org)	Branch D
Organization 5 Acronym (for **Primary Org** being analyzed)	Section E
Current Year (Do not change from 2020)	2020
Prior Year (When comparing prior year data for **Org 5/Primary Org**)	2019

STEP 2: Enter Participation Rates and/or Custom Message

Enter Participation Rates and/or Custom Message

STEP 3: Copy & Paste 2020 Data

ACTION: Open the OPM 2020 FEVS "All Items All Levels" report for the organization(s) you wish to analyze.

Follow the instruction at the top of the "2020 Core Questions" and "2020 Telework & Work-Life" tabs below to copy and paste information from the OPM 2020 FEVS "All Items All Levels" report to this EVS ART template.

RESULT: The Dashboard, Positive, Neutral, Negative, and Telework & Work-Life worksheets will autofill with 2020 survey data.

STEP 4: Copy & Paste Org 5/Primary Org PRIOR Year Data (if available)

ACTION: Open the PRIOR year FEVS "All Levels" report that you wish to use to do a year to year comparison for Org 5.

Follow the instruction at the top of the "Org 5 PRIOR Year Data" tab below to copy and paste information from the prior year OPM FEVS "All Items All Levels" reports to this EVS ART template.

RESULT: The Dashboard, Positive, Neutral, and Negative worksheets will autofill with Org 5's PRIOR year survey data.

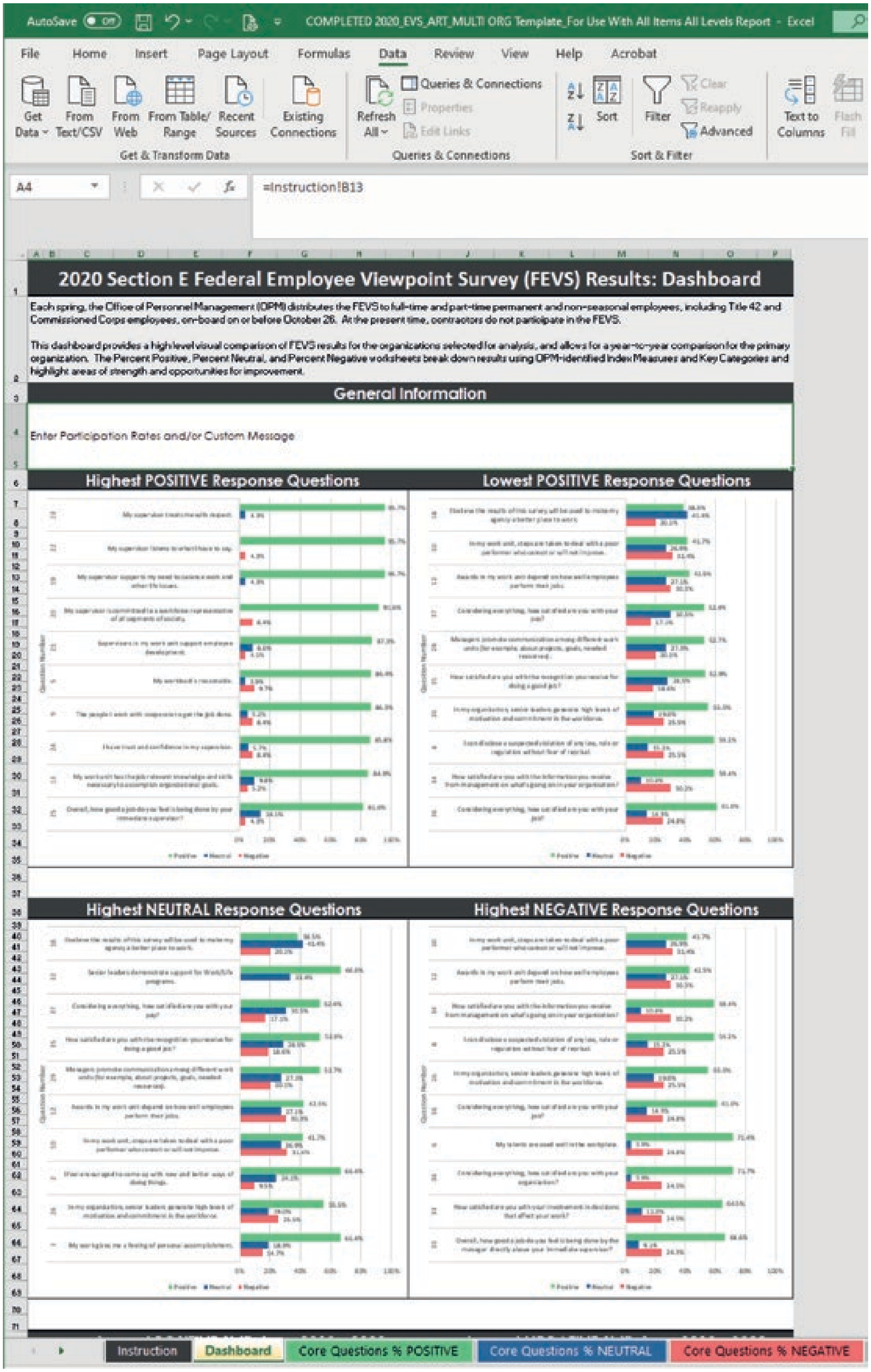
STEP 5: Activate the Dashboard!

ACTION: Click the "**Data**" tab in the toolbar above and select "**Refresh All**". Then go to the "**Dashboard**" tab to begin reviewing your 2020 results.

RESULT: The Dashboard graphs and heat map will autofill with the 2020 results. When prior year data is present, the largest positive and negative shifts will be shown and the difference between 2020 and prior year will be reflected in the heat map.

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

FIGURE 9.13 Landing Page of the EVS ART Dashboard



=Instruction!B13

2020 Section E Federal Employee Viewpoint Survey (FEVS) Results: Dashboard

Each spring, the Office of Personnel Management (OPM) distributes the FEVS to full-time and part-time permanent and non-seasonal employees, including Title 42 and Commissioned Corps employees, on-board on or before October 26. At the present time, contractors do not participate in the FEVS.

This dashboard provides a high level visual comparison of FEVS results for the organizations selected for analysis, and allows for a year-to-year comparison for the primary organization. The Percent Positive, Percent Neutral, and Percent Negative worksheets break down results using OPM-identified Index Measures and Key Categories and highlight areas of strength and opportunities for improvement.

General Information

Enter Participation Rates and/or Custom Message

Highest POSITIVE Response Questions

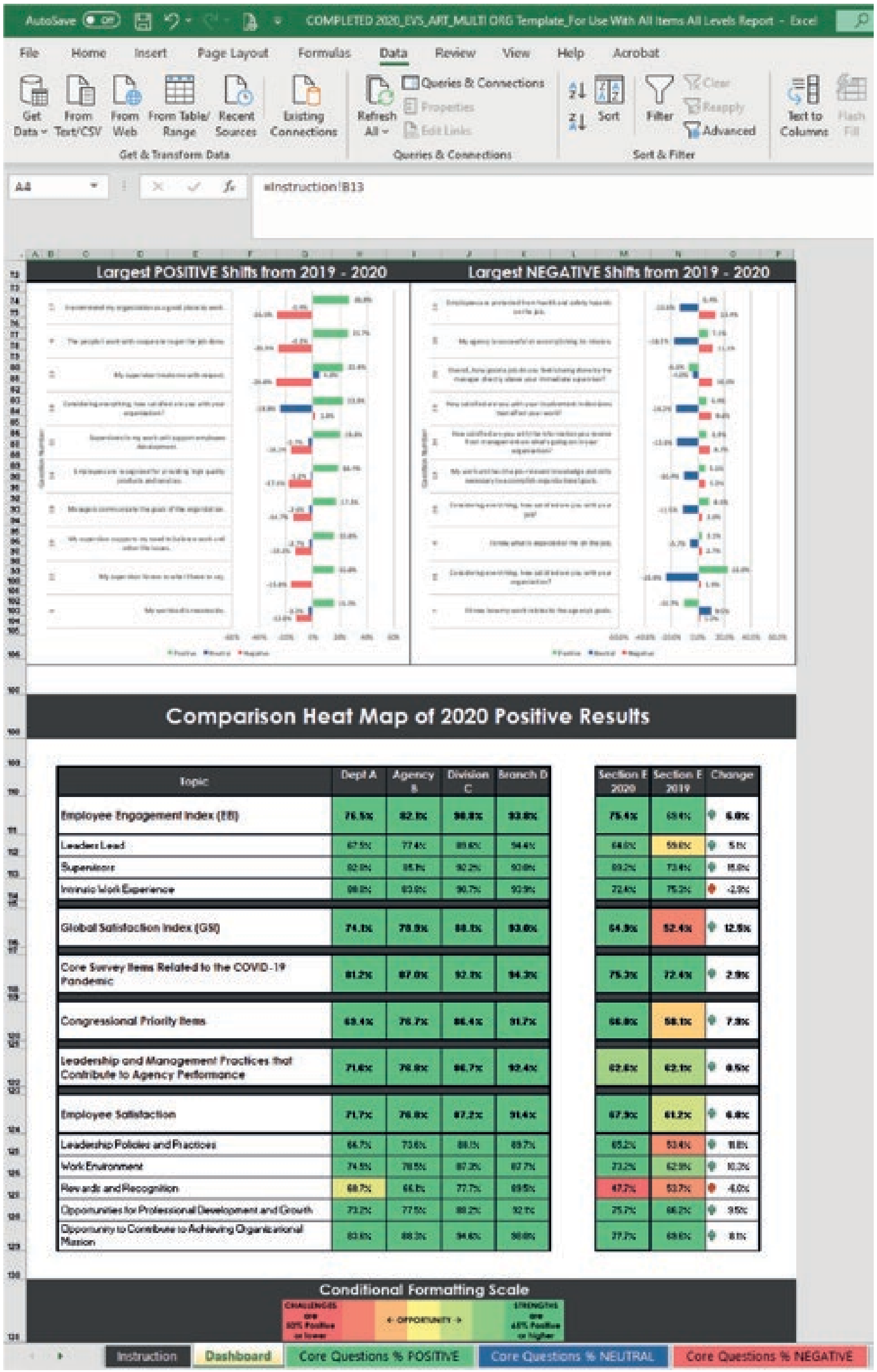
Lowest POSITIVE Response Questions

Highest NEUTRAL Response Questions

Highest NEGATIVE Response Questions

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

FIGURE 9.14 Results Comparison in the EVS ART Dashboard



=Instruction!B13

Largest POSITIVE Shifts from 2019 - 2020

Largest NEGATIVE Shifts from 2019 - 2020

Comparison Heat Map of 2020 Positive Results

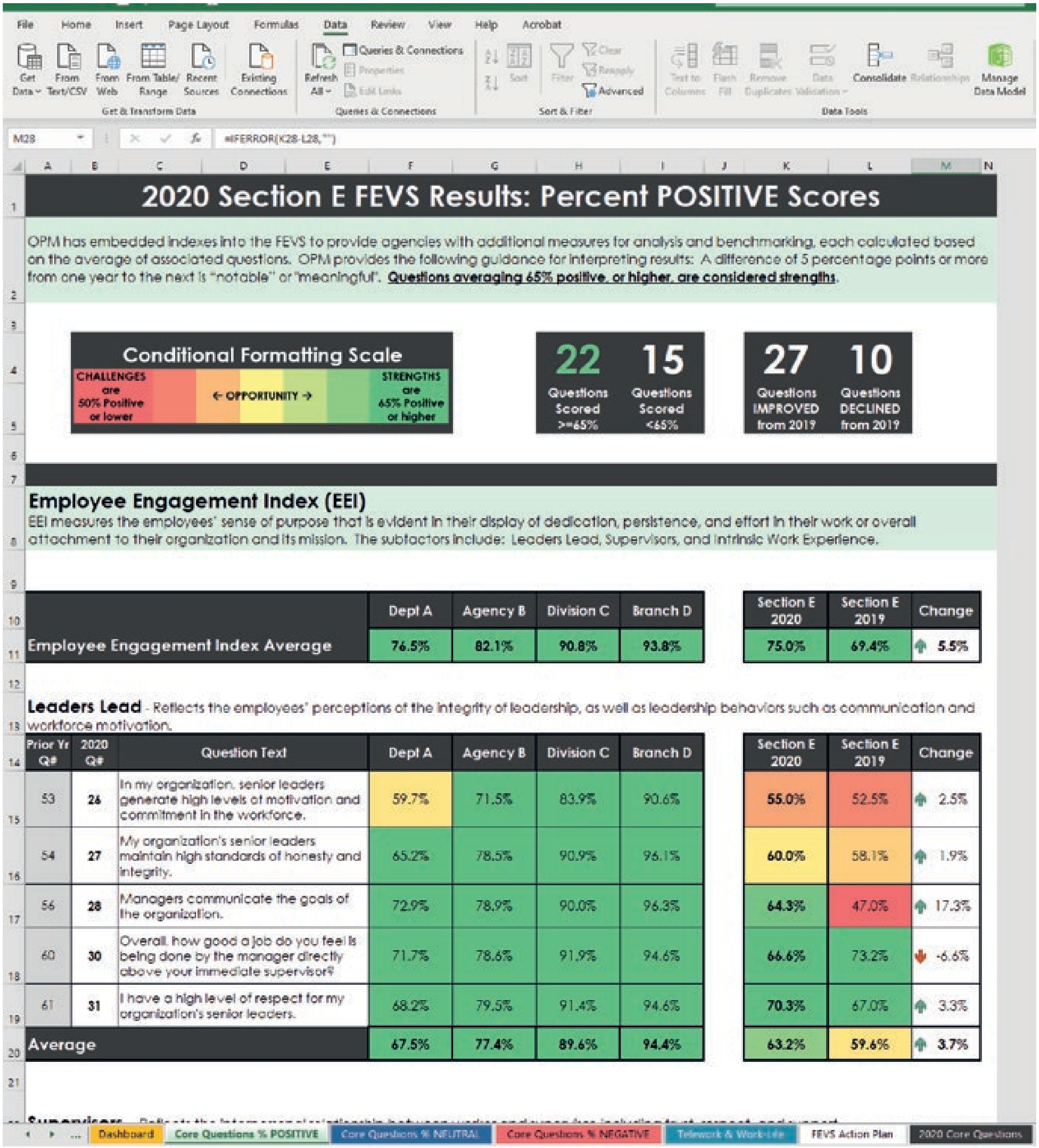
Topic	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
Employee Engagement Index (EEI)	76.5%	82.1%	90.8%	93.8%	75.4%	69.4%	6.0%
Leaders Lead	67.5%	77.4%	89.6%	94.4%	64.6%	59.6%	5.1%
Supervisors	82.0%	85.1%	92.2%	93.8%	89.2%	73.4%	15.8%
Intrinsic Work Experience	80.0%	83.9%	90.7%	93.9%	72.4%	75.3%	-2.9%
Global Satisfaction Index (GSI)	74.1%	78.9%	88.1%	93.6%	64.9%	52.4%	12.5%
Core Survey Items Related to the COVID-19 Pandemic	81.2%	87.0%	92.1%	94.3%	75.3%	72.4%	2.9%
Congressional Priority Items	69.4%	76.7%	86.4%	91.7%	66.0%	58.1%	7.9%
Leadership and Management Practices that Contribute to Agency Performance	71.6%	76.8%	86.7%	92.4%	62.6%	62.1%	0.5%
Employee Satisfaction	71.7%	76.8%	87.2%	91.4%	67.3%	61.2%	6.1%
Leadership Policies and Practices	64.7%	73.6%	88.1%	89.7%	65.2%	53.4%	11.8%
Work Environment	74.5%	78.5%	87.3%	87.7%	73.2%	62.9%	10.3%
Rewards and Recognition	60.7%	66.1%	77.7%	89.5%	47.7%	53.7%	-6.0%
Opportunities for Professional Development and Growth	73.2%	77.5%	88.2%	92.1%	75.7%	66.2%	9.5%
Opportunity to Contribute to Achieving Organizational Mission	83.6%	88.3%	94.6%	98.0%	77.7%	69.6%	8.1%

Conditional Formatting Scale

CHALLENGES are 50% Positive or lower | ← OPPORTUNITY → | STRENGTHS are 65% Positive or higher

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

FIGURE 9.15 Percentage of Positive Employee Engagement Scores from the Federal Employee Viewpoint Survey



M28 =IFERROR(K28-L28,"")

2020 Section E FEVS Results: Percent POSITIVE Scores

OPM has embedded indexes into the FEVS to provide agencies with additional measures for analysis and benchmarking, each calculated based on the average of associated questions. OPM provides the following guidance for interpreting results: A difference of 5 percentage points or more from one year to the next is "notable" or "meaningful". **Questions averaging 65% positive, or higher, are considered strengths.**

Conditional Formatting Scale

CHALLENGES are 50% Positive or lower ← OPPORTUNITY → STRENGTHS are 65% Positive or higher

22 Questions Scored >=65%

15 Questions Scored <65%

27 Questions IMPROVED from 2019

10 Questions DECLINED from 2019

Employee Engagement Index (EEI)

EEI measures the employees' sense of purpose that is evident in their display of dedication, persistence, and effort in their work or overall attachment to their organization and its mission. The subfactors include: Leaders Lead, Supervisors, and Intrinsic Work Experience.

	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
Employee Engagement Index Average	76.5%	82.1%	90.8%	93.8%	75.0%	69.4%	↑ 5.5%

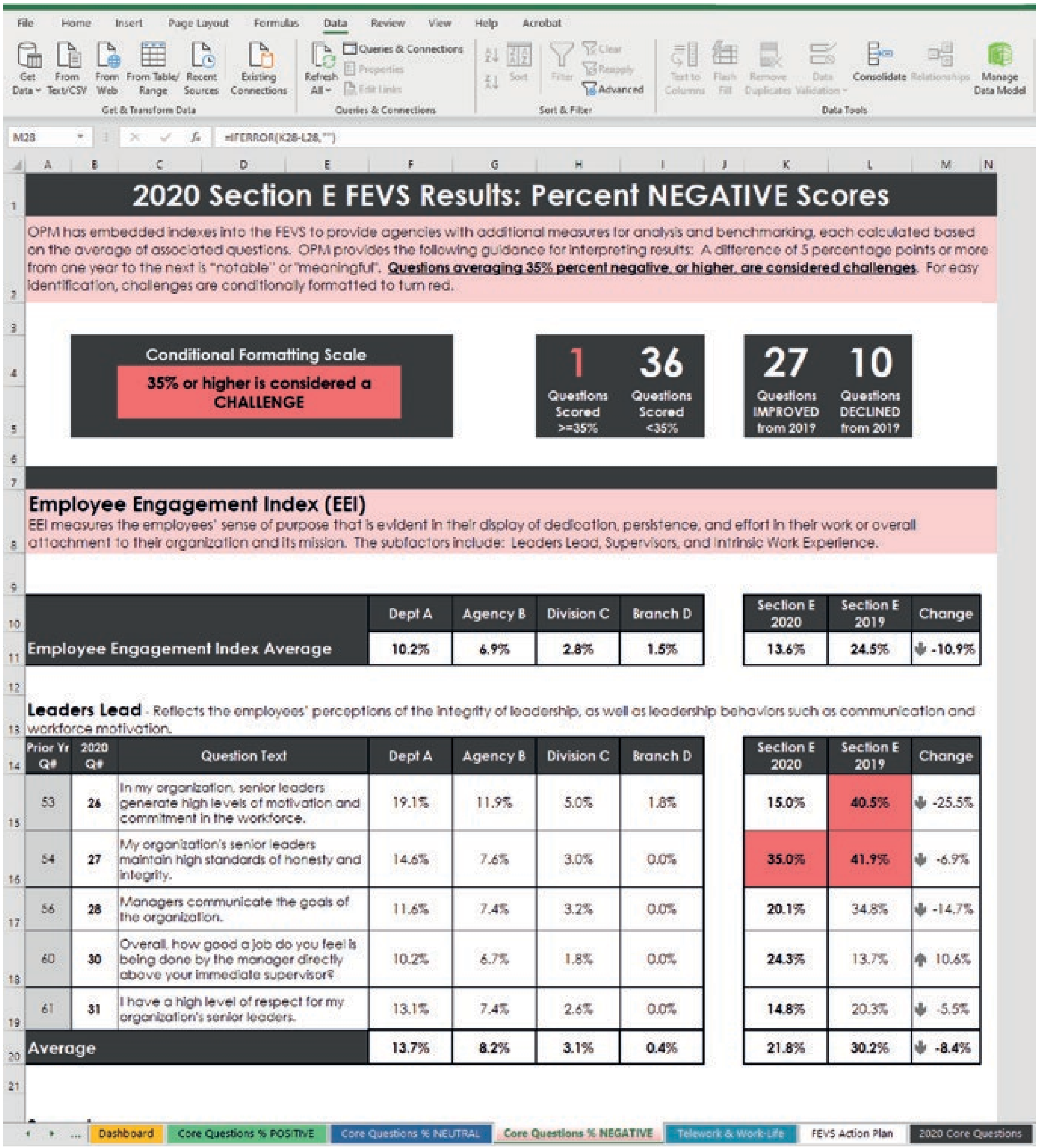
Leaders Lead - Reflects the employees' perceptions of the integrity of leadership, as well as leadership behaviors such as communication and workforce motivation.

Prior Yr Q#	2020 Q#	Question Text	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
53	26	In my organization, senior leaders generate high levels of motivation and commitment in the workforce.	59.7%	71.5%	83.9%	90.6%	55.0%	52.5%	↑ 2.5%
54	27	My organization's senior leaders maintain high standards of honesty and integrity.	65.2%	78.5%	90.9%	96.1%	60.0%	58.1%	↑ 1.9%
56	28	Managers communicate the goals of the organization.	72.9%	78.9%	90.0%	96.3%	64.3%	47.0%	↑ 17.3%
60	30	Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor?	71.7%	78.6%	91.9%	94.6%	66.6%	73.2%	↓ -6.6%
61	31	I have a high level of respect for my organization's senior leaders.	68.2%	79.5%	91.4%	94.6%	70.3%	67.0%	↑ 3.3%
Average			67.5%	77.4%	89.6%	94.4%	63.2%	59.6%	↑ 3.7%

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

EVS ART is conditionally formatted so that themes are easily identified. Users do not have to know how the tool works to be able to interpret the story or determine where they need to focus, where they have strengths, and where there are opportunities for improvement.

FIGURE 9.16 Percentage of Negative Employee Engagement Scores from the Federal Employee Viewpoint Survey



M28 =IFERROR(K28-L28,"")

2020 Section E FEVS Results: Percent NEGATIVE Scores

OPM has embedded indexes into the FEVS to provide agencies with additional measures for analysis and benchmarking, each calculated based on the average of associated questions. OPM provides the following guidance for interpreting results: A difference of 5 percentage points or more from one year to the next is "notable" or "meaningful". **Questions averaging 35% percent negative, or higher, are considered challenges.** For easy identification, challenges are conditionally formatted to turn red.

Conditional Formatting Scale

35% or higher is considered a CHALLENGE

1	36	27	10
Questions Scored >=35%	Questions Scored <35%	Questions IMPROVED from 2019	Questions DECLINED from 2019

Employee Engagement Index (EEI)

EEI measures the employees' sense of purpose that is evident in their display of dedication, persistence, and effort in their work or overall attachment to their organization and its mission. The subfactors include: Leaders Lead, Supervisors, and Intrinsic Work Experience.

	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
Employee Engagement Index Average	10.2%	6.9%	2.8%	1.5%	13.6%	24.5%	↓ -10.9%

Leaders Lead - Reflects the employees' perceptions of the integrity of leadership, as well as leadership behaviors such as communication and workforce motivation.

Prior Yr Q#	2020 Q#	Question Text	Dept A	Agency B	Division C	Branch D	Section E 2020	Section E 2019	Change
53	26	In my organization, senior leaders generate high levels of motivation and commitment in the workforce.	19.1%	11.9%	5.0%	1.8%	15.0%	40.5%	↓ -25.5%
54	27	My organization's senior leaders maintain high standards of honesty and integrity.	14.6%	7.6%	3.0%	0.0%	35.0%	41.9%	↓ -6.9%
56	28	Managers communicate the goals of the organization.	11.6%	7.4%	3.2%	0.0%	20.1%	34.8%	↓ -14.7%
60	30	Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor?	10.2%	6.7%	1.8%	0.0%	24.3%	13.7%	↑ 10.6%
61	31	I have a high level of respect for my organization's senior leaders.	13.1%	7.4%	2.6%	0.0%	14.8%	20.3%	↓ -5.5%
Average			13.7%	8.2%	3.1%	0.4%	21.8%	30.2%	↓ -8.4%

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

It became clear that one should look beyond whether scores are positive or not. Often, federal leaders focus solely on questions that average 65 percent positive or lower. While this is important, going beyond to review both neutral and negative scores can provide clarity (figure 9.16). For instance, there is a big difference between a *low positive score with a high neutral score* and a *low positive score with a* *high negative score*. While a low positive score is not preferable, if it is paired with a high neutral score, it could indicate an opportunity for communication and clarification, whereas a low positive score paired with a high negative score clearly indicates a problem area.

Action-Planning Tab

Effective action planning can transform data into meaningful change. The EVS ART action-planning tab is designed to help initiate the process and determine next steps (see figure 9.17). After reviewing the results, users can

- Identify focus areas (these areas can align with OPM index measures and key categories or can be customized to reflect areas of interest),
- Enter related FEVS question numbers (data will automatically populate based on the question number selected),
- Brainstorm initiatives and interventions geared toward improving focus areas, considering both the potential impact and available resources,
- Designate a lead person or office to address each focus area, and
- Assign target completion dates.

Implementation and Reform Sequence

When initiating the development of the tool, the team first identified the questions that made up each of the FEVS index measures. This was a bigger challenge than anticipated because no one document contained all the information needed, so they created their own. The team scoured the OPM's FEVS technical guides going back to 2012 to identify each measure, its definition, and the associated survey questions. They compiled a master document with this information that is still in use today.

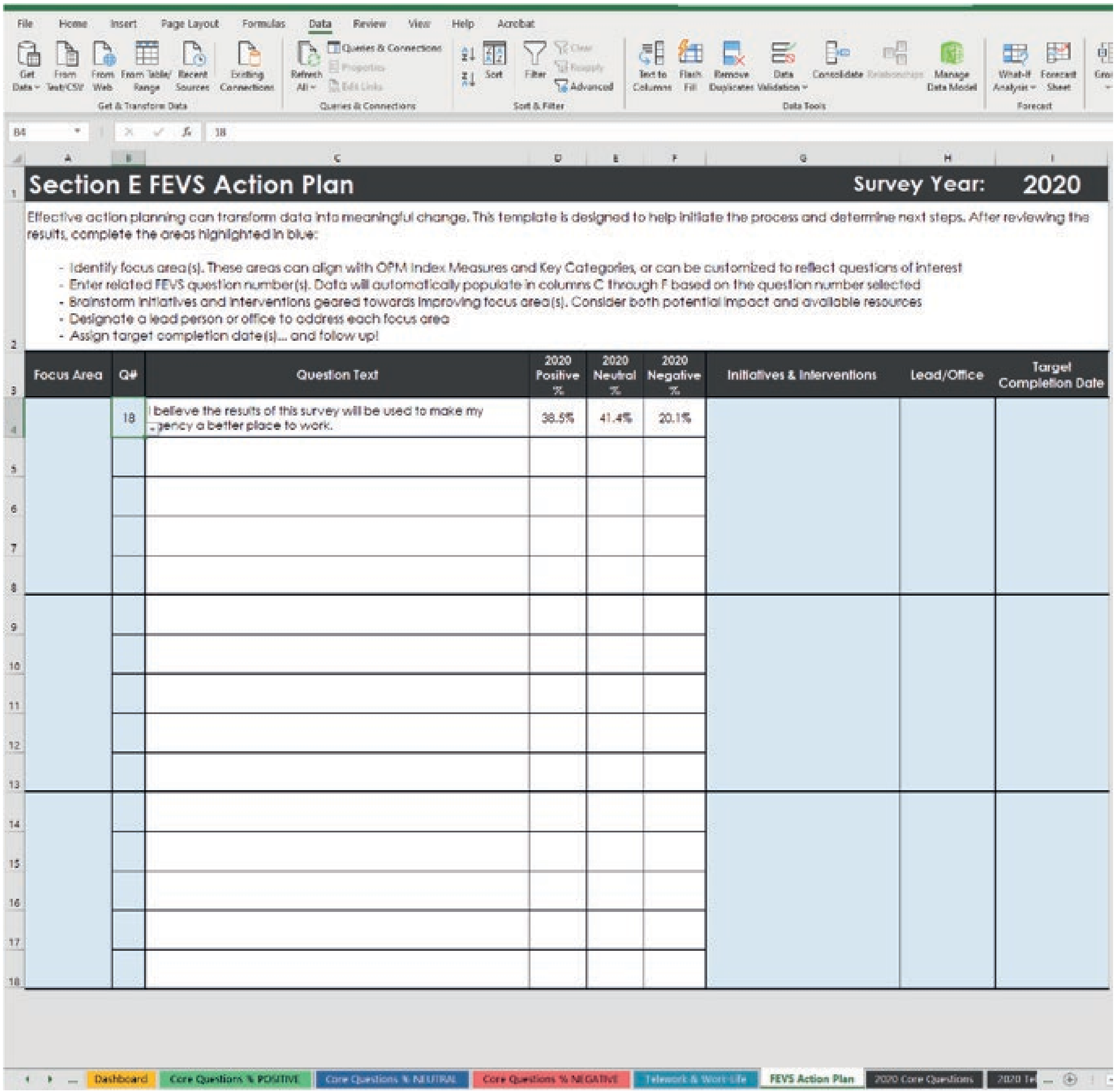
The team also faced their own learning curve. They had a creative vision of what they wanted to accomplish, what they wanted the tool to look like, and what they wanted it to do, but they did not necessarily have the expertise to accomplish it—or so they thought. So the team began to work backward, peeling back the layers of what they anticipated the final product would look like, then researching and teaching themselves how to accomplish each step along the way.

Whether it was the visual appearance and flow or the inner workings of many hidden pivot tables and charts, each task was new, each was important, and each was tackled and then painstakingly built out, tested, adjusted, and then tested again. With each success came a small victory that fueled the next challenge. The analyst team looked for gaps, identified opportunities for improvement, and created efficiencies—and this project provided all of that and more. They knew that what they were creating could feasibly make a difference in the way the FEVS was used and valued governmentwide.

Upon completion, the NIDDK team recognized that the dashboard could be useful in other contexts and decided to share it broadly. Little did they know that getting the word out and giving the tool to other departments and agencies would prove to be more of a challenge than building the tool itself. First and foremost, the creation of EVS ART began as a grassroots effort, far removed from those who managed and administered the FEVS. The NIDDK team began sharing their tool across their agency, but the department had little influence in sharing it broadly.

When the team gained the attention of the US Office of Management and Budget (OMB) and the OPM, all of that changed. The NIDDK team was invited to present to the OMB and the OPM. The OMB was impressed with EVS ART and praised the work done by the NIDDK.[14] The OMB and the OPM organized a venue during which the NIDDK shared the tool with federal chief human capital officers (CHCOs) governmentwide. With the amplification of this extraordinary tool, the team received numerous requests for

FIGURE 9.17 Sample Action-Planning Tab in the EVS ART



Section E FEVS Action Plan **Survey Year: 2020**

Effective action planning can transform data into meaningful change. This template is designed to help initiate the process and determine next steps. After reviewing the results, complete the areas highlighted in blue:

- Identify focus area(s). These areas can align with OPM Index Measures and Key Categories, or can be customized to reflect questions of interest
- Enter related FEVS question number(s). Data will automatically populate in columns C through F based on the question number selected
- Brainstorm initiatives and interventions geared towards improving focus area(s). Consider both potential impact and available resources
- Designate a lead person or office to address each focus area
- Assign target completion date(s)... and follow up!

Focus Area	Q#	Question Text	2020 Positive %	2020 Neutral %	2020 Negative %	Initiatives & Interventions	Lead/Office	Target Completion Date
	18	I believe the results of this survey will be used to make my agency a better place to work.	38.5%	41.4%	20.1%			

Source: Screenshot of EVS ART 2020, NIDDK.
Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

demonstration of EVS ART from agencies and departments outside of their own. This was a challenge in itself for the team of three because many of the organizations expressing interest in the tool were within the same agency or department but siloed from one another, resulting in multiple requests from each. Additionally, due to turnover in political leadership, there were requests to return to organizations to share with new leaders the capabilities of EVS ART and the progress recognized by the NIDDK when their FEVS results were used to inform change.

The enormity of the US federal government made it more and more difficult to manage such requests. The NIDDK team established an online presence, which allowed federal employees to access the tool and its training resources. The OPM also offered EVS ART as a resource to agencies and departments as part of work being done under the President's Management Agenda. The collaboration between the NIDDK, the OPM, and the OMB blossomed, and since 2017, the NIDDK team has conducted hundreds of presentations, trainings, and customized workshops that have reached personnel in each of the 15 executive departments. These meetings and training sessions continue because the team has found that organizations at all levels within departments are interested in learning about successful practices.

It was important to the team that other federal employees knew of EVS ART's availability, and equally important that they benefited from it, but it was also important that no outside entity unduly profit from its use. EVS ART was created by federal employees, for federal employees, using resources readily available and with no contractor support. Realizing its benefit and potential, the NIDDK elected to share its tool with any and all federal entities expressing interest, free of charge. Its view as a steward of federal funds was that it had done the work and, by sharing its tool, others could avoid duplicating its efforts and could create extraordinary efficiencies within their own organizations. Many organizations had spent weeks, if not months, and sometimes thousands of federal dollars on outside consultants to do what they could now do for themselves in minutes using EVS ART. The NIDDK has received numerous requests from outside vendors and consultants related to the use of its tool in support of work they are doing for other federal organizations—and even requests for unlocked versions that they can modify for their own use with federal clientele. This goes against the grain of the NIDDK's vision for sharing the tool at no cost. The team does not want federal funds to be used, perhaps unknowingly, to pay for a tool available for free.

Feedback, Flexibility, and Continuous Improvement

End users of EVS ART have expressed gratitude for the tool.[15] Having this tool helps leadership to see data in one place, or by field office if they like. This tool gives the flexibility to do that, and quickly, economizing time. The way the analysis and reports are organized makes the data clearer, which makes for faster analysis of employee feedback and allows leadership to address the question "what now?" so that agencies can develop a plan of action based on employee responses.

EVS ART was designed to provide users with many ways to view data. It offers a dashboard, heat maps, breakouts by index measure, and bar charts. However, there is always the desire to display data in different ways. Early on when the team received requests from users to modify the tool, they provided unlocked versions to those requesting to make modifications. After seeing the inner workings of EVS ART, and the thought that went into the creation of the tool, a user remarked that "it look[ed] easier than it really is," and this is true.

The team learned, through trial and error, that it was not wise to share unlocked versions of the tool. There are numerous pivot tables and charts and thousands of links and formulas in each of the templates. Breaking any one of them could compromise the analysis. Because of this, they decided to no longer provide unlocked versions and instead to collect the feedback received and use that information to improve the templates each year.

LESSONS LEARNED

The project taught the implementation team a set of lessons:

- **Cost does not equal worth.** A tool does not have to be expensive to provide extraordinary value.
- **Keep the solution simple.** While the inner workings of a tool may venture over to the complex side, do not make the act of engaging with the analysis complex for users, or they will not use it.
- **Make the solution accessible to all.** This means two things. One, ensure that the tool is accessible to those with disabilities, and two, make it available to the broadest audience possible. If people do not

know about the tool, they will continue to spend unnecessary time re-creating analyses and unnecessary money on contracts to conduct analyses, or they may simply do nothing with the valuable information they have at their fingertips.

- **Ensure that the output of the tool is intuitive and useful.** Do not make users reanalyze the analysis—the tool should do the work for them the first time. Provide results in a format that can be utilized for presentation.
- **Tell the story.** Do not overwhelm end users. Offer a high-level visual overview and then direct them down an intuitive path to more specific details.
- **Be transparent.** When working with results, whether positive or negative, do not shy away from difficult conversations. Survey takers already know whether something is working well or not. Be up front, dig deeper, and let them know that their input will drive change.
- **Tie actions back to survey feedback.** When creating initiatives based on feedback obtained through a survey, it is important to tie the organization's actions back to the voice of the people. This will increase engagement, add validity to the survey, and in most cases, increase future participation.

What was the most basic lesson learned? Great things can come from grassroots efforts.

IMPACT OF THE SOLUTION

The introduction of EVS ART created immediate efficiencies in both the time and cost of completing the FEVS analysis. Colleagues at the Centers for Disease Control and Prevention (CDC) experienced a significant reduction in the time spent conducting FEVS analysis. Prior to EVS ART, they produced 24 reports in 72 workdays at a cost of approximately US$30,861. Reporting can now be done in one workday at a cost of approximately US$1,129—a savings of US$29,732 and a 96 percent reduction in both time and cost. These efficiencies have allowed the CDC to increase its reporting sixfold to 150 individual analyses—meaning that 126 additional managers now receive their own customized FEVS results.

An NIH analyst who once spent 30 hours creating one report at an average cost of US$1,350 can now complete an analysis in less than 5 minutes at a cost of US$3.75. Simplifying the analysis process means that frontline managers can access meaningful data to better inform policies, programs, and initiatives much sooner. They can also have confidence that the information they are using to create or bolster initiatives is coming directly from those whom their actions impact most.

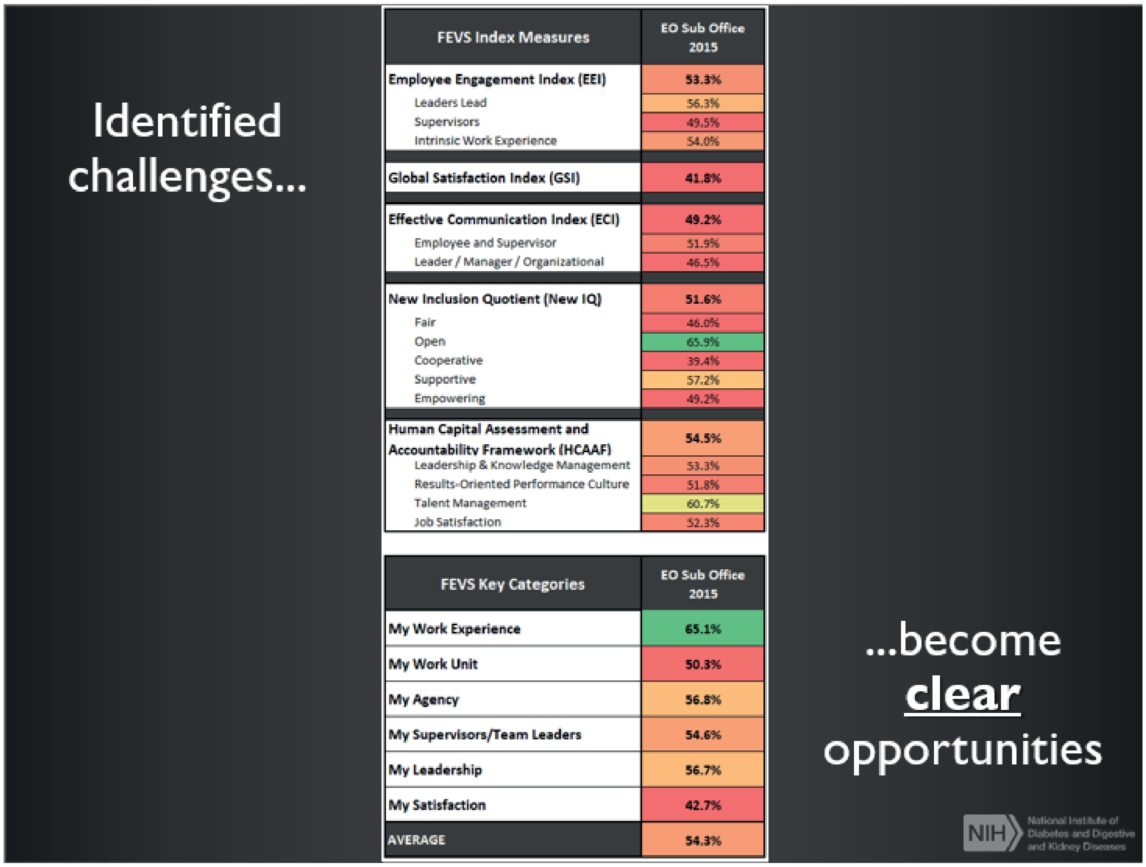
BEYOND ANALYTICS: CREATING MEASURABLE AND SUSTAINABLE CHANGE

While the efficiencies created by EVS ART have helped save both time and money, the most important aspect, by far, has been the increased ability to identify themes and measure organizational change (see figure 9.18).

One example of a success story concerns the transformation of an underperforming organization. This organization was forward facing and interfaced with all 1,300 institute employees. To remedy its underperformance, the NIDDK Executive Officer stepped in with a multipronged approach and, over the course of a year,

- Put in place new standards and forms of accountability, including metrics to measure productivity (industry standards),

FIGURE 9.18 Identifying Challenges through the EVS ART



Identified challenges…

FEVS Index Measures	EO Sub Office 2015
Employee Engagement Index (EEI)	**53.3%**
Leaders Lead	56.3%
Supervisors	49.5%
Intrinsic Work Experience	54.0%
Global Satisfaction Index (GSI)	**41.8%**
Effective Communication Index (ECI)	**49.2%**
Employee and Supervisor	51.9%
Leader / Manager / Organizational	46.5%
New Inclusion Quotient (New IQ)	**51.6%**
Fair	46.0%
Open	65.9%
Cooperative	39.4%
Supportive	57.2%
Empowering	49.2%
Human Capital Assessment and Accountability Framework (HCAAF)	**54.5%**
Leadership & Knowledge Management	53.3%
Results-Oriented Performance Culture	51.8%
Talent Management	60.7%
Job Satisfaction	52.3%

FEVS Key Categories	EO Sub Office 2015
My Work Experience	**65.1%**
My Work Unit	**50.3%**
My Agency	**56.8%**
My Supervisors/Team Leaders	**54.6%**
My Leadership	**56.7%**
My Satisfaction	**42.7%**
AVERAGE	**54.3%**

…become **clear** opportunities

NIH National Institute of Diabetes and Digestive and Kidney Diseases

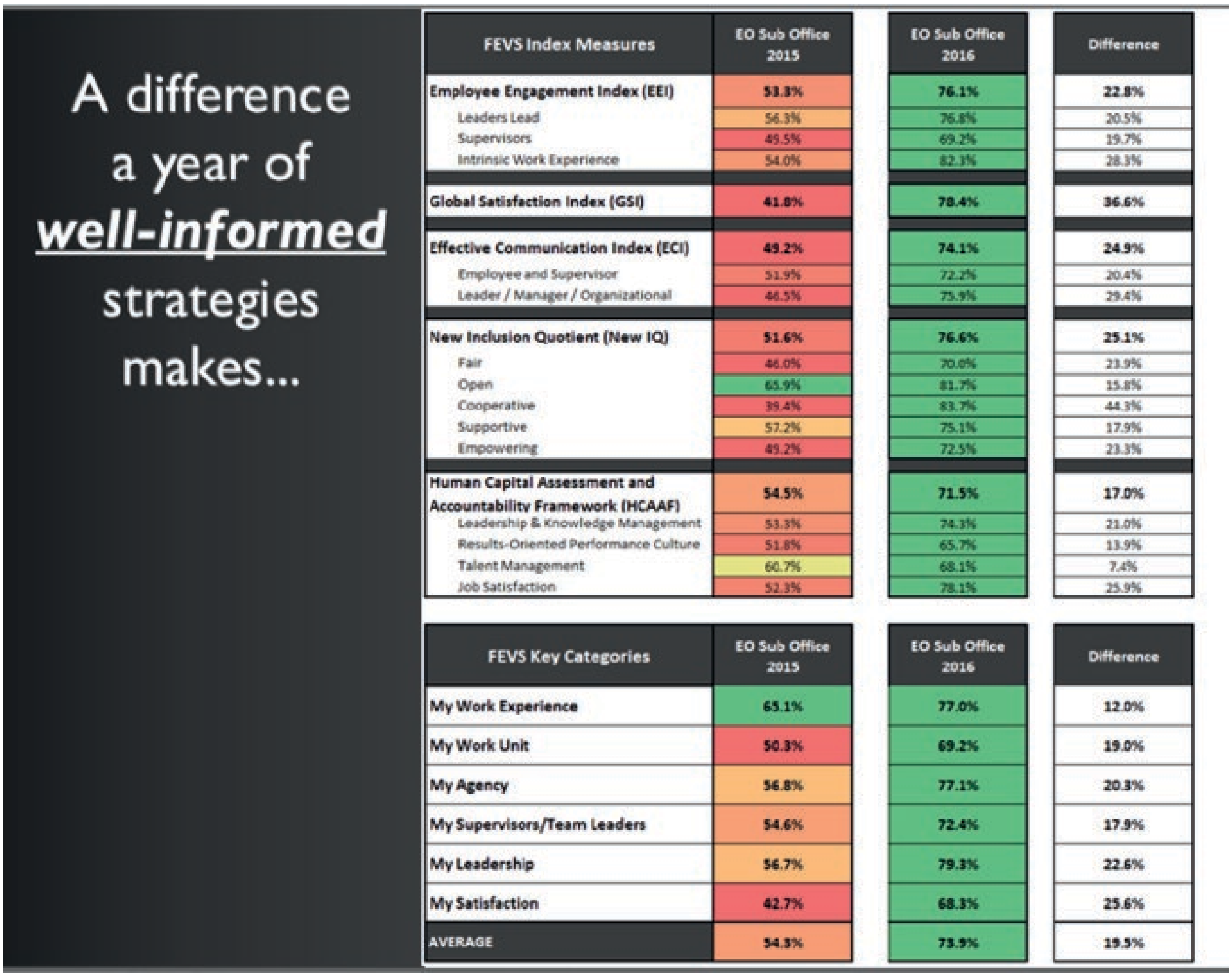
Source: Original figure for this publication, NIDDK.
Note: EO = executive office; EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

- Worked closely with leaders to create a new vision for the group,
- Changed out leaders who did not embrace the new vision,
- Taught necessary competencies to supervisors,
- Created opportunities for high performers,
- Ensured that mediocrity was not acceptable and that there were consequences for poor performance, not only for employees but also for leaders, and
- Worked closely with the employees of the organization so they knew in real time what changes were happening and why, ensuring that each employee within the organization had a voice.

Over the course of a year, the organization was transformed. Employees knew it because service improved, complaints were greatly reduced, and partnerships began to form. By using EVS ART, the NIDDK was able to prove that its targeted interventions were working. Figure 9.19 illustrates the transformation from one year to the next. The employee engagement index went up by 22.8 percentage points, the global satisfaction index went up 36.6 percentage points, and the new inclusion quotient increased from 51.6 percent to 76.6 percent positive.[16]

NIDDK staff recognized the transformation, and confidence in the organization returned. The work continues to pay off, and five years later, the success of the interventions is still clearly demonstrated (see figure 9.20).

FIGURE 9.19 Changes in Federal Employee Viewpoint Survey Index Measures, 2015–16



A difference a year of ***well-informed*** strategies makes…

FEVS Index Measures	EO Sub Office 2015	EO Sub Office 2016	Difference
Employee Engagement Index (EEI)	**53.3%**	**76.1%**	**22.8%**
Leaders Lead	56.3%	76.8%	20.5%
Supervisors	49.5%	69.2%	19.7%
Intrinsic Work Experience	54.0%	82.3%	28.3%
Global Satisfaction Index (GSI)	**41.8%**	**78.4%**	**36.6%**
Effective Communication Index (ECI)	**49.2%**	**74.1%**	**24.9%**
Employee and Supervisor	51.9%	72.2%	20.4%
Leader / Manager / Organizational	46.5%	75.9%	29.4%
New Inclusion Quotient (New IQ)	**51.6%**	**76.6%**	**25.1%**
Fair	46.0%	70.0%	23.9%
Open	65.9%	81.7%	15.8%
Cooperative	39.4%	83.7%	44.3%
Supportive	57.2%	75.1%	17.9%
Empowering	49.2%	72.5%	23.3%
Human Capital Assessment and Accountability Framework (HCAAF)	**54.5%**	**71.5%**	**17.0%**
Leadership & Knowledge Management	53.3%	74.3%	21.0%
Results-Oriented Performance Culture	51.8%	65.7%	13.9%
Talent Management	60.7%	68.1%	7.4%
Job Satisfaction	52.3%	78.1%	25.9%

FEVS Key Categories	EO Sub Office 2015	EO Sub Office 2016	Difference
My Work Experience	**65.1%**	**77.0%**	**12.0%**
My Work Unit	**50.3%**	**69.2%**	**19.0%**
My Agency	**56.8%**	**77.1%**	**20.3%**
My Supervisors/Team Leaders	**54.6%**	**72.4%**	**17.9%**
My Leadership	**56.7%**	**79.3%**	**22.6%**
My Satisfaction	**42.7%**	**68.3%**	**25.6%**
AVERAGE	**54.3%**	**73.9%**	**19.5%**

Source: Original figure for this publication, NIDDK.
Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

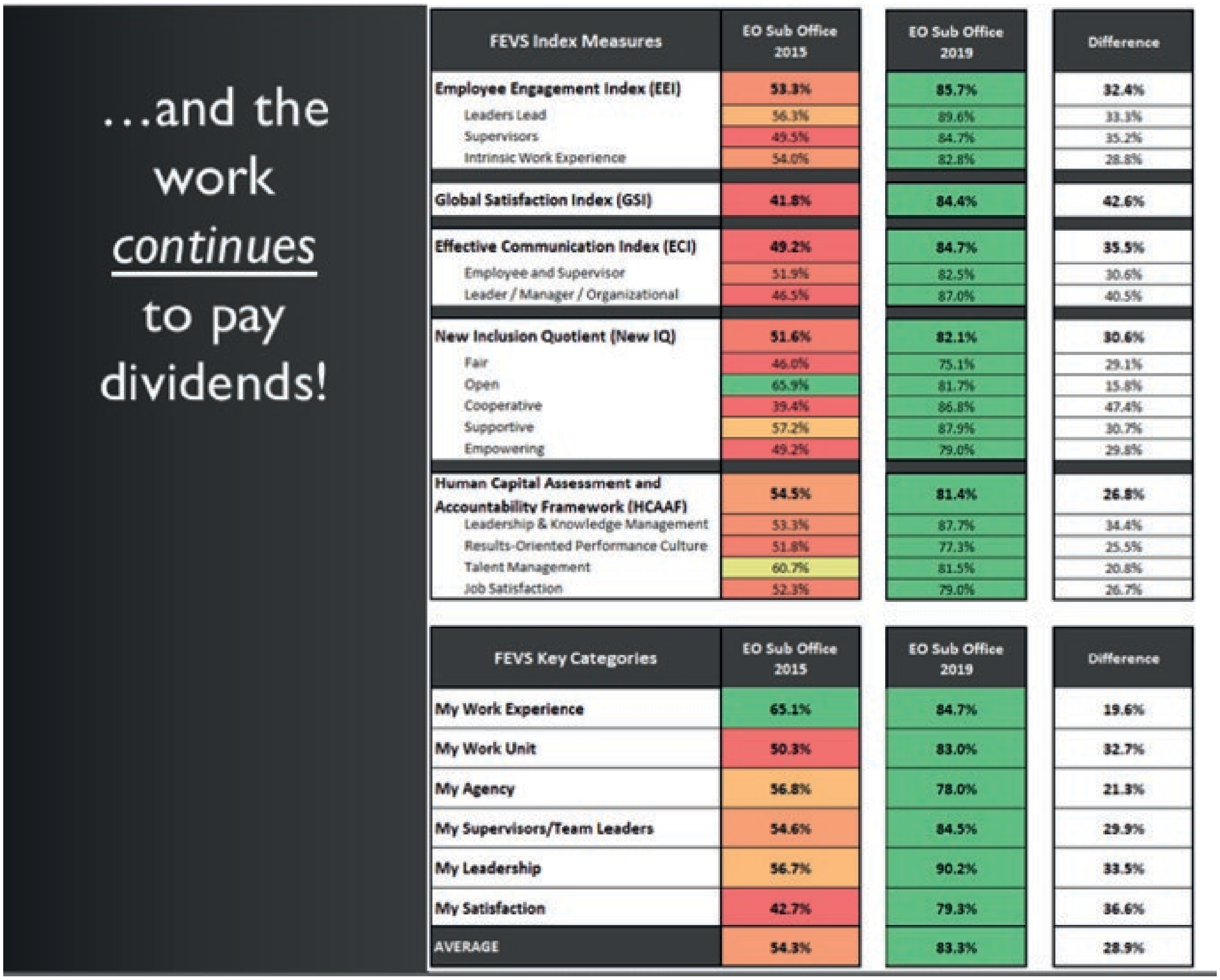
Performance Management in Practice

The same success has played out across the institute. In addition to targeted interventions, to be a truly performance-centric organization, performance management must be incorporated into an organization's culture continuously. NIDDK leadership routinely finds opportunities across the institute to highlight the importance of performance standards and conversations.

At the NIDDK, people throughout the organization shared via the FEVS that they wanted discussions with their supervisors about their performance to be more worthwhile: they wanted their supervisors to provide them with more constructive suggestions about how to improve their job performance and to give them meaningful recognition when they had done a good job. To address this, the NIDDK Executive Officer initiated the following practices:

- Reviewing performance ratings across the entire organization to make sure that all supervisors were interpreting "outstanding" rating requirements, versus "excellent" and "satisfactory" ones, in the same way and that, where appropriate, they were giving lower ratings when deserved rather than ignoring underperformance

FIGURE 9.20 Changes in Federal Employee Viewpoint Survey Index Measures, 2015–19



…and the work continues to pay dividends!

FEVS Index Measures	EO Sub Office 2015	EO Sub Office 2019	Difference
Employee Engagement Index (EEI)	**53.3%**	**85.7%**	**32.4%**
Leaders Lead	56.3%	89.6%	33.3%
Supervisors	49.5%	84.7%	35.2%
Intrinsic Work Experience	54.0%	82.8%	28.8%
Global Satisfaction Index (GSI)	**41.8%**	**84.4%**	**42.6%**
Effective Communication Index (ECI)	**49.2%**	**84.7%**	**35.5%**
Employee and Supervisor	51.9%	82.5%	30.6%
Leader / Manager / Organizational	46.5%	87.0%	40.5%
New Inclusion Quotient (New IQ)	**51.6%**	**82.1%**	**30.6%**
Fair	46.0%	75.1%	29.1%
Open	65.9%	81.7%	15.8%
Cooperative	39.4%	86.8%	47.4%
Supportive	57.2%	87.9%	30.7%
Empowering	49.2%	79.0%	29.8%
Human Capital Assessment and Accountability Framework (HCAAF)	**54.5%**	**81.4%**	**26.8%**
Leadership & Knowledge Management	53.3%	87.7%	34.4%
Results-Oriented Performance Culture	51.8%	77.3%	25.5%
Talent Management	60.7%	81.5%	20.8%
Job Satisfaction	52.3%	79.0%	26.7%

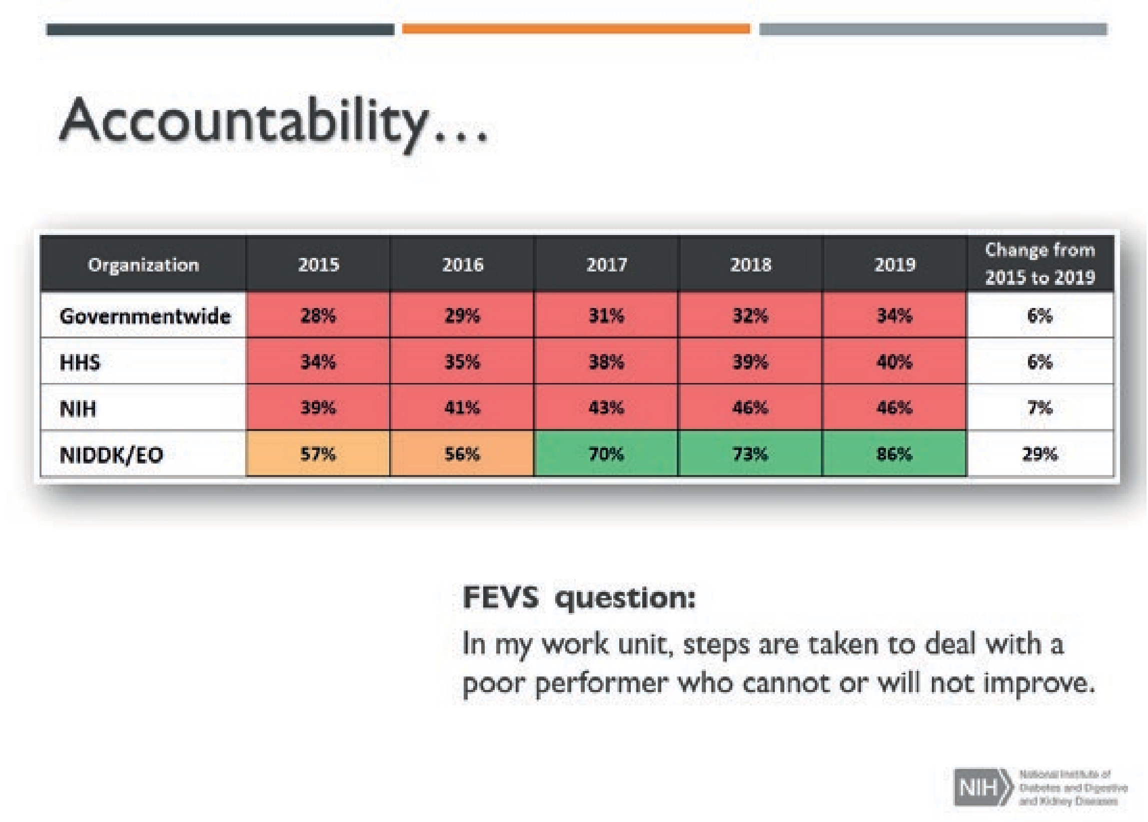
FEVS Key Categories	EO Sub Office 2015	EO Sub Office 2019	Difference
My Work Experience	**65.1%**	**84.7%**	**19.6%**
My Work Unit	**50.3%**	**83.0%**	**32.7%**
My Agency	**56.8%**	**78.0%**	**21.3%**
My Supervisors/Team Leaders	**54.6%**	**84.5%**	**29.9%**
My Leadership	**56.7%**	**90.2%**	**33.5%**
My Satisfaction	**42.7%**	**79.3%**	**36.6%**
AVERAGE	**54.3%**	**83.3%**	**28.9%**

Source: Original figure for this publication, NIDDK.
Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

- Reviewing all awards and retention incentives to make sure there was equity and fairness in who received awards and in what amounts
- Sending mid-year and end-of-year communications to the NIDDK's supervisors, reiterating what employees had said and emphasizing that staff played an active role in their performance conversations
- Sending communications to staff reminding them that what they said was important and encouraging them to play an active role in their performance conversations
- Sharing the document "Performance Management Tips and Templates" with both supervisors and staff to equip them with the tools they needed to have more robust performance conversations.

Over time, the people of the organization saw a noticeable change in performance management, which the NIDDK has validated using the FEVS and EVS ART. Traditionally, one of the lowest-scoring questions across government has been "In my work unit, steps are taken to deal with a poor performer who cannot or will not improve." This is one of the most difficult questions to tackle across the government. Many

FIGURE 9.21 Improving Measures of Accountability at the National Institute of Diabetes and Digestive and Kidney Diseases



Accountability…

Organization	2015	2016	2017	2018	2019	Change from 2015 to 2019
Governmentwide	28%	29%	31%	32%	34%	6%
HHS	34%	35%	38%	39%	40%	6%
NIH	39%	41%	43%	46%	46%	7%
NIDDK/EO	57%	56%	70%	73%	86%	29%

FEVS question:
In my work unit, steps are taken to deal with a poor performer who cannot or will not improve.

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Source: Original figure for this publication, NIDDK.
Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

federal leaders have said that it should be removed from the FEVS because, due to the confidential nature of employee relations, it is nearly impossible to demonstrate that actions are being taken.[17]

However, the NIDDK proved that it is possible. Leaders throughout the institute devoted resources to assist supervisors and employees early on when there were problems with performance or conduct. The key was creating a culture where early intervention occurs and clear standards and accountabilities are established and transparent. When this was done, staff began to notice underperforming organizations improve (see figure 9.21).

CHALLENGES FACED AND THE PATH FORWARD

The biggest challenge for the NIDDK team has been balancing their traditional responsibilities with the demands of creating, modifying, and supporting a tool that has gained in popularity. Since its inception, EVS ART has been enhanced to expand its capabilities several times due to the NIDDK's desire to strive for continuous improvement based on feedback received from users. The FEVS itself has undergone changes over the last three years, and EVS ART has required substantial modification to adapt to those changes as well. The team has learned that, while there is great satisfaction in being able to provide their federal colleagues with a tool that evolves with their needs, this also means that their work is never really done.

One function not yet incorporated by the tool's creator is the ability of the tool to determine statistical significance in changes from one year to the next, or between organizations of similar or different sizes. The addition of this capability could help to "win over" survey cynics. Last, with the topics of diversity, equity, inclusion, and accessibility at the forefront of conversations within the United States, it would be helpful to have the ability to compare data using different demographics, such as race and ethnicity. This is something that the NIDDK team is actively working on. While not all users have access to this level of data, the team would like to provide a similar user-friendly reporting format to those who do have access.

CONCLUSION

All leaders aspire to create and sustain high-functioning organizations. How can organizations achieve high-functioning performance, much less provide measurable evidence that they have reached this goal?

The synergy between technology for data analytics and the voice of the people can be powerful. It can inform a leader's strategy and resource allocation and provide evidence that an organization's performance and engagement activities are paying off. The NIDDK now uses the FEVS as a launchpad for moving forward, not as a report card looking back. It is with this in mind that it put into effect the year-round campaign "You Speak … We Listen … Things Happen!" to reiterate to employees that it is constantly listening to their voices and taking their feedback into account in the planning of programs and initiatives. Leadership incorporates this campaign into email communications, posters, and all-hands meetings to remind employees that their voices make a difference.

The NIDDK Executive Officer also conducts workshops to build communities and connect with staff. Early on, these workshops were used as part of targeted interventions. Now, as a very high-functioning organization, the NIDDK has transitioned to more strategic initiatives. It does this by harnessing the talents of staff who have relevant interests and technical expertise that extend beyond their functional areas to deliver workshops that continue to strengthen employee development—in lieu of bringing in outside facilitators. It focuses on career development, offering world cafés that allow staff one-on-one interaction with senior leaders from across the NIH who volunteer to share experiences, as well as specialized workshops on resilience, problem solving, conducting difficult conversations, and managing up.

Another part of the NIDDK's success has been in creating many strategic initiatives. Some of the more novel programs it has put in place, which have resulted in an increase in FEVS scores and in employee engagement across the institute, include using crowdsourcing to initiate conversations and capture ideas, incorporating pulse surveys into town halls, and conducting stay and exit interviews with staff. In addition, it has created a novel awards program to recognize "rising stars," innovative problem solving, and personification of the organization's core values. It has also focused on the professional and career development of staff through the launch of a formal mentoring program, a shadowing program, a new supervisors program, and the novel Career Climbers Cohort, which was specifically designed for staff who were junior in experience or brand new to the federal workforce. Each of these initiatives, programs, and activities has been informed by the institute's employees. This largely explains the institute's success in the FEVS's "Belief in Action" question: "I believe the results of this survey will be used to make my agency a better place to work." Across government, this question has traditionally scored incredibly low—but at the NIDDK, that has changed.

In 2015, the NIDDK was able to do its first deeper-dive analysis using an early version of EVS ART. Armed with this information, they set out to create employee-informed change, and this did not go unnoticed. Between 2015 and 2016, the NIDDK Executive Office's positive responses to the "Belief in

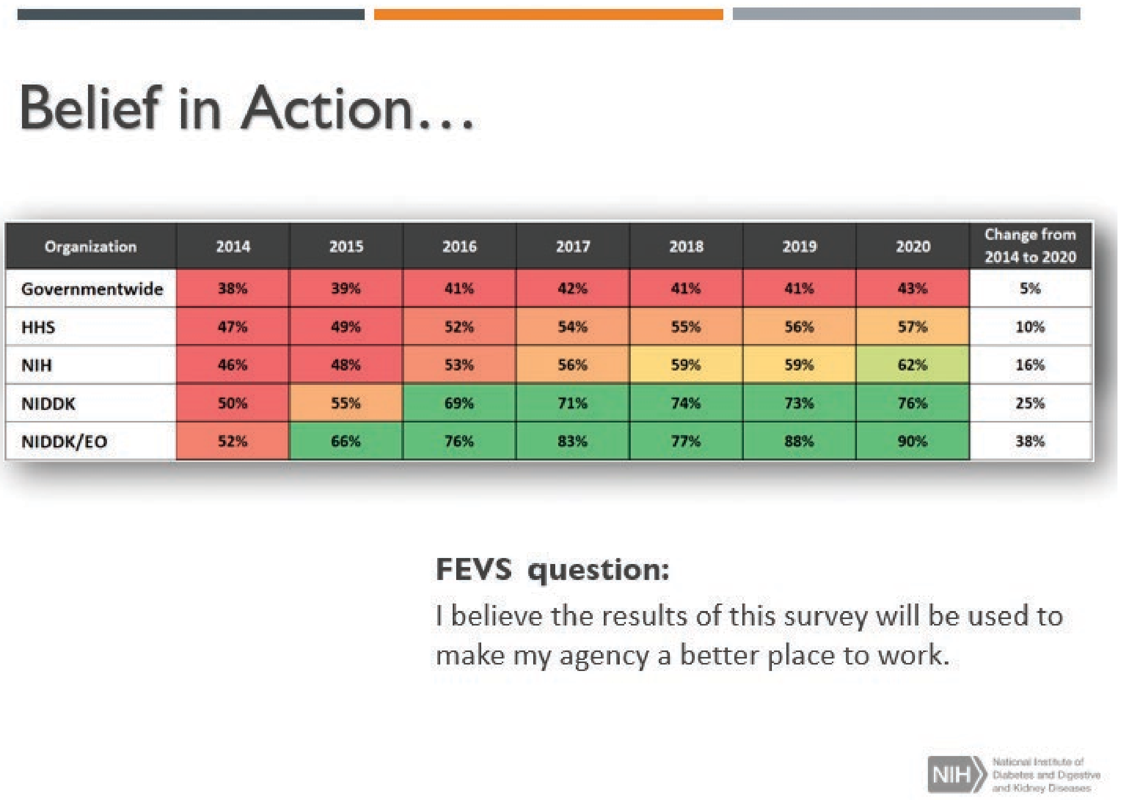
Action" question jumped by 14 percentage points, from 52 percent to 66 percent. In 2020, this same office recognized a "Belief in Action" score that was 90 percent positive—a jump of 38 percentage points from 2014 (see figure 9.22).

With the increase in "Belief in Action" scores, survey response rates increased as well. The NIDDK's overall employee participation increased from 36.8 percent to 74.5 percent (see figure 9.23).

Very basic things are needed to help ensure an organization's success. An organization requires reliable and real-time data, the ability to keep a finger on its own pulse, and the ability to tie organizational interventions and strategic initiatives back to employees' voices. Data are only meaningful when they are accounted for, acted upon, and shared with staff. They must be incorporated into the organization's culture and practices on a daily basis. The result is an amazing ripple effect (figure 9.24).

In closing, EVS ART is an incredible resource, but it is important to remember that the tool itself cannot create change—it can only inform it. The magic lies in what is done with the information it provides. The importance of leadership buy-in and action, at all levels, is critical, and a leader's level of buy-in can either help or hinder an organization's success. When leaders effectively use employee feedback to create timely, well-informed, and meaningful initiatives, the rest will begin to fall into place—and that is a wonderful cycle to be in.

FIGURE 9.22 "Belief in Action" Scores from the Federal Employee Viewpoint Survey, 2014–20



Belief in Action…

Organization	2014	2015	2016	2017	2018	2019	2020	Change from 2014 to 2020
Governmentwide	38%	39%	41%	42%	41%	41%	43%	5%
HHS	47%	49%	52%	54%	55%	56%	57%	10%
NIH	46%	48%	53%	56%	59%	59%	62%	16%
NIDDK	50%	55%	69%	71%	74%	73%	76%	25%
NIDDK/EO	52%	66%	76%	83%	77%	88%	90%	38%

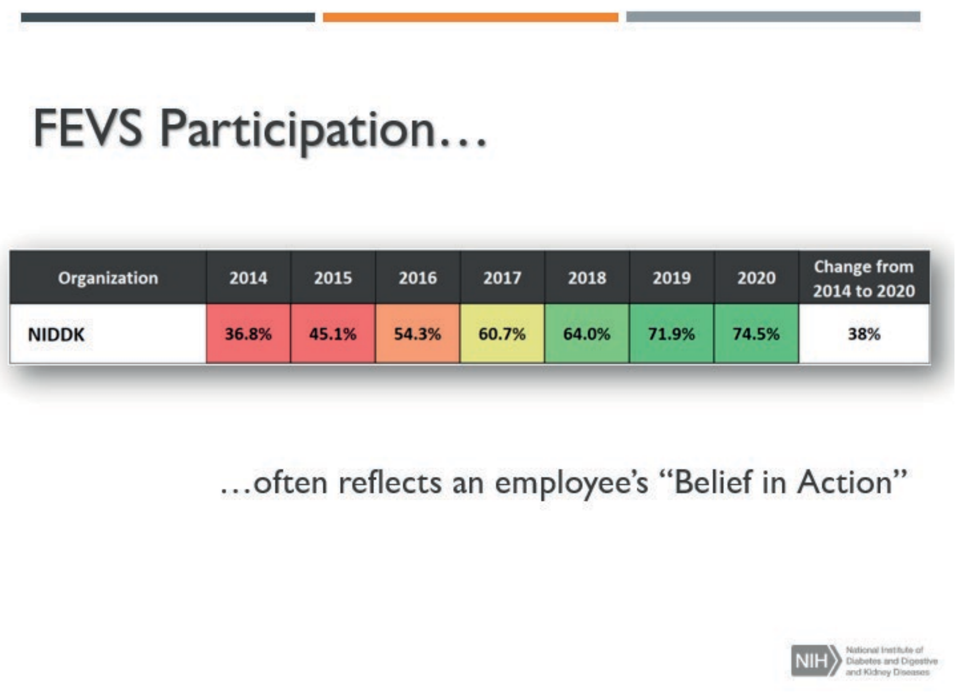
FEVS question:

I believe the results of this survey will be used to make my agency a better place to work.

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Source: Original figure for this publication, NIDDK.
Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

FIGURE 9.23 Federal Employee Viewpoint Survey Participation Rates, 2014–20



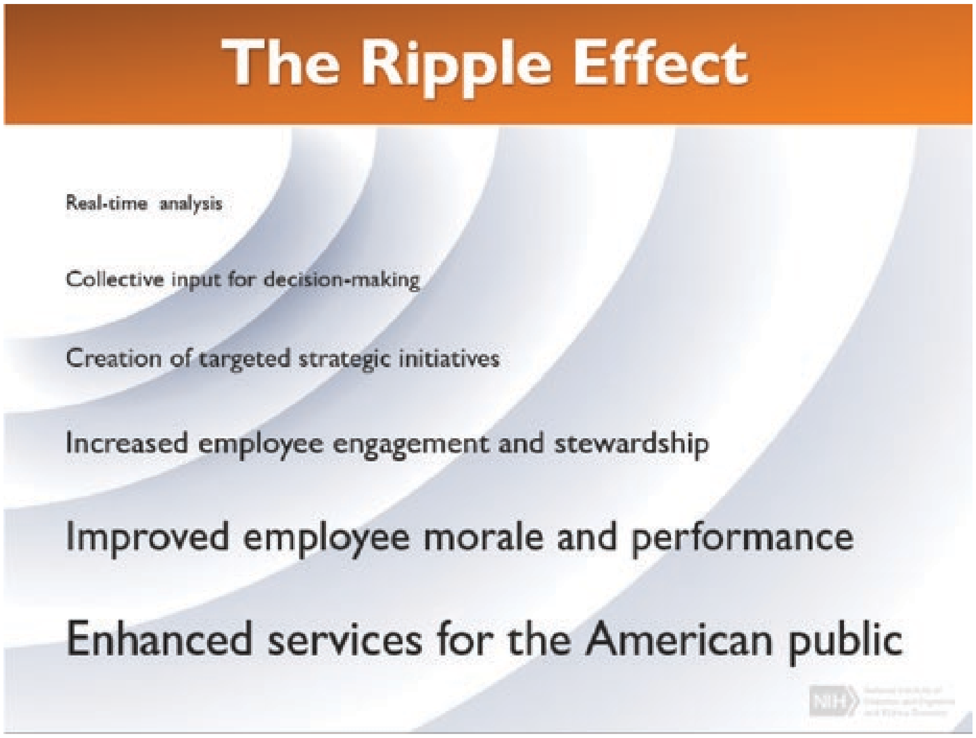
FEVS Participation…

Organization	2014	2015	2016	2017	2018	2019	2020	Change from 2014 to 2020
NIDDK	36.8%	45.1%	54.3%	60.7%	64.0%	71.9%	74.5%	38%

…often reflects an employee's "Belief in Action"

Source: Original figure for this publication, NIDDK.
Note: FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

FIGURE 9.24 The Ripple Effect



The Ripple Effect

Real-time analysis

Collective input for decision-making

Creation of targeted strategic initiatives

Increased employee engagement and stewardship

Improved employee morale and performance

Enhanced services for the American public

Source: Original figure for this publication, NIDDK.
Note: NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

NOTES

The authors are grateful for contributions by government counterparts in Brazil—Luciana Andrade (Anvisa) and Matheus Soldi Hardt (EloGroup)—Luxembourg—Ludwig Balmer (CGPO), Marc Blau (CGPO), and Danielle Bossaert (Observatory of the Civil Service)—and the United States—Camille Hoover (NIDDK) and Robin Klevins (NIDDK).

1. For an overview of different statistical capacities produced by the World Bank, see the Statistical Performance Indicators (SPI) page on the World Bank website, https://www.worldbank.org/en/programs/statistical-performance-indicators.
2. There is an important distinction between data infrastructure and information systems. Data infrastructure often entails both the digital and physical infrastructure required to store data, while the information system refers specifically to the software architecture in which data are stored.
3. For additional details, see chapter 12.
4. Ghost workers can appear in many forms: employees without a legal appointment who are nevertheless being paid; dead employees who continue to appear on the payroll and whose salaries are drawn by local administrative staff or shared with descendants; employees who draw both a pension and regular salaries; employees who draw multiple salaries from different ministries or agencies; employees who have been dismissed or have retired or resigned from service but continue to draw salaries; employees who are not working or showing up but continue to be paid; and employees who use false or multiple identities to draw multiple salaries.
5. More information about the FEVS can be found on the website of the OPM, https://www.opm.gov/fevs/.
6. See case study 9.3.
7. The project is described in more detail on the World Bank website, https://projects.worldbank.org/en/projects-operations/project-detail/P176877.
8. A sheet is a page that contains the charts, key performance indicators, and tables that compose a dashboard. An application contains multiple sheets.
9. More information about the FEVS can be found on the website of the OPM, https://www.opm.gov/fevs/. For summary statistics on full-time permanent, nonseasonal federal employees, see OPM (2017).
10. This staff time estimate does not include the time spent administering the survey or analyzing its results.
11. Index measures are aggregates of positive questions regarding perceptions of employee engagement. Key categories are generally described as survey modules---for instance, work experience and relationship to supervisors.
12. Some indexes contain four, and others, as many as 39 answers.
13. This act requires all electronic and information technology that is created by the federal government to be accessible to people with disabilities. Compliance allows users with assistive technology, such as screen readers, to use the tool.
14. One participant noted that the dashboard was "proof of concept that with strategic initiatives and targeted interventions, a federal leader can affect positive change and realize significant measurable improvement from 1 year to the next."
15. A colleague from the Department of Homeland Security "shed a tear when we learned of [the] tool, watched the video, used it, and saw how fast the analysis was performed." A senior user from the Centers for Disease Control and Prevention (CDC) shared that "EVS ART is a great tool to convey more data in less time and in a more meaningful way." And a senior colleague from the Office of the Secretary of the Department of Health and Human Services stated that "EVS ART will allow us more time to analyze the data and focus more time on strategic planning."
16. These are not jumps in specific questions but changes to the average of index measures. Many of the specific questions within the index measures went up by 40–50 percentage points in just one year.
17. Note that the specific question has been removed for 2022 and exchanged for a multiple-choice question. The wording of the question has been changed as well to address these concerns.

REFERENCES

Bozeman, Barry, and Stuart Bretschneider. 1986. "Public Management Information Systems: Theory and Prescription." *Public Administration Review* 46: 475–87. https://doi.org/10.2307/975569.

Caudle, Sharon L., Wilpen L. Gorr, and Kathryn E. Newcomer. 1991. "Key Information Systems Management Issues for the Public Sector." *MIS Quarterly* 15 (2): 171–88. https://doi.org/10.2307/249378.

Diamond, Jack. 2013. *Good Practice Note on Sequencing PFM Reforms*. Washington, DC: Public Expenditure and Financial Accountability (PEFA). https://www.pefa.org/resources/good-practice-note-sequencing-pfm-reforms.

Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. 2016. *The Age of Analytics: Competing in a Data-Driven World*. McKinsey Global Institute. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

Kamensky, John. 2019. "Why Engagement Matters and How to Improve It." *Government Executive*, April 19, 2019. https://www.govexec.com/management/2019/04/why-engagement-matters-and-how-improve-it/156424/.

Newcomer, Kathryn E., and Sharon L. Caudle. 1991. "Evaluating Public Sector Information Systems: More Than Meets the Eye." *Public Administration Review* 41 (5): 377–84. https://doi.org/10.2307/976406.

NIH (National Institutes of Health). 2018. "FEVS Privacy." NIH Videos, May 4, 2018. Video, 2:14. https://www.youtube.com/watch?v=k2umYftXKCI.

OPM (Office of Personnel Management). 2017. "Profile of Federal Civilian Non-Seasonal Full-Time Employees." US Office of Personnel Management, US Government, September 30, 2017. https://www.opm.gov/policy-data-oversight/data-analysis-documentation/federal-employment-reports/reports-publications/profile-of-federal-civilian-non-postal-employees/.

OPM (Office of Personnel Management). 2021. *Governmentwide Management Report: Results from the 2020 OPM Federal Employee Viewpoint Survey*. Washington, DC: US Office of Personnel Management, US Government. https://www.opm.gov/fevs/reports/governmentwide-reports/governmentwide-management-report/governmentwide-report/2020/2020-governmentwide-management-report.pdf.

Runkler, Thomas A. 2020. *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. Cham, Switzerland: Springer Vieweg.

World Bank. 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank. https://www.worldbank.org/en/publication/wdr2016.

World Bank. 2021a. *Europe and Central Asia Economic Update, Spring 2021: Data, Digitalization, and Governance*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/35273.

World Bank. 2021b. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. https://www.worldbank.org/en/publication/wdr2021.

CHAPTER 10

Government Analytics Using Human Resources and Payroll Data

Rafael Alves de Albuquerque Tavares, Daniel Ortega Nieto, and Eleanor Florence Woodhouse

SUMMARY

This chapter presents a microdata-based approach for governments to improve their strategic human resource management and fiscal planning. Using a series of examples from Latin American countries, the authors demonstrate how basic statistics created using payroll and human resource management data can help policy makers gain insight into the current and future state of their government's wage bill. The authors argue that this constitutes an important first step toward tapping the potential of existing bodies of payroll and human resources microdata that are currently underused. This approach can help policy makers make difficult decisions by breaking down the causes of problems and putting numbers to the ways in which certain policy choices translate into longer-term consequences.

ANALYTICS IN PRACTICE

- *Data collection practices.* It is recommended that where possible, governments centralize their human resources (HR) data collection systems and render these data accessible to insights teams. If such data do not exist, even in a disparate fashion, we strongly advise governments to begin collecting, in a centralized manner, payroll and human resources management information system (HRMIS) microdata. We advise governments to make these data public where possible (anonymizing the data, naturally) to improve transparency.
- *Fiscal planning.* We advocate for better integration of HR data analysis with fiscal planning. To be able to leverage payroll and HRMIS microdata, governments must encourage civil servants from the treasury and HR department(s) to collaborate more closely. This could be achieved by allocating dedicated

The authors' names are listed alphabetically. Rafael Alves de Albuquerque Tavares is a consultant at the World Bank. Daniel Ortega Nieto is a senior public sector management specialist at the World Bank. Eleanor Florence Woodhouse is an assistant professor in the Department of Political Science and School of Public Policy, University College London.

portions of civil servant workload to the task of sharing and analyzing data or by creating dedicated interdepartmental roles to push forward and undertake the collection and analysis of payroll microdata for strategic human resource management (SHRM).

- *Service delivery.* By better integrating HR data and wage bill planning, policy makers can improve service delivery to citizens. For example, projections of which categories of public servants will retire or transfer across posts allow managers to identify where additional resources will be required to ensure the continuity of service provision. This logic can be extended to the integration of personnel data with the wider dataverse available to policy makers. For example, data on demographic changes among citizens allow policy makers to predict changes in service demands. The interaction of these analytics on the demand and supply sides of service delivery allows policy makers to use their resources intelligently.
- *Insulation of SHRM and fiscal planning.* Political considerations can impede the implementation of successful SHRM and fiscal planning. We recommend that governments insulate certain aspects of planning offices' work from the ebb and flow of politics. This could go hand-in-hand with our second lesson, to carve out explicit portfolios or roles dedicated to collecting and analyzing HR microdata, by ensuring that this work is undertaken by public servants reporting to an independent agency rather than to a minister.

INTRODUCTION

This chapter offers policy makers ways to use HR microdata to improve SHRM and fiscal planning. More specifically, practical examples are presented of how to use payroll and HRMIS data to strengthen wage bill projections, gain better insights into the dynamics of the public sector labor market, and strengthen evidence-based personnel policy. The approach offers ways to tap the potential of HRMIS data that are widely available but underused. The types of analysis that we propose can support public sector managers in their decision-making by rendering explicit some of the consequences of key human resource management (HRM) choices and simulating distinct scenarios.

The approach described uses administrative data related to individual employment and compensation to model the dynamics of the public sector workforce and its associated costs. By applying an analytical lens to the administrative data the public sector holds on its employees, these data become a means of better understanding the characteristics of public administration. This includes determining simple statistics, such as the ratio of pay and allowances across distinct groups of employees, the different job placements and training opportunities secured by officials across time and institutional environments, and extrapolations of core variables, such as the wage bill under current laws and regulations.

With these generally straightforward statistics—to which any government with an HRMIS should have access—significant improvements can be made to addressing potential HRM shortcomings and related fiscal issues, including strategic workforce planning, replacement rates, salary inequalities within and across government agencies, the distribution of pay-for-performance benefits, the retirement of personnel, and projections of payroll costs, among others.[1]

Data analytics based on personnel data have proliferated in recent years and enable organizations to understand analytics across the HRM cycle (Davenport 2019)—from the attractiveness of distinct positions advertised by an organization (measured by the number of applicants) to diversity and inclusion (measured by, for instance, the ethnic diversity in different ranks of an organization's hierarchy), to name just two examples. Yet, as outlined in chapter 9 of the *Handbook*, many government organizations lack an HRMIS with which to register these data. For this reason, we limit the HRMIS analysis in this chapter to personnel data that are often more readily available and registered by governments—such as age (by registering date of birth) and gender—while acknowledging that this only presents a small fraction of the HRMIS data analytics possible with more widely available data.

Common Sources of Microdata

Two key terms used throughout this chapter are the *government wage bill* and *payroll and HRMIS microdata*. Harborne, Bisca, and Dorotinsky (2017, 267n48) define the government wage bill as

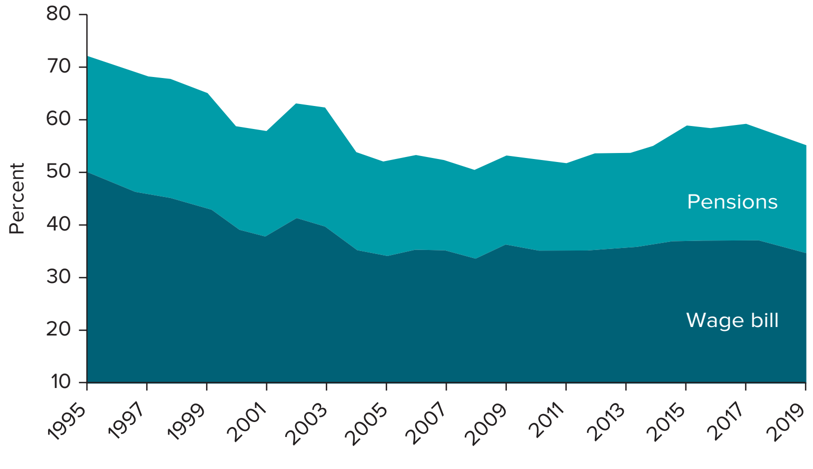
> the sum of wages and salaries paid to civilian central government and the armed forces. Wages and salaries consist of all payments in cash (no other forms of payment, such as in-kind, are considered) to employees in return for services rendered, before deduction of withholding taxes and employee pension contributions. Monetary allowances (e.g., for housing or transportation) are also included in the wage bill.

Pensions, by contrast, are generally not included in the wage bill. Pensions remain, however, an important share of a government's payroll. Indeed, a recent report has found that subnational governments in Brazil have been under growing pressure from pension expenses, at times spending the same amount of money on pensions as on the wage bill (World Bank 2022). Figure 10.1 presents the percentage of the state budgetary ceiling allocated to pensions and the wage bill in Brazil. In 2019, over 20 percent of the budgetary ceiling was spent on pensions, almost the same proportion as the wage bill itself. The same type of analysis that we perform for the wage bill in this chapter could be replicated for pensions. For example, projections of pension expenses could aid governments in making strategic decisions to prepare for and potentially reduce their burden.

Payroll and HRMIS microdata are two separate data sources that we leverage in our analyses. Both capture individual-level information about employees and should be easily available to most governments. Payroll data include information pertaining to an employee's contract (job position, contract type, etc.), personal details (date of birth, national insurance number, address, etc.), salary and tax details (amount and date of payments, tax codes, etc.), and leave, holidays, and benefits. These data are generally collected by the HR or finance department that administers the salaries of public employees. For this reason, these data are automatically updated because they must reflect promotions or changes in role or leave allocations.

HRMIS data, on the other hand, can be used to enrich payroll data because they also capture information such as an employee's gender and educational qualifications or prior professional experience. However, they tend not to be updated in the same way as payroll data because they are usually taken as a

FIGURE 10.1 Wage Bill and Pensions as a Percentage of Subnational States' Budget, Brazil, 1995–2019



Source: World Bank 2022.

snapshot at the recruitment stage and thus capture an employee at only a single point in time (for example, if the employee earns a degree after starting a position, this will not necessarily be reflected in the HRMIS data).

When we refer to HR data more broadly, we refer to the combination of payroll and HRMIS data for active (or currently employed) civil servants.[2] While many governments have access to data collected via their HRMIS, they sometimes struggle to extract and use them to their full potential. This can be due to a number of issues, including outdated HRMIS or analytical capacity, the decentralization of HR departments (which means that central government administrators only have access to partial data), or a lack of long-term strategic HR planning. In the section "Payroll and HRMIS Microdata and Related Challenges," we offer insights into how to get these data into shape for analysis, and in the section "Descriptive Statistics," we discuss how to undertake simple and effective analyses using said data to improve SHRM.

Capitalizing on Government Microdata

We focus on SHRM and the government wage bill for a number of reasons. First, the wage bill has considerable fiscal impact because it represents a significant portion of government spending: around one-fifth of total spending, according to the International Monetary Fund (IMF) (Gupta et al. 2016, 2), or around 9–10 percent of gross domestic product (GDP) and roughly a quarter of general government expenditures, according to the World Bank (Hasnain et al. 2019, 8). Second, it is likely that across the globe, pressures on wage spending will increase in the coming years and decades because "advanced economies are facing fiscal challenges associated with aging populations while also needing to reduce high public debt levels" and because "emerging markets and low-income countries have pressures to expand public service coverage in the context of revenue and financing constraints and the need for higher public investment" (Gupta et al. 2016, 1). Thus, in order to ensure that they are able to continue to deliver essential public services when facing increasing financial constraints, governments must invest in fiscal planning and SHRM.

The approach we propose allows government organizations to better leverage their HR data and make use of evidence for decision-making. Such strategic use of HR data can also have a significant fiscal impact, helping to avoid short-termism and here-and-now pressures that may cast a long shadow over government organizations' ability to undertake their work and offer the best services to citizens under government budget constraints. A case in the Brazilian state of Alagoas offers a good example, illustrating the potential of using payroll microdata to provide empirical evidence for the pros and cons of policy decisions. Here, estimates of a decreasing pupil-per-teacher ratio helped inform the government's decision to recruit fewer teachers while maintaining the quality of the public education delivered to its citizens by opening up fiscal space to better provide other, more needed public services.

One of the major advantages of applying an analytical lens to SHRM and wage bill data is that it supports governments to improve workforce and fiscal planning jointly and in a coordinated way. These two aspects of government work should occur in tandem, but in practice, this is rarely the case. *Workforce planning* "is a core HRM process that helps to identify, develop and sustain the necessary workforce skills" and that "ensures that the organisation has the right number of people with the right skills in the right place at the right time to deliver short and long-term organisational objectives" (Huerta Melchor 2013, 7). The ultimate goal of public sector workforce planning is to optimize the number and type of staff employed and the budget of the department or government in question. By the *type* of staff, we mean their professional skill set: are they able to contribute to completing the mission of the organization they serve? Identifying the needs of an organization and the HR required to achieve its goals is the heart of strategic workforce planning (Jacobson 2009; Kiyonaga 2004; Selden 2009): "a goal of workforce planning is to identify the gap between those needs and the available labor supply for government to continue providing quality services and fulfill its mission" (Goodman, French, and Battaglio 2015, 137). *Fiscal planning*, by contrast, refers to the way in which governments use their spending and taxation to influence the economy. As such, it can be improved by developing a better understanding of when certain groups of employees are going to be hired or retire, for example, allowing for more accurate revenue forecasting, which influences the budget approved by the government. One area that the IMF has identified as important for improving fiscal planning is precisely strengthening links between wage bill management—specifically, wage determination processes—and fiscal frameworks (Gupta et al. 2016, 2).

Strengthening Traditional Approaches

One additional application of our microdata-driven approach is to help bridge the gap between *macroanalysis* and traditional *functional reviews*, two common approaches to the analysis of the government wage bill and the distribution of work functions across the civil service, respectively. The former relies on macro-level analysis that leverages indicators such as the wage bill as a share of GDP and government employment per capita to gauge the appropriate size and cost of the civil service. By relying on macro indicators, these analyses have often led to simplistic policy prescriptions in the context of fiscal crises.

The latter strain of analysis relies on functional reviews. Using mostly legal documents, regulations, and interviews, these reviews scrutinize the goals, tasks, and resources of units inside the government to improve efficiency and effectiveness. Functional reviews thus have multiple goals but generally aim to assess how work is distributed across the civil service and to identify potential duplication of work through the functions performed by different departments. The analysis may produce results that are not integrated with an overarching strategy of reforming the civil service based on fiscal constraints.

By undertaking microdata analyses, one can complement functional reviews by not only looking at government functions but also gaining greater insight into other relevant dimensions of government organization, such as staffing and competencies. For instance, if one undertook a functional review and discovered that two departments perform similar functions, a parallel microdata-powered analysis could identify the distribution of competencies across the two departments. Perhaps one department has a natural advantage in taking full responsibility for the function because of the greater strength of its staff. Or perhaps there needs to be a redistribution of staff to more effectively distinguish the roles and activities of the two departments.

Micro-level analysis can be used to help reconcile and complement the fiscally oriented nature of macroanalysis and the flexible and detailed nature of functional reviews. This can be done through the use of simple descriptive statistics, such as the drivers of payroll growth (variation in total payroll, wages, and number of employees), the distribution of the workforce according to levels in the career ladder, and progressions and promotions over time and how much they cost, among others, and via a model-based simulation of the wage bill with the fiscal impacts of policies that improve and consolidate wage bill spending. One contribution that our chapter makes is to demonstrate some of the potential uses of and synergies between payroll and HRMIS data. By breaking down data silos, governments can start to better leverage data that are already at their disposal to gain insights into how to manage certain processes, such as adjusting the wage bill and improving fiscal planning.

In short, our chapter aims to lay out a practical, practitioner-friendly approach to the government wage bill that can improve SHRM and fiscal planning with relatively little technical expertise and data that should be accessible (with a relatively low cost of extraction) to any government with an HRMIS. This approach offers significant advantages in helping governments to use the untapped potential of lakes of payroll and HRMIS microdata and, more broadly, to use evidence in order to navigate difficult policy decisions.

STRATEGIC HUMAN RESOURCE MANAGEMENT AND FISCAL PLANNING

The public administration literature on SHRM focuses on identifying how it is used across different levels of government (Choudhury 2007; Goodman, French, and Battaglio 2015; Jacobson 2010), evaluating the effectiveness of different types of SHRM (Selden 2009; Selden and Jacobson 2007), and determining which factors influence the successful implementation of SHRM strategies (Goodman, French, and

Battaglio 2015; Pynes 2004). However, it is widely recognized that there is a paucity of empirical research on public sector SHRM (Choudhury 2007; Goodman, French, and Battaglio 2015; Reitano 2019), with much of the existing literature being normative in nature, relying on small samples, or being significantly dated. Moreover, the extant literature has a strong focus on the United States, with little to no evidence from the rest of the world.[3] Broadly, SHRM and wage bill data are underused as a source of analytics data for better understanding the characteristics and nature of public administration and public service.

One central finding of the existing literature is that many local governments do not have workforce plans in action (Jacobson 2010). In their survey of the largest US municipal governments, Goodman, French, and Battaglio (2015, 147) find that "very few local governments make use of comprehensive, formal workforce plans."[4] This is confirmed by other studies focusing on specific geographical regions, such as Jacobson (2010) and Frank and Zhao (2009). Local governments have been shown to lack the technical know-how and resources required to undertake SHRM (Choudhury 2007; Huerta Melchor 2013; Jacobson 2010). Small local governments, in particular, often lack the fiscal, professional, and technical expertise to innovate successfully (French and Folz 2004). For this reason, local governments may shy away from more complex econometric approaches to processes such as budget forecasting because they lack the know-how (Frank and Zhao 2009; Kavanagh and Williams 2016). This is precisely where our approach comes into its own. With very few, simple statistics that any public organization with an HRMIS should have access to, local and national HR departments can make a marked improvement in the use of their SHRM data.

Although the lack of capacity for SHRM seems to be most acute at the local level, it has also been documented in national governments. The Organisation for Economic Co-operation and Development (OECD) describes how its member states have "experienced problems with developing the necessary institutional capacity to engage in workforce planning both at the level of the central HRM body and the budget authority, and at the level of HR departments, professionals and front line managers" (Huerta Melchor 2013, 15). Strategic human capital management was identified by the US General Accounting Office (GAO) in 2001 as a governmentwide high-risk area because many agencies were experiencing "serious human capital challenges" and the combined effect of these challenges placed "at risk the ability of agencies to efficiently, economically, and effectively accomplish their missions, manage critical programs, and adequately serve the American people both now and in the future" (GAO 2001b). Strategic human capital management remains "high risk" to this day and is proving difficult to improve upon, with "skills gaps . . . identified in government-wide occupations in fields such as science, technology, engineering, mathematics, cybersecurity, and acquisitions" and "emerging workforce needs in the wake of the COVID-19 pandemic" (GAO 2021). For this reason, simple, timely ways to improve SHRM—such as the approach that we propose—are urgently needed.

Another important obstacle to successful SHRM and fiscal planning highlighted by the existing literature is political considerations. Successful SHRM requires support and planning from top management because data have to be systematically collected and analyzed over long periods of time. If elected figures are more interested in satisfying concerns "here and now" and are unwilling to invest in longer-term HRM and fiscal strategies, this can pose a significant challenge. This is especially true in smaller local governments, where leadership tends to be more centralized and informal and where, frequently, no separate personnel departments exist (Choudhury 2007, 265). Thus, local governments appear more susceptible to a lack of long-term planning because they are more likely to lack technical know-how or to face direct political pressures (Kong 2007; Wong 1995). It seems especially important, then, to take into consideration the nature and size of a government when examining SHRM (Reitano 2019). As Choudhury (2007, 265) notes, "the conditions of effective human resource management at the federal, state, or large urban levels often are not a good fit for smaller jurisdictions." That said, we believe that our approach can cut across different levels and sizes of government because it relies on data that should be widely available to small and large governments alike.

The extant literature has also paid significant attention to what Goodman, French, and Battaglio (2015, 147) refer to as the "perfect storm" of "human capital crisis that looms for local governments due to the number of employees who will be eligible for retirement or early retirement in the near future," which "offers significant opportunity for the use of workforce planning to help with forecasting the labor pool and fine tuning recruitment efforts." Such a storm is still brewing in many countries around the world, both at the local and national levels. A significant number of studies explore the issue, which was becoming evident already in the early 2000s, with predictions that over 50 percent of US government senior management would retire as the baby boomer generation came to retirement age (Dychtwald, Erickson, and Morison 2004; GAO 2001a; Jacobson 2010; Kiyonaga 2004; Pynes 2009; Wilkerson 2007). Today, the issue of retirement, and the subsequent talent shortage due to a smaller pool of younger public officials available to replace retiring officials, is aggravated by significant budget constraints in the public sector. Agencies are "freezing recruitment and not replacing employees who retire. The problem is that countries continue cutting budgets without scaling back agencies' and ministries' missions, compromising the ability to serve the public" (Huerta Melchor 2013, 15). This makes SHRM all the more important because governments need to use their available resources as wisely as possible to continue to deliver essential services to the public.

Another obstacle to successful SHRM that has been identified by the existing literature is a lack of adequate data (Anderson 2004). For example, in the empirical context of Queensland, Australia, Colley and Price (2010, 203) argue that there were "inadequate workforce data to support workforce planning and thereby identify and mitigate workforce risks." Several other studies echo the finding that public organizations in many countries find it difficult to obtain an accurate picture of their workforce composition (OECD 2007; Pynes 2004; Rogers and Naeve 1989). Colley and Price (2010, 204) note that "there is general agreement in the public service HR literature that the ideal is a centralised whole-of-service database to meet the common workforce planning needs of agencies. However, establishing such databases is time-consuming and costly, which limits its appeal to an incumbent government focused on short term budget and election cycles." Again, then, we see that political short-termism can obstruct successful SHRM before one even considers the lack of technical expertise or time and capacity that HR professionals may suffer (as we saw earlier in this section). Our proposed approach speaks to this obstacle to SHRM because it requires only a few basic statistics to better leverage HR data.

In addition to the direct challenges of enacting SHRM, SHRM and fiscal planning also interact in important ways. In order to enact more effective and sustainable fiscal planning, there are numerous ways in which the management of government wages can be improved and better take fiscal concerns into consideration. For example, the IMF notes that wage bill increases have been shown to be associated with worsening fiscal balances: "rather than crowding out other items in the budget, increases in the wage bill have on average been associated with increases in other government spending and with a deterioration of the overall balance" (Gupta et al. 2016, 14). For this reason, policy makers should be especially wary of increasing the wage bill when the budget is tight. Furthermore, if SHRM is not undertaken so as to employ the right type and amount of workers, this can have a negative fiscal impact. If there is a wage premium in the public sector, this can "increase private production costs, including wage costs, as well as result in additional 'deadweight losses' associated with distortionary taxation" (15). In fact, wage penalties can also have detrimental fiscal effects because difficulty recruiting and retaining qualified workers adversely affects the quality of publicly provided goods and services and can also contribute to corruption (Hasnain et al. 2019, 8). For this reason, public sector salaries should be calibrated to those of the private sector for comparable jobs and adjusted according to broader changes in the population, society, and the economy at large (Somani 2021). Indeed, advanced economies have been found to struggle to adjust employment levels in response to demographic changes—such as the decline in school-aged children, which led to an oversupply of teachers (Gupta et al. 2016, 20)—which can lead to significant fiscal concerns that could be avoided with a more forward-thinking HRM strategy.

PAYROLL AND HRMIS MICRODATA AND RELATED CHALLENGES

Before delving into what analysis can be done with payroll and HRMIS microdata, it is important to further discuss the kind of data we are talking about and the type of variables one can extract from such data sources. We describe payroll microdata first, before turning to HRMIS microdata. Payroll microdata are drawn from the administrative data sets that governments use to follow and register the monthly compensation of civil servants and their underlying items. They usually cover most of the government's contracts with its employees and sometimes contain demographic characteristics of civil servants and their occupational information (for example, the department or unit where the civil servant is located, the type of contract, the date of their entry in the civil service, etc.). In some contexts, sets of information are collected independently by different teams. HRMIS microdata, on the other hand, as anticipated in the introduction, are additional data, often collected by recruitment units, that can enrich payroll data with information about employees' gender, education level, and professional sector, for example. To undertake our analyses, we combine these two types of microdata.

In table 10.1, we present an example of a hypothetical combined payroll-HRMIS microdata set with the main variables (columns) and observations (lines) needed for the type of analysis we propose in this chapter. This table represents the minimum data required to undertake the analyses we propose. Each line represents an individual and that individual's respective contract with the government, and each column points to some relevant variable for analysis, such as the unit where the civil servant is located, age, gender, date of entry in the civil service, type of contract, and so on. An individual might have more than one contract with the government: for example, a teacher with two part-time job positions. Ideally, the database should have information about the government's employees for the last 10 years so that one can retrieve variables of interest based on historical data (for example, the average number of years of service before retirement).

Ideally, governments should have the aforementioned information for all their public employees readily available, but based on our experience working with several governments from Latin America and the Caribbean (LAC) countries, we know that governments face challenges when it comes to their wage bill microdata. These challenges can be organized along two dimensions. First, governments may not be able to collect information about all their employees, potentially leading aggregate figures to be wrong or biased. This can happen if wage bill microdata collection is not centralized and the information of some units or departments is missing in the data. In table 10.1, this would be reflected in fewer observations (lines) in the data than in the actual government bureaucracy. A second dimension relates to the number of different aspects that are collected to describe the bureaucracy. In table 10.1, these are captured in the number of columns in the data set. For example, in a recent analysis undertaken in the context of a project with a LAC country, the wage bill data did not have information about when public employees started their careers in the civil service, making it difficult to determine how experience in a position, measured by years of service, was related to wage levels and, as a consequence, the total cost of hiring a new civil servant for that position. With these issues in mind, practitioners should be cautious about what the available wage bill microdata can tell them about the current situation of bureaucracy in aggregate terms and about which aspects can be explored to provide insights for governments to better manage their SHRM and fiscal planning.

In figure 10.2, we propose a simple wage bill microdata "quality ladder" to help practitioners separate good data from bad data. We organize the ladder into five levels, with the first level representing the lowest-quality microdata and the fifth level the highest-quality microdata. At level 0, there is a missed opportunity for HRMIS data analysis because the *minimum required data* are not available (see table 10.1 for reference). This is because the information on public employees is scarce, inaccurate, inconsistent, and scattered across government units or career lines, such that almost any indicator or statistic based on such data would be wrong or biased. Statistically, it is impossible to draw inferences from incomplete data, especially where there are worries that the missingness is correlated with relevant features of the underlying values of the variables in the data. To see this, you need only think of some reasons why a government agency would not report HR microdata: because they lack the capacity or manpower to do so (in this case, only agencies with greater capacity would present their data, offering a skewed vision of the performance of the government at large) or because they are not mandated to do so and thus will not spend precious resources reporting HR data (again, in this case, drawing inferences from such data would give a misleading impression of the government at large because only the agencies with reporting mandates would provide their microdata for analysis).

TABLE 10.1 Example of Payroll + Human Resources Microdata Set Showing Minimum Data Required

March 2020																					
Year	Month	Individual ID	Job ID	Date of birth	Gender	Education	Date of entry	Type of contract	Area	Job position	Weekly working hours	Career level	Date of last progression	Base salary	Allowance 1	Allowance 2	Allowance 3	Vacation	Pension contribution	Gross wage	Net wage
2020	March	100,001	1	1987-03-05	Female	Secondary	2015-01-01	Statutory	Education		20	III	2016-03-01	3,500	0	0	0	0	440	3,500	3,060
2020	March	100,001	2	1987-03-05	Female	Secondary	2010-11-10	Statutory	Health		20	IV	2013-03-01	1,000	0	100	0	0	110	1,100	990
2020	March	100,004	1	1980-06-04	Female	Superior	2008-03-02	Temporary	Safety		30	VI	2020-03-05	4,000	0	0	0	0	440	4,000	3,560
2020	March	100,005	1	1985-02-03	Female	No schooling	2009-05-03	Political appointee	Other		40	III	2020-03-31	2,500	200	0	0	0	275	2,700	2,425

March 2021																					
Year	Month	Individual ID	Job ID	Date of birth	Gender	Education	Date of entry	Type of contract	Area	Job position	Weekly working hours	Career level	Date of last progression	Base salary	Allowance 1	Allowance 2	Allowance 3	Vacation	Pension contribution	Gross wage	Net wage
2021	March	100,001	1	1987-03-05	Female	Secondary	2015-01-01	Statutory	Education		30	III	2016-03-01	3,500	0	0	0	0	440	3,500	3,060
2021	March	100,002	1	1980-06-05	Male	Primary	2010-11-10	Statutory	Health		40	IV	2013-03-01	1,000	0	100	0	0	110	1,100	990
2021	March	100,004	1	1980-06-04	Female	Superior	2008-03-02	Temporary	Safety		30	VI	2020-03-05	4,000	0	0	0	0	440	4,000	3,560
2021	March	100,005	1	1985-02-03	Female	No schooling	2009-05-03	Political appointee	Other		40	III	2020-03-31	2,500	200	0	0	0	275	2,700	2,425

Source: Original table for this publication.

At level 1, some analysis can be performed for the units or careers for which there are data available. However, for the reasons outlined above, such analyses must be applied only to the units or career lines for which data are available, and careful consideration must be given to why and how the missingness in the data is occurring. A good example of this is a situation where the wage bill data gathering is decentralized and some government units collect data while others do not. For instance, if only the education and health departments could fill table 10.1 with information about their employees, the analysis should be restricted to these units, and the government should start collecting data from other units of the government.

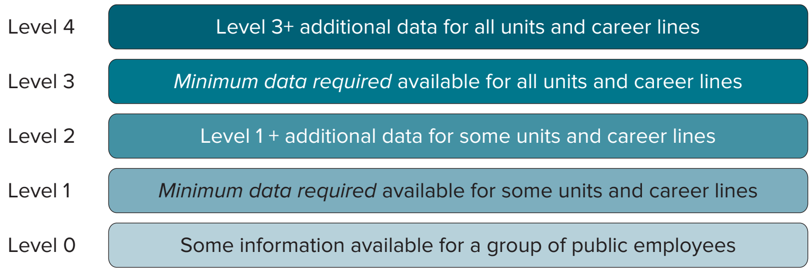
At level 2, not only is the basic information shown in table 10.1 readily available, but one is also able to connect these data with additional data sources and explore specific features of job contracts. Using the above example, this would be the case if the wage bill data for teachers could be connected to students' performance in standardized tests, allowing for the analysis of teachers' productivity in the public sector.

Level 3 illustrates a situation in which the information outlined in table 10.1 is collected for a large part of the bureaucracy in such a way that one can undertake an aggregate analysis of wage bill expenditures based on the microdata. In the section "Wage Bill Projections," we present an example of such an aggregate analysis, with a projection of the wage bill for future years based on data from the Brazilian federal government. We would like to note that levels 2 and 3 of the quality ladder can be ranked differently depending on the objectives of the analyses to be performed. For example, when analyzing the impact or value added of teachers on students' performance, having a productivity measure in the wage bill data for teachers can be especially useful. Given the fiscal nature of the analyses undertaken in this chapter, having a wage bill data set that allows the analyst to create aggregate figures is particularly important. Because of this, we have decided to rank a comprehensive data set for all civil servants without productivity measures above a data set with partial productivity measures in our quality ranking.

In level 4, one can not only undertake the analysis described in level 3 but can also merge other available data sources and connect them with the overall fiscal landscape of the government. Building on the example in level 2, one could assess both the fiscal impacts and the productivity impacts of adding a pay-for-performance scheme to teachers' compensation based on the performance of students on standardized tests.

Building an HRMIS that climbs the ladder described in figure 10.2 can be politically costly and requires sustained investment in the technical skills that underlie data management. The benefit is the improved understanding of the public sector that such an effort provides. The next section outlines the basic analytics for which such databases provide the foundation. Without the qualities outlined in figure 10.2, these analytics are undermined and can be distortionary. But with a sound foundation of quality and comprehensive data collection, these descriptives can support substantial fiscal efficiencies and improved service delivery. In the country cases described in the following section, these investments have paid off many times over.

FIGURE 10.2 Human Resources Microdata Quality Ladder


Level 4
Level 3+ additional data for all units and career lines
Level 3
Minimum data required available for all units and career lines
Level 2
Level 1 + additional data for some units and career lines
Level 1
Minimum data required available for some units and career lines
Level 0
Some information available for a group of public employees

Source: Original figure for this publication.

DESCRIPTIVE STATISTICS

In this section, we present descriptive statistics that can help policy makers gain insight into the current and future state of their government's wage bill. Along with each insight, we present examples from wage bill analyses that we undertook in different LAC countries. As mentioned before, the data required for these analyses should be available to any government that has an HRMIS. That said, we recognize that there are sometimes significant challenges to obtaining these data—especially in contexts where these data sets are not held centrally—and organizing them in order to undertake these analyses. We posit that there is great untapped potential in the payroll and HRMIS data that governments collect and propose a way to start using these data lakes, where they exist. Where they do not exist, we recommend starting to centralize HR microdata to undertake these types of analyses.

We present our proposed descriptive statistics in three groups. The first provides a general overview of the wage bill and HRM systems to give the reader a sense of how HR practices can impact the wage bill. The second addresses how these HR microdata can be used to identify inequalities in terms of representation within the public sector. Finally, the third proposes a way to address some of these inequalities by adopting a forward-looking perspective that applies changes to fiscal policy to avoid such inequalities or inefficiencies in the future.

General Overview of the Wage Bill and HRM

We first address how HRM practices can impact the wage bill and offer some examples of the insights that can be gained by better exploiting payroll and HRMIS microdata.

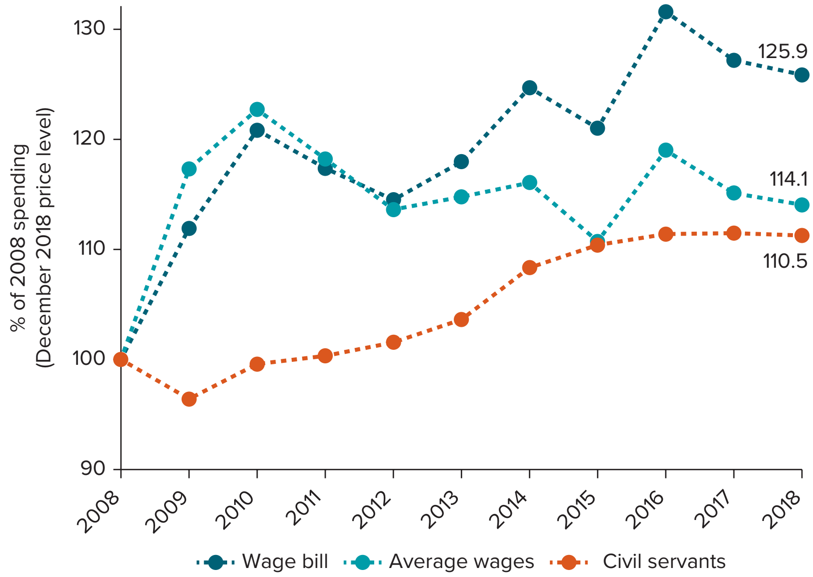
Drivers of Payroll Growth

Changes in wage bill expenditures can be attributed to changes in employment levels and changes in the average wages of civil servants. A wage bill increase resulting from increased employee hiring is usually accompanied by an expansion in the coverage of public services. Wage increases do not have an immediate impact on the provision of public services, but they may have a medium- and long-term impact on the attraction, retention, and motivation of civil servants that could enhance the productivity of the public service and lead to better service provision. Figure 10.3 presents a simple way of analyzing what is driving wage bill variation. By setting the starting year as a baseline, we can see in this example from the Brazilian federal government's wage bill that most of the increase in wage bill expenditures came from increases in civil servants' compensation. In fact, between 2008 and 2017, spending on Brazilian federal executive personnel rose by 2.9 percent per year in real terms. This growth was made up of a 1.8 percent increase in average salaries and a 1.2 percent increase in the number of public servants. This kind of figure can also be applied to analyze specific sectors and career lines in the government, such as the education sector and, within that, teachers. Undertaking a sector- or career-specific analysis is also a way of providing insights with partial data, since one should be cautious when making aggregate claims from microdata if not all wage bill data are available.

Breakdown of the Wage Bill by Sector

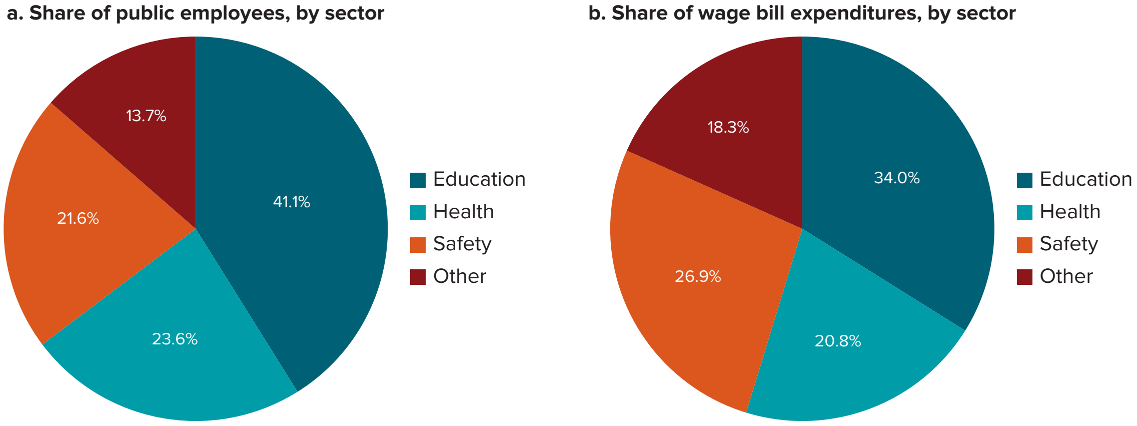
Breaking down the change in overall wage bill expenditures into changes in the number of civil servants and in average wages can also lend itself to understanding how civil servants and wage bill expenditures are distributed among priority areas. Extending the analysis to the sector level can shed light on the needs and targets of the government in areas such as education, health, and security. For example, in the case of the Brazilian state of Rio Grande do Norte (see figure 10.4), 86 percent of civil servants are distributed in

FIGURE 10.3 Drivers of Wage Bill Variation, Brazilian Federal Government, 2008–18



Source: World Bank 2019.

FIGURE 10.4 Wage Bill Breakdown, by Sector, Brazilian State of Rio Grande do Norte, 2018



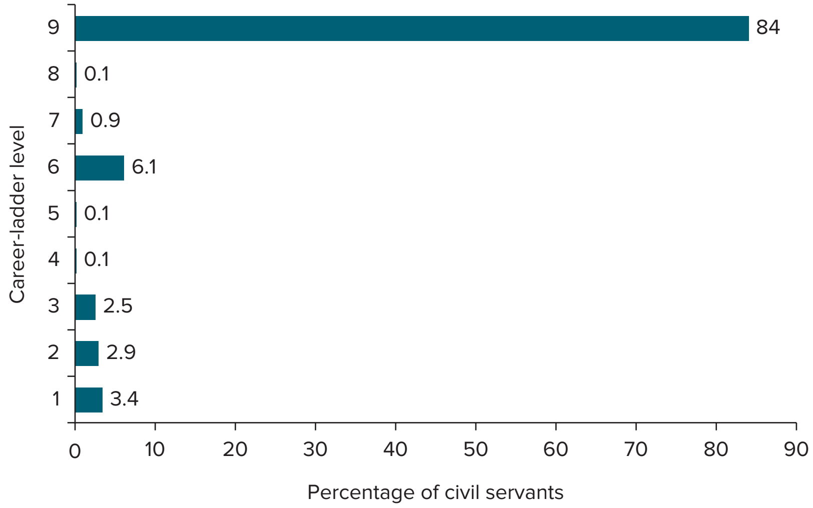
Source: Original figure for this publication.

priority areas, while the wage bill expenditures for these same sectors amount to 82 percent of the total wage bill spending. In particular, the education sector employs 41 percent of the public servants and accounts for 34 percent of the total wage bill.

Distribution of Civil Servants by Career-Ladder Level

Progressions and promotions along career ladders are a common path for governments to connect higher labor productivity to wage increases. Based on this link between productivity and wages, we can expect that a longer tenure in the civil service reflects a knowledge gain that should equip employees with better tools

FIGURE 10.5 Distribution of Civil Servants, by Career-Ladder Level, Brazilian Federal Government, 2018



Source: Original figure for this publication.
Note: The figure shows the career-ladder levels for the Tax Auditor career track in the Brazilian federal government.

with which to deliver public services. By analyzing how civil servants are distributed along career-ladder levels, policy makers can assess whether the ladder structure of civil service careers reflects increases in productivity. Ideally, we should expect to see a smooth distribution of civil servants across the different levels. In figure 10.5, we use as an example the career of tax auditors in the Brazilian federal government. We can see that more than 80 percent of public employees are in the final step of their careers, which suggests that there may be margin for improving the design of the career structure and the requirements for progression or promotion to better reflect labor productivity gains.

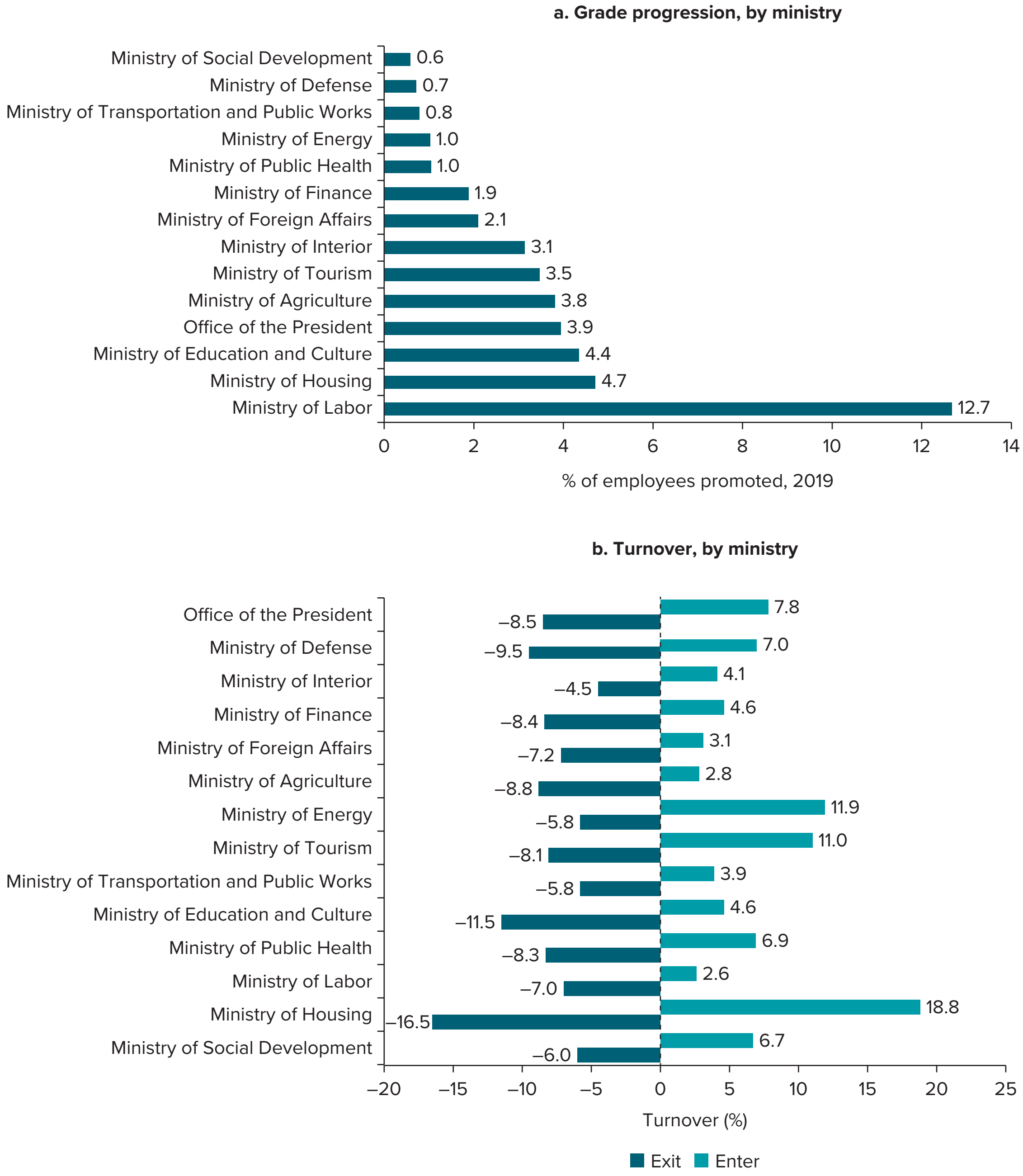
Strict Progression Rules and High Turnover of Civil Servants

On the other hand, situations where the rules for career progression and promotion are too strict may lead to difficulty retaining public employees, along with their acquired knowledge and expertise. To illustrate such a situation, we can examine the case of Uruguay's central administration, where civil servants are assigned to one scale (*escalafón*) and ministry. As movement across ministries and scales is rare and can only take place with special authorization, grade progression is the only career path available for civil servants. As a result, this limited room for vertical promotions may end up hindering productivity and motivation, as well as increasing turnover. In figure 10.6, we can see the share of employees who were promoted in 2019 (figure 10.6, panel a), and the turnover of employees (figure 10.6, panel b) by ministry in Uruguay's central administration. Less than 5 percent of employees were promoted to a higher grade in almost all ministries, while 7 percent of employees entered the central administration in 2019, and 6 percent exited that same year. In some ministries, the exit rate was even higher than the entry rate. This high turnover can be interpreted as a sign of the challenges in retaining civil servants. It also represents a hidden cost for the government due to the loss of expertise and the cost of training new staff.

Distribution of Pay-for-Performance Allowances

Pay-for-performance is a useful tool to stimulate productivity in the civil service. In theory, it rewards high-performing public employees and inspires low performers to perform better. However, there is much debate regarding the extent to which performance pay succeeds in improving civil service performance

FIGURE 10.6 Grade Progressions and Turnover, Uruguay Central Government, 2019

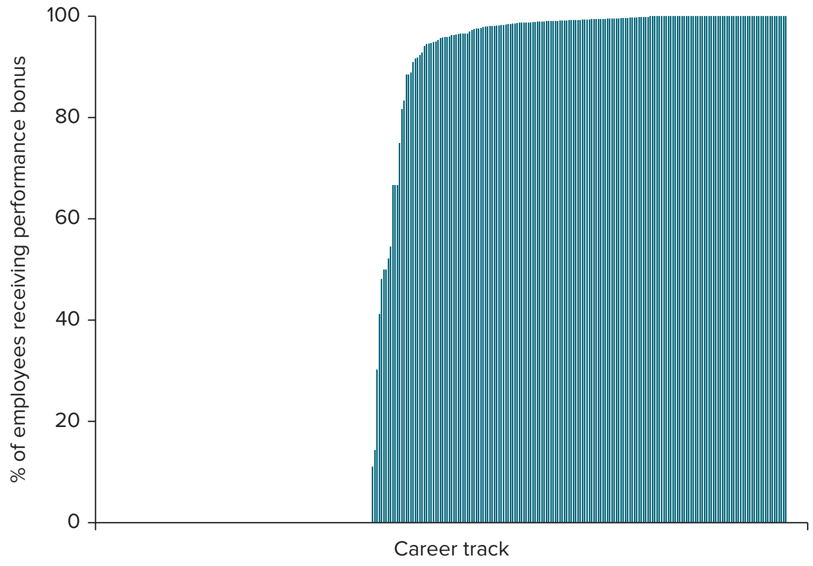

a. Grade progression, by ministry

b. Turnover, by ministry

Source: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, 2019.
Note: In panel b, the year of entry of an employee is the first year when her/his ID appears on the list of employees related to a specific job category.

(cf. Hasnain, Manning, and Pierskalla 2014). We posit that our approach can help policy makers understand whether pay-for-performance is working in the context in which they work. For example, one problem arises when *all* employees receive performance payment. In figure 10.7, using data from the Brazilian federal government, we display on the *x* axis all careers (each vertical line represents a specific career track) and on the *y* axis the percentage of each profession that received a performance bonus. We show that in 2017, at least 90 percent of employees received performance-related payment in 164 of the 187 careers that offered such schemes. This could indicate that the pay-for-performance scheme in question is not successful in differentiating between good and bad performers.

FIGURE 10.7 Distribution of Pay-for-Performance Allowances, Brazilian Federal Government, 2017



Source: Original figure for this publication.
Note: The *x* axis shows all career tracks in the Brazilian federal civil service, ranked by the *y*-axis variable.

Inequality in the Public Sector Wage Bill

Having given a general overview of key features of the public service, we turn to the use of HRMIS data to understanding inequalities in the public service. Such inequalities may come in different forms and have correspondingly different impacts on the efficiency or other qualities of the state.

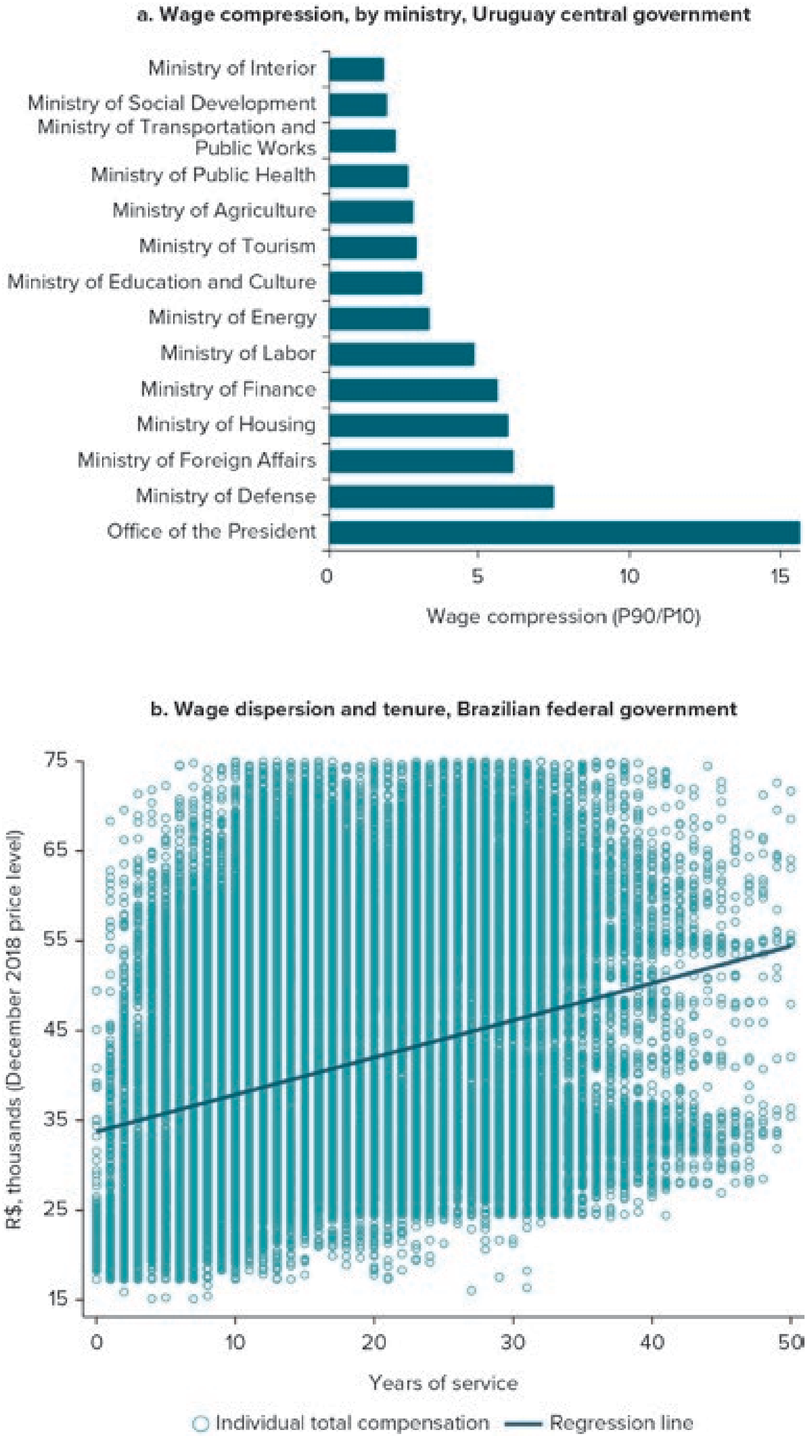
Representativeness

Many governments strive to recruit officials in a way that ensures the administration as a whole is broadly representative of the population it serves: for example, by having personnel from across the country's regions in rough proportion with the distribution of the population across those regions. Normatively, such considerations are important, given that in a democratic setting, bureaucracies should represent the populations they serve. Moreover, it has been empirically demonstrated that more representative bureaucracies—dependent on the policy domain—can affect important phenomena such as citizens' trust in the government and willingness to cooperate with the state (see, for example, Riccucci, Van Ryzin, and Lavena 2014; Theobald and Haider-Markel 2009; Van Ryzin, Riccucci, and Li 2017). Though there may be good reason for this principle not to hold strictly, HRMIS data allow the degree of representativeness of the administration to be accurately articulated and to act as the foundation of an evidence-based debate on the matter.

Pay Inequity

Inequality in payments in the public sector can reflect underlying differences in responsibilities or can be a sign that inconsistent compensation rules are being applied. For example, we expect the government to reward managers and high-performing employees with better compensation than entry-level civil servants, but we do not expect it to award significantly different levels of compensation to employees with the same attributes, jobs, and tenure, following the generally observed principle of equal pay for equal jobs. In the case of the Brazilian federal government tax auditors (see figure 10.8b), we can see that there is huge wage dispersion for similar workers. Gross pay can vary fivefold for workers with similar levels of experience, which is largely a result of nonperformance-related payments and is not related to base salary.

FIGURE 10.8 Measuring Pay Inequity in the Uruguayan and Brazilian Governments



a. Wage compression, by ministry, Uruguay central government

b. Wage dispersion and tenure, Brazilian federal government

Source: Panel a: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, February 2020. Panel b: Original figure for this publication.

Related to this is the need for governments to devise pay schedules that incentivize officials to keep exerting effort to rise up the career ladder while also being aware that equity in pay is a key issue for some officials' motivation. To measure inequality due to differences in responsibilities and career level, we can analyze the pay scale compression of the government's units.[5] Higher wage compression (a smaller wage gap between management level and entry level) is associated with greater difficulty in motivating personnel to progress through the public service because increased responsibility is not adequately compensated. For example, in the case of Uruguay (see figure 10.8a), wage compression in the central administration is low by international standards but varies greatly across ministries. Having low wage compression by international standards is good for equity, but the implications for civil servants' productivity and motivation are unclear. Low pay compression can generate positive attitudes across civil servants if responsibilities are also spread accordingly across the civil service, but it might also indicate that the salary structure is not sufficiently able to incentivize and reward workers' efforts or reward workers who have additional responsibilities.

Pay Inequity Based on Increasing Wage Components

A good compensation system should allow the government to select high-quality candidates and offer incentives to align each public servant's interests with those of society. Desirable characteristics of a payment system include the ability to link wage gains with skills and performance and the transparency of the wage components. Having a large number of salary components can hinder transparency and generate inequalities. For example, in the case of Uruguay's central administration, there are 297 different salary components, of which 53 are "basic" and 244 are "personal."[6] Each entity has some discretion to define the compensation its employees receive, thereby reducing transparency and potentially creating payment inequalities. From figure 10.9, we can see that this discretion is reflected in the distribution of personal payments (figure 10.9, panel b), which, unlike the distribution of basic payments (figure 10.9, panel a), follows a nonstandard distribution. The nonstandard distribution of personal payments suggests both a lack of transparency and an unequal pay structure, based on the increase of payment line items.

Wage Inequality by Gender

Gender equality is a key indicator of progress toward making the public sector workforce more diverse, representative, and innovative, and better able to provide public services that reflect citizens' needs. According to the OECD (2019), women are overrepresented in the public sector workforce of OECD countries.

However, this is not true across the globe; in fact, the Worldwide Bureaucracy Indicators show that public sector gender equity is correlated with country income (Mukhtarova, Baig, and Hasnain 2021). Part of the issue lies in providing similar levels of compensation for women and men where some systems discriminate against women. In some cases, the wage gap can discourage women from entering the civil service or applying for higher positions in an organization. In this sense, identifying potential gender wage gaps in the public sector is important to fostering the diversity of public employees. In figure 10.10, we analyze the gender wage gap in Uruguay's public sector workforce. The results suggest that overall, after controlling for working hours, age, type of contract, grade, tenure, and occupation, there is not a statistically significant gender wage gap, but this varies across ministries.

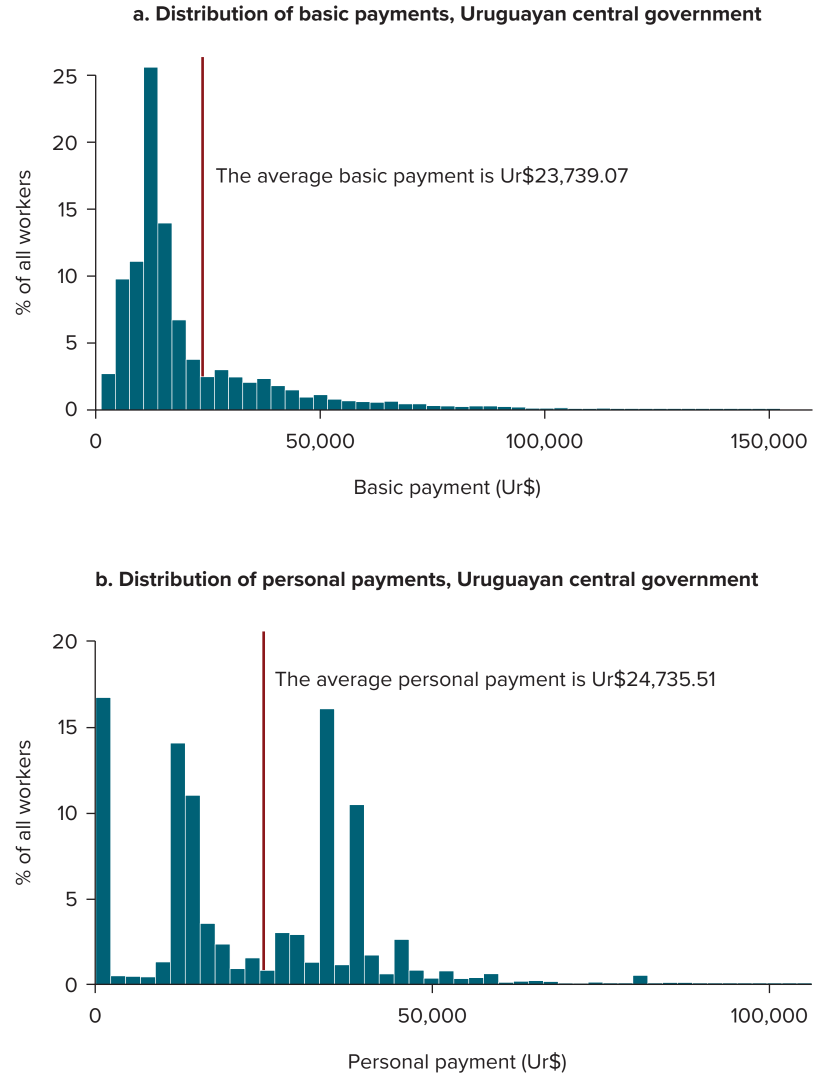
While there are many other margins of potential inequality in the service, and between the public service and the rest of society, these examples showcase the power of government microdata in identifying the extent and distribution of inequities across public administrations.

Fiscal Analysis

Having considered what the wage bill is, how HRM can affect it, and how HR practices can affect the character and equity of the bureaucracy, we now turn our attention to how such practices can affect the fiscal health of a polity.

Setting compensation schemes, including initial wages and wage increases related to progressions and promotions, is a key tool to attract, retain, and motivate civil servants. But it can also be a cause of long-term fiscal imbalance because public sector employees usually work for more than 15 years.[7] For example, careers with high starting salaries may attract qualified candidates, but when combined with slow or small wage increases related to progressions, this can lead to demotivated public employees. In such a situation, a reform

FIGURE 10.9 Inequity of Pay in Wage Components, Uruguay, 2019


a. Distribution of basic payments, Uruguayan central government

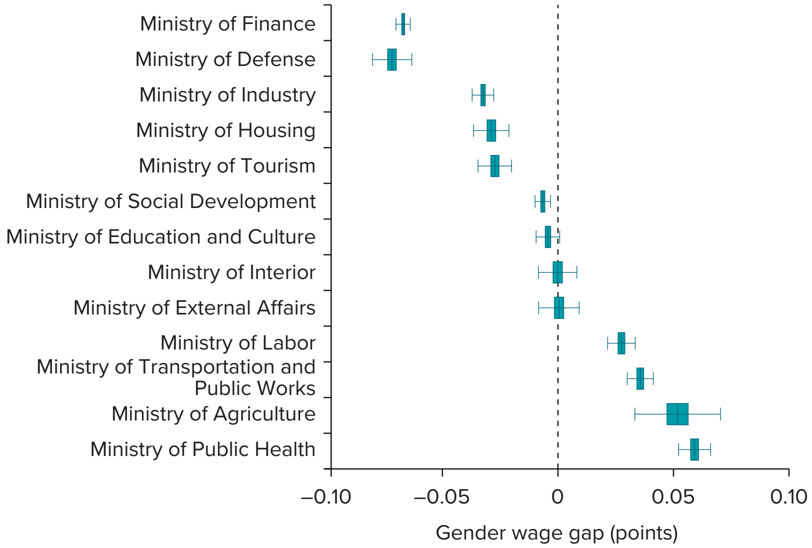
b. Distribution of personal payments, Uruguayan central government

Source: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, February 2020.

that kept starting salary levels high and increased the additional pay related to progressions and promotions might cause the wage bill to be fiscally unsustainable. By understanding the fiscal impact of current career compensation schemes and potential reforms, policy makers can better manage the public sector's HR in the long term. In figure 10.11, we present examples of how these compensation features can be visualized. In the case of the Brazilian state of Mato Grosso (figure 10.11, panel b), we find that for some of the careers, the first three progressions more than double public employees' salaries.

Besides starting salaries and wage increases, another important piece of information for policy makers implementing strategic workforce planning is when public officials retire. Getting a clearer picture of when public employees retire is of critical importance for strategic workforce planning and fiscal planning. One needs to understand who will retire and when in order to plan successfully for incoming cohorts of civil servants, both in terms of their numbers and the competencies they will need. When large numbers of public servants are all due to retire at the same time, this can offer a window of opportunity for policy reform. For example, in the case of the Brazilian federal administration, the World Bank projected, using 2017 data, that 22 percent of public servants would have retired by 2022 and that 40 percent would have retired by 2030 (see figure 10.12). This situation presented an opportunity for administrative reform to restructure career systems

FIGURE 10.10 Gender Gap, by Ministry, Government of Uruguay, 2010–20



Source: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, 2010–20.
Note: The graph shows regression coefficients and 95 percent confidence intervals for the interaction between the female dummy and for the ministry fixed effect. Each point represents the average salary difference with respect to the Ministry of Public Health, after controlling for workers' characteristics.

and rationalize the number of existing civil servants in order to better plan, both in terms of the workforce and in fiscal terms. The use of HR microdata to undertake this analysis helped to inform the debate about civil service reform.[8]

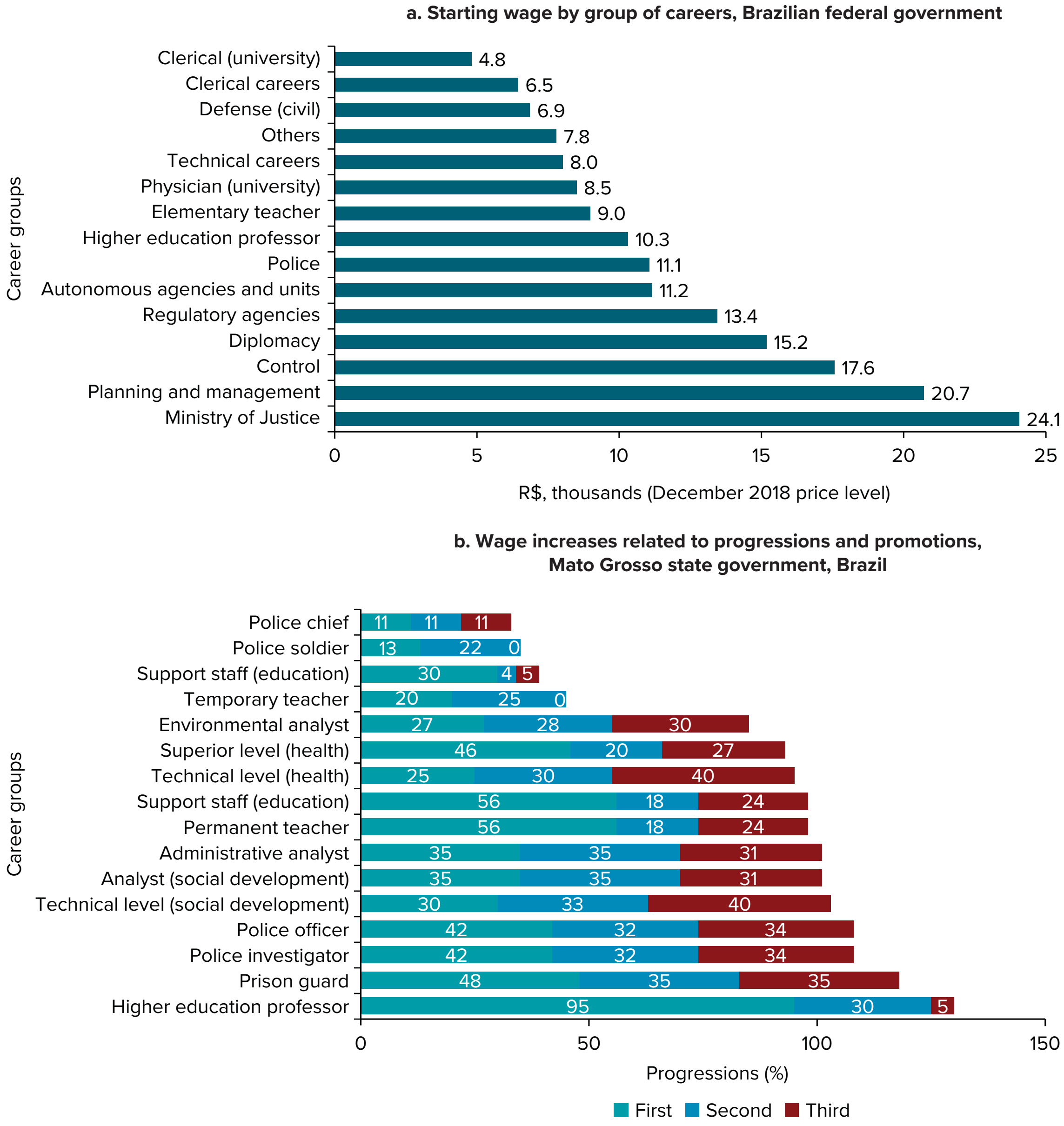
WAGE BILL PROJECTIONS

In this section, we present an HR-microdata-based model on the basis of the building blocks presented so far. With information about initial wages, wage increases related to career progressions, and expected dates of retirement, policy makers can project the expected fiscal impact of civil service reforms, the design of new careers, and fiscal consolidation policies. Using counterfactual scenarios can also help governments promote diversity and reduce inequalities in the civil service, fostering policies and services that better reflect citizens' needs.

Payroll and HRMIS microdata represent an important tool for the analysis of HR and fiscal policies. They can help policy makers lay out the trade-offs among competing policy objectives. For example, in the Brazilian state of Maranhão, the government sought to understand the fiscal impacts of wage increases for teachers along with increased recruitment of police personnel. By representing graphically the relevant statistics and comparing, first, the decreasing trend of the pupil-per-teacher ratio and its effect on the demand for new teachers and, second, levels of violence in the state when compared with its peers and the ratio of policemen per inhabitant, decision-makers obtained a more realistic picture of the available employment policies. In this section, we use some of the figures from the previous section to lay out the building blocks of a policy-oriented model for projecting wage bill expenditures. This model can help policy makers make difficult choices more transparent by showing the real costs and benefits of potential civil service reforms.

In practice, this is how we make the projections. First, we set up the HR microdata in a structure similar to the one described in the section "Payroll and HRMIS Microdata and Related Challenges"

FIGURE 10.11 Career Types and Wages, by Career Group, Brazil



a. Starting wage by group of careers, Brazilian federal government

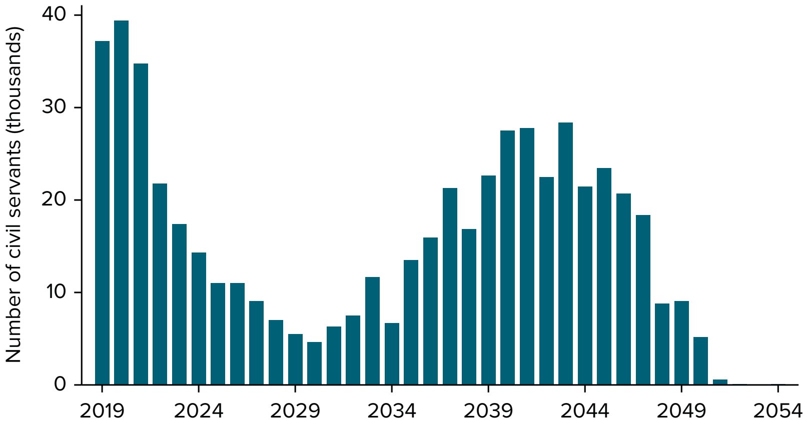
b. Wage increases related to progressions and promotions, Mato Grosso state government, Brazil

Source: Original figure for this publication.

and reported in table 10.1. Ideally, the database should contain payroll and HR information for the last 10 years. If monthly data are not available, it is possible to use a representative month of the year.[9] The wage bill data from previous years are then used to estimate some of the parameters of the model, and the most recent month or year data are used as a starting point for the projections.

Second, with the microdata set up, we group civil servants according to similarities in job position or common legal framework. The inputs of the government's HR managers are critical to this first part of the model because the number of groups set should both reflect the bulk of civil service careers and allow for more fine-grained policy options. In this sense, there is no "magic number" of groups; the number is based on context. In practice, we tend to cluster civil servants in a range of 5–20 groups.

FIGURE 10.12 Retirement Projections, Brazilian Federal Government, 2019–54



Source: Original figure for this publication based on Brazilian government data from 2008–18.

For example, in the case of some Brazilian states, we defined seven main groups: teachers, military police, investigative police, physicians, education support staff, health support staff, and others. These groups were defined to reflect the main public services Brazilian states are responsible for: public security, secondary education, and mid- to high-complexity health care. In another example, for the Brazilian federal government, we defined 15 career groups, which included university professors because Brazilian public universities are mostly federal.

Third, after setting the clusters of careers, we estimate some basic parameters for these groups using the microdata from previous years: the number of retirees by year for the following years, average tenure when retiring, initial wages, years between progressions or promotions, real increases in salaries related to progression or promotion legislation, real increases in salaries not related to progression or promotion legislation, and the attrition rate, which is the ratio of new hires to leavers. Some of these parameters were shown in the previous section. For example, figure 10.12 shows estimates for the number of retirees by year for the Brazilian federal government with data from 2008 to 2018.

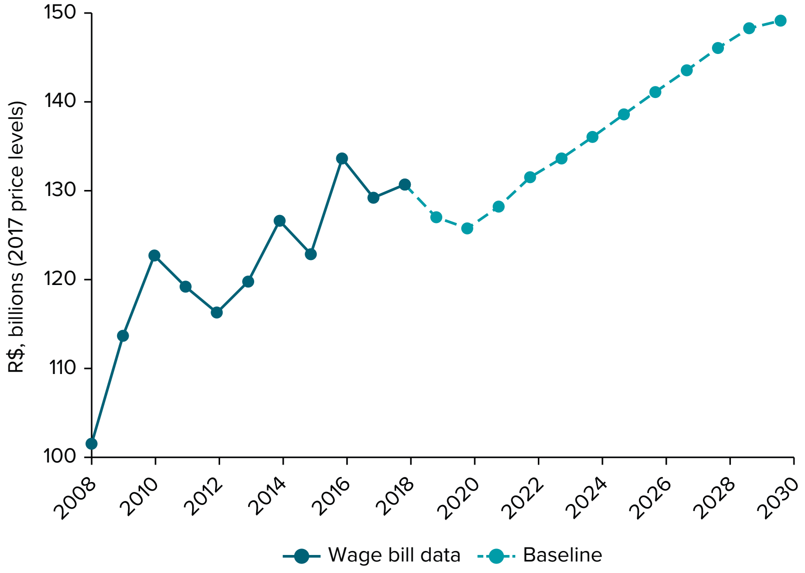
Fourth, we use the most recent month of the wage bill database and our estimated parameters to track the career path of current employees until their retirement and of the new civil servants who will replace retiring civil servants. Because of the fiscal nature of the model, the wage bill estimates tend to be less accurate for long-term projections. Based on experiences with LAC governments, we recommend using at most a 10-year span for projections. Using the estimated parameters, we come up with a baseline projection: the trajectory of wage bill expenditures assuming "business as usual," as extrapolated from the data on past years. In other words, we project the expected wage bill spending if we assume the same wage increases as in past years, the same expected tenure before retirement, and the same replacement rate of new civil servants per retiring employee.

Finally, after making a baseline projection of the wage bill, we are able to simulate reforms that implement changes to the estimated parameters. For example, if the government wants to analyze the fiscal impacts of a reform that increases the recruitment of teachers, we simply change the rate of replacement of the career group of teachers. In another example, if the government wants to consolidate wage bill expenditures by freezing wages for the next two years, we change the parameter for salary increases that are not related to progressions or promotions. The list of potential policy scenarios includes hiring freezes or targeted pay increases for specific classes of employees. The model is meant to be flexible to adapt to the government's needs so policy makers can test different reform options and hypotheses.

Example from the Brazilian Federal Government

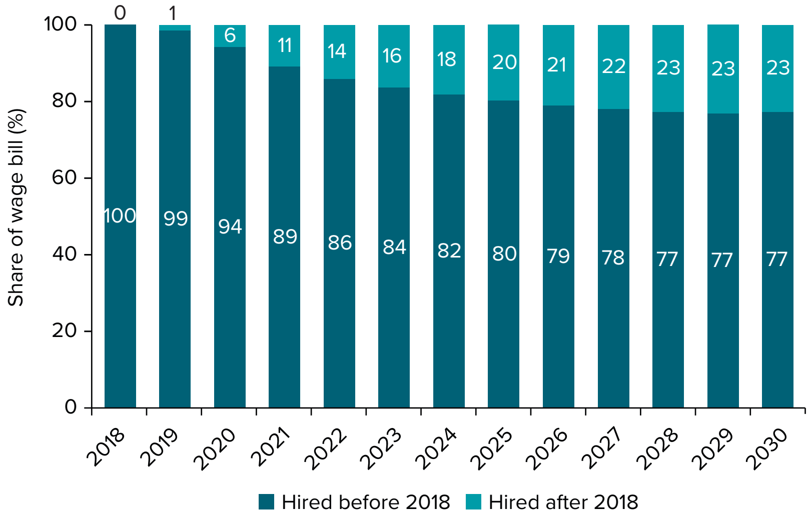
To exemplify the use of the model, in this section, we present wage bill projections for the Brazilian federal government for the period 2019–30, which were undertaken using HR microdata from 2008 to 2018. For example, figures 10.11, panel a and 10.12 in the previous section are graphical representations of the starting wages and the number of retirees by year, respectively. Figure 10.13 presents the baseline projection of the wage bill, and figure 10.14 provides a decomposition of the wage bill projection across current and new employees. Brazil is something of an outlier among LAC countries in that it has very high-quality administrative data, making it a good example of the more advanced types of analyses one can undertake with HR microdata once a comprehensive, centralized data collection system has been put in place.

FIGURE 10.13 Baseline Wage Bill Projection, Brazilian Federal Government, 2008–30



Source: Original figure for this publication based on Brazilian government data from 2008–18.

FIGURE 10.14 Decomposition of Wage Bill Projection between Current and New Employees, Brazilian Federal Government, 2018–30



Source: Original figure for this publication based on Brazilian government data from 2008–18.

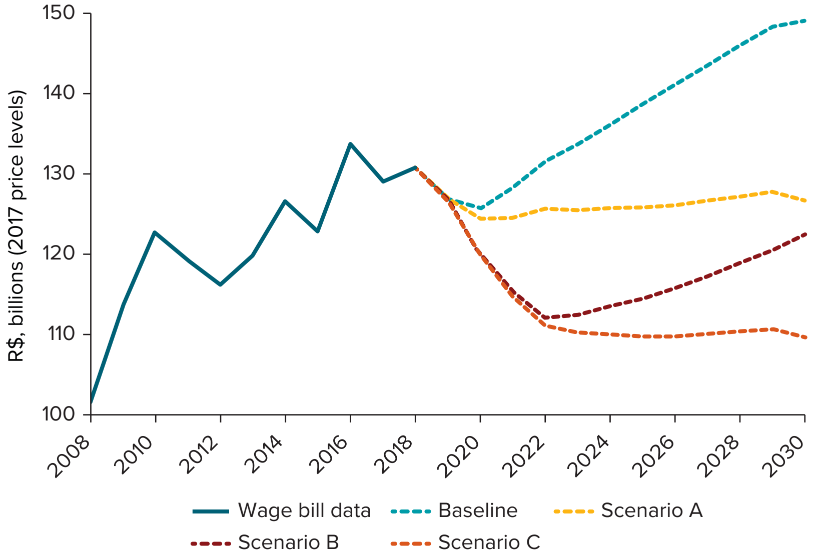
After projecting a baseline scenario for wage bill expenditures in the coming decade, we are able to compare it to different policy scenarios. To better organize reform options, we can separate them into pay-related and employment-related reforms. In the context of Brazil, the federal government's main objective was to simulate reforms that could lead to fiscal savings. We presented nine policy options, two of them related to employment reforms and the other seven related to pay policies. Based on these specific policies, we projected the following scenarios, each with a set of pay-related and employment-related policies:

- Scenario A: Replacement of 100 percent of retiring employees and no real salary increases for 10 years.
- Scenario B: Replacement of 90 percent of retiring employees and no nominal salary increases for the first three years.
- Scenario C: Replacement of 80 percent of retiring employees and no nominal salary increases for the first three years, and after that, no real salary increases for the next seven years.

Figure 10.15 provides a graphical presentation of the baseline projection along with the three outlined reform scenarios. In scenario A, a policy of no real wage increases is implemented starting in 2019. Since the *y* axis measures wage bill expenditures in real prices for 2017, the policy of correcting salaries only for inflation leads to an almost steady line in the chart. Scenarios B and C implement tighter policies, with a nominal freeze in salaries for the first three years starting in 2019, along with fewer hires of new employees to replace retiring civil servants. The bulk of the difference in savings between scenarios B and C comes from the years after 2022, in which scenario B returns to the baseline wage bill expenditures, while in scenario C, salaries are corrected for inflation.

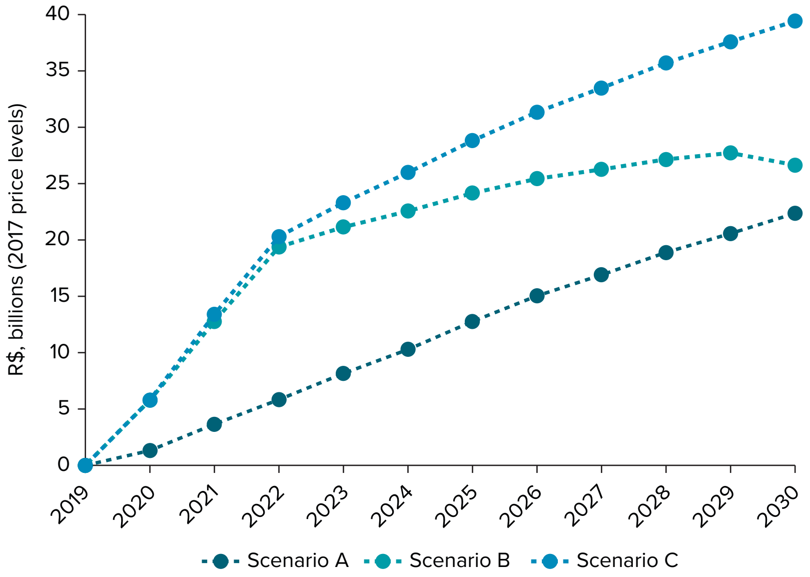
To put these different scenarios in perspective and compare their effectiveness in providing fiscal savings, we show in figure 10.16 the fiscal savings accumulated throughout the years in each reform scenario. In 2018, wage bill expenditures in the Brazilian federal civil service amounted to a total of R$131 billion. The projections of the model used in this analysis indicate that in 2026, scenario A saves approximately 12 percent of the 2018 wage bill expenditures, scenario B saves 19 percent, and scenario C saves 24 percent. Besides these differences in total savings, in scenarios B and C, the government achieves larger savings in the short term while compensating with smaller savings after a few years, whereas, in scenario A, the total savings are spread out over the years.

FIGURE 10.15 Wage Bill Projection and Policy Scenarios, Brazil, 2008–30



Source: Original figure for this publication.

FIGURE 10.16 Cumulative Fiscal Savings from Policy Scenarios, Brazil, 2019–30



Source: Original figure for this publication based on Brazilian government data from 2019.

Experimenting with combinations of policies before implementation to understand their fiscal impact has the potential to save a significant proportion of a government's wage bill. Similarly, such extrapolations can be extended to the descriptive analysis outlined in previous sections so governments can better understand how personnel policy reform will impact the character of the public service. With enough good-quality data, governments can leverage their SHRM and wage bill data for evidence-based planning of their fiscal expenditures and personnel dynamics into the future.

CONCLUSION

We have presented a microdata-based approach for governments to improve their SHRM and develop realistic civil service compensation and employment strategies. We have also demonstrated how such strategies can allow policy makers to make better fiscal choices. We have used a series of examples from LAC countries to demonstrate how the use of relatively basic payroll and HRMIS statistics can help policy makers gain insight into the current and future state of their government's wage bill. We posit that this constitutes an important first step toward tapping the potential of existing bodies of payroll and HRMIS microdata that are currently underused. We believe that our approach can help policy makers make difficult decisions by breaking down the causes of problems and putting numbers to the ways in which certain policy choices will translate into longer-term consequences. On the basis of our experience using HR microdata for such analyses, we have a series of practical recommendations to make.

The first recommendation pertains to the collection of the data required to undertake the analyses we propose. Although, in theory, any government with an HRMIS should have access to these data, we know from our experience working with governments that extracting and cleaning these data can be a difficult task. As such, we recommend that where possible, governments centralize their HR data collection systems and render these data accessible to insights teams. If such data do not exist, even in a disparate fashion, we strongly advise governments to begin collecting, in a centralized manner, payroll and HRMIS microdata. If governments are able to break down existing inter- and intradepartmental data silos and embed data analytics into their institutional culture, they stand to gain a much clearer idea of—among many other phenomena—the composition of their workforce, how to use their workforce more effectively, and how to plan, budget, and staff for future challenges. This is a central recommendation from our experience working with these microdata. As we laid out above, the quality and coverage of the data at one's disposal affect the usefulness of the analyses one can undertake and, consequently, the power of the insights one can gain.

The second recommendation is that the analysis of HR data be better integrated with fiscal planning. Our approach can both complement and help to bridge functional reviews and macroanalyses and, for this reason, can reconcile the fiscally oriented nature of macroanalyses with the detail of functional reviews. For this to be effective, however, governments must encourage civil servants from the treasury and HR department(s) to collaborate more closely. This could be achieved by allocating dedicated portions of civil servant workload (from both the treasury and the HR department) to the task of sharing and analyzing data in collaboration, or by creating dedicated interdepartmental roles to push forward and undertake the collection and analysis of HR microdata for SHRM. By better integrating HR data and wage bill planning, policy makers can also improve the services that are delivered to citizens. In the example we mentioned in the introduction, policy makers in Alagoas incorporated demographic changes into their projections of how many teachers to hire (given the falling pupil-per-teacher ratio caused by lower fertility rates) and were thereby able to identify an area in which they could achieve substantial savings and better target their HR strategy to hire different categories of civil servants that were not oversupplied. In this way, the state was able to provide better-quality services to its citizens by hiring civil servants in areas where greater personnel were needed, rather than in the education sector, where there was an excess of teachers.

The third recommendation relates to how political considerations can impede the implementation of successful SHRM and fiscal planning. We recommend that governments, in addition to centralizing HR data collection systems, seek to insulate certain aspects of planning offices' work from the ebb and flow of politics. This could go hand-in-hand with our second recommendation, to carve out explicit portfolios or roles dedicated to collecting and analyzing HR microdata, by ensuring that this work is undertaken by public servants reporting to an independent agency rather than to a minister.

All three recommendations pertain to how governments can better institutionalize SHRM and improve their analytical capabilities with data that should be relatively easy to collect and use. By developing a culture of centralizing and sharing such data—always anonymized, stored, and shared with full respect for employees' privacy and rights—governments can improve their ability to identify and resolve issues pertaining to the workforce and fiscal planning alike, as we have laid out. Moreover, such analyses are simple to undertake, meaning that governments can leverage these data through existing staff with even minimal data literacy, without hiring a significant number of data specialists. We hope we have illustrated the benefits of combining HRMIS and payroll data to inform SHRM and fiscal planning and that we have inspired practitioners to exploit these data's potential for more and better evidence-based policy making.

NOTES

This chapter is based on technical support provided to several governments across Latin America. The team was led by Daniel Ortega Nieto. Our thanks go to Vivian Amorim, Paulo Antonacci, Francisco Lima Filho, Sara Brolhato de Oliveira, Alison Farias, and Raphael Bruce for their part in the work presented here. The findings, interpretations, and conclusions expressed in this chapter are entirely those of the authors.

1. The IMF, in fact, estimates that over 130 countries report comprehensive government finance statistics and that, on average, countries have about 25 years of data at their disposal (Gupta et al. 2016, 11).
2. Some governments also gather and record data on pensioners and survivors. Having this additional data can be useful, especially to improve the government's understanding of retirees' profiles and the overall fiscal impact of pensions. Given that this subject opens a whole set of new analyses, however, we do not comprehensively discuss the use of pension data in this chapter.

3. The studies that exist are limited analyses looking at very specific issues, often from the health care sector, with the notable exception of Colley and Price (2010), who examine the case of the Queensland public service.
4. Of the HRM professionals they surveyed, 47 percent reported engaging in little or no work-force planning for their municipalities, and only 11 percent reported that their municipalities had a centralized, formal workforce plan (Goodman, French, and Battaglio 2015, 148).
5. Wage compression is generally defined as the ratio between high earners and low earners in a specific organization. In this chapter, we define wage compression as the ratio between the 90th percentile and the 10th percentile of the wage distribution of the organization.
6. The salary structure in the public administration consists of multiple salary components, grouped into "basic" and "personal" components. Basic payments are determined based on the specific position (*plaza*), which represents the set of tasks, responsibilities, and working conditions associated with each civil servant, including *sueldos al grado* and *compensaciones al cargo*. All civil servants also receive personal payments, which are specific to each individual employee.
7. For example, a Brazilian federal government employee works for an average of 30 years before retiring.
8. See "Banco Mundial aponta urgˆencia de uma reforma administrativa," *Valor Econômico,* October 10, 2019, https://valor.globo.com/brasil/noticia/2019/10/10/banco-mundial-aponta-urgencia-de-uma-reforma-administrativa.ghtml.
9. The "representative month" should allow for the extrapolation of monthly wage bill expenditures and the number of civil servants for the whole year.

REFERENCES

Anderson, Martin W. 2004. "The Metrics of Workforce Planning." *Public Personnel Management* 33 (4): 363–78. https://doi.org/10.1177/009102600403300402.

Choudhury, Enamul H. 2007. "Workforce Planning in Small Local Governments." *Review of Public Personnel Administration* 27 (3): 264–80. https://doi.org/10.1177/0734371X06297464.

Colley, Linda, and Robin Price. 2010. "Where Have All the Workers Gone? Exploring Public Sector Workforce Planning." *Australian Journal of Public Administration* 69 (2): 202–13. https://doi.org/10.1111/j.1467-8500.2010.00676.x.

Davenport, Thomas. 2019. "Is HR the Most Analytics-Driven Function?" *Harvard Business Review*, April 18, 2019. https://hbr.org/2019/04/is-hr-the-most-analytics-driven-function.

Dychtwald, Ken, Tamara Erickson, and Bob Morison. 2004. "It's Time to Retire Retirement." *Public Policy and Aging Report* 14 (3): 1–28. https://doi.org/10.1093/ppar/14.3.1.

Frank, Howard A., and Yongfeng Zhao. 2009. "Determinants of Local Government Revenue Forecasting Practice: Empirical Evidence from Florida." *Journal of Public Budgeting, Accounting & Financial Management* 21 (1): 17–35. https://doi.org/10.1108/JPBAFM-21-01-2009-B002.

French, P. Edward, and David H. Folz. 2004. "Executive Behavior and Decision Making in Small U.S. Cities." *The American Review of Public Administration* 34 (1): 52–66. https://doi.org/10.1177/0275074003259186.

GAO (US General Accounting Office/Government Accountability Office). 2001a. *Federal Employee Retirements: Expected Increase over the Next 5 Years Illustrates Need for Workforce Planning.* Report to the Chairman, Subcommittee on Civil Service and Agency Organization, Committee on Government Reform, House of Representatives, GAO-01-509. Washington, DC: US General Accounting Office. https://www.gao.gov/assets/gao-01-509.pdf.

GAO (US General Accounting Office/Government Accountability Office). 2001b. *High-Risk Series: An Update.* GAO-01-263. Washington, DC: US General Accounting Office. https://www.gao.gov/assets/gao-01-263.pdf.

GAO (US General Accounting Office/Government Accountability Office). 2021. *High-Risk Series: Dedicated Leadership Needed to Address Limited Progress in Most High-Risk Areas*. GAO-21-119SP. Washington, DC: US Government Accountability Office. https://www.gao.gov/products/gao-21-119sp.

Goodman, Doug, P. Edward French, and R. Paul Battaglio Jr. 2015. "Determinants of Local Government Workforce Planning." *The American Review of Public Administration* 45 (2): 135–52. https://doi.org/10.1177/0275074013486179.

Gupta, Sanjeev, David Coady, Manal Fouad, Richard Hughes, Mercedes Garcia-Escribano, Teresa Curristine, Chadi Abdallah, et al. 2016. "Managing Government Compensation and Employment-Institutions, Policies, and Reform Challenges." IMF Policy Paper, April 8, 2016, International Monetary Fund, Washington, DC.

Harborne, Bernard, Paul M. Bisca, and William Dorotinsky, eds. 2017. *Securing Development: Public Finance and the Security Sector*. Washington, DC: World Bank. http://hdl.handle.net/10986/25138.

Hasnain, Zahid, Nick Manning, and Jan Henryk Pierskalla. 2014. "The Promise of Performance Pay? Reasons for Caution in Policy Prescriptions in the Core Civil Service." *World Bank Research Observer* 29 (2): 235–64. https://doi.org/10.1093/wbro/lku001.

Hasnain, Zahid, Daniel Oliver Rogger, Daniel John Walker, Kerenssa Mayo Kay, and Rong Shi. 2019. *Innovating Bureaucracy for a More Capable Government*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/249891549999073918/Innovating-Bureaucracy-for-a-More-Capable-Government.

Huerta Melchor, Oscar. 2013. "The Government Workforce of the Future: Innovation in Strategic Workforce Planning in OECD Countries." OECD Working Papers on Public Governance 21, OECD Publishing, Paris. https://doi.org/10.1787/5k487727gwvb-en.

Jacobson, Willow S. 2009. "Planning for Today and Tomorrow: Workforce Planning." In *Public Human Resource Management: Problems and Prospects*, edited by Steven W. Hays, Richard C. Kearney, and Jerrell D. Coggburn, 5th ed., 179–202. New York: Longman.

Jacobson, Willow S. 2010. "Preparing for Tomorrow: A Case Study of Workforce Planning in North Carolina Municipal Governments." *Public Personnel Management* 39 (4): 353–77. https://doi.org/10.1177/009102601003900404.

Kavanagh, Shayne C., and Daniel W. Williams. 2016. *Informed Decision-Making through Forecasting: A Practitioner's Guide to Government Revenue Analysis*. Chicago: Government Finance Officers Association.

Kiyonaga, Nancy B. 2004. "Today Is the Tomorrow You Worried about Yesterday: Meeting the Challenges of a Changing Workforce." *Public Personnel Management* 33 (4): 357–61. https://doi.org/10.1177/009102600403300401.

Kong, Dongsung. 2007. "Local Government Revenue Forecasting: The California County Experience." *Journal of Public Budgeting, Accounting & Financial Management* 19 (2): 178–99. https://doi.org/10.1108/JPBAFM-19-02-2007-B003.

Mukhtarova, Turkan, Faisal A. Baig, and Zahid Hasnain. 2021. "Five Facts on Gender Equity in the Public Sector." *Governance for Development* (blog). *World Bank Blogs*, September 27, 2021. https://blogs.worldbank.org/governance/five-facts-gender-equity-public-sector.

OECD (Organisation for Economic Co-operation and Development). 2007. *Ageing and the Public Service: Human Resource Challenges*. Paris: OECD Publishing. https://doi.org/10.1787/9789264029712-en.

OECD (Organisation for Economic Co-operation and Development). 2019. "Gender Equality in Public Sector Employment." In *Government at a Glance 2019*, 88–89. Paris: OECD Publishing. https://doi.org/10.1787/9735a9f2-en.

Pynes, Joan E. 2004. "The Implementation of Workforce and Succession Planning in the Public Sector." *Public Personnel Management* 33 (4): 389–404. https://doi.org/10.1177/009102600403300404.

Pynes, Joan E. 2009. "Strategic Human Resources Management." In *Public Human Resource Management: Problems and Prospects*, edited by Steven W. Hays, Richard C. Kearney, and Jerrell D. Coggburn, 5th ed., 95–106. New York: Longman.

Reitano, Vincent. 2019. "Government and Nonprofit Personnel Forecasting." In *The Palgrave Handbook of Government Budget Forecasting*, edited by Daniel Williams and Thad Calabrese, 361–76. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-18195-618.

Riccucci, Norma M., Gregg G. Van Ryzin, and Cecilia F. Lavena. 2014. "Representative Bureaucracy in Policing: Does It Increase Perceived Legitimacy?" *Journal of Public Administration Research and Theory* 24 (3): 537–51. https://doi.org/10.1093/jopart/muu006.

Rogers, James, and C. M. Naeve. 1989. "One Small Step towards Better Government." *Public Utilities Fortnightly* 3: 9–12.

Selden, Sally Coleman. 2009. *Human Capital: Tools and Strategies for the Public Sector*. Washington, DC: CQ Press.

Selden, Sally Coleman, and Willow Jacobson. 2007. "Government's Largest Investment: Human Resource Management in States, Counties, and Cities." In *In Pursuit of Performance: Management Systems in State and Local Government*, edited by Patricia W. Ingraham, 82–116. Baltimore, MD: The Johns Hopkins University Press.

Somani, Ravi. 2021. "The Returns to Higher Education and Public Employment." *World Development* 144: 105471. https://doi.org/10.1016/j.worlddev.2021.105471.

Theobald, Nick A., and Donald P. Haider-Markel. 2009. "Race, Bureaucracy, and Symbolic Representation: Interactions between Citizens and Police." *Journal of Public Administration Research and Theory* 19 (2): 409–26. https://doi.org/10.1093/jopart/mun006.

Van Ryzin, Gregg G., Norma M. Riccucci, and Huafang Li. 2017. "Representative Bureaucracy and Its Symbolic Effect on Citizens: A Conceptual Replication." *Public Management Review* 19 (9): 1365–79. https://doi.org/10.1080/14719037.2016.1195009.

Wilkerson, Brian. 2007. *Effective Succession Planning in the Public Sector.* Arlington, VA: Watson Wyatt Worldwide.

Wong, John D. 1995. "Local Government Revenue Forecasting: Using Regression and Econometric Revenue Forecasting in a Medium-Sized City." *Journal of Public Budgeting, Accounting & Financial Management* 7 (3): 315–35. https://doi.org/10.1108/JPBAFM-07-03-1995-B001.

World Bank. 2019. *Gestão de Pessoas e Folha de Pagamentos no Setor Público Brasileiro: O Que os Dados Dizem*. Washington, DC: World Bank.

World Bank. 2021. *Uruguay Public Expenditure Review: Civil Service Diagnosis*. Washington, DC: World Bank.

World Bank. 2022. *Subnational Civil Servant Pension Schemes in Brazil: Context, History, and Lessons of Reform.* Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/37267.

CHAPTER 11

Government Analytics Using Expenditure Data

Moritz Piatti-Fünfkirchen, James Brumby, and Ali Hashim

SUMMARY

Government expenditure data hold enormous potential to inform policy around the functioning of public administration. Their appropriate use for government analytics can help strengthen the accountability, effectiveness, efficiency, and quality of public spending. This chapter reviews where expenditure data come from, how they should be defined, and what attributes make for good-quality expenditure data. The chapter offers policy makers an approach to reviewing the adequacy and use of government expenditure data as a means to strengthen the effectiveness of government analytics that builds on these data. Case studies are used to illustrate how this can be done.

ANALYTICS IN PRACTICE

- Be clear about how *expenditure* is defined. *Expenditure* is a term that is often interpreted and used loosely. This can lead to confusion and misunderstandings in analytical assessments. The literature around public expenditure data is clear on a series of standard definitions and offers guidance as to their application. It is recommended to take advantage of this, where feasible, and to minimize the ambiguous use of terms, to the extent possible.
- Understand and document the origins of government expenditure data. Government expenditure data have enormous potential to inform the accountability, efficiency, impact, and equity of operations. It is important to understand and document how transactions across spending items in government are created, what control protocols they are subject to, how this information is stored, and how microdata are aggregated for analysis.
- Do not take data at face value. The usefulness of analysis from government expenditure data hinges upon the quality of the underlying microdata. It is recommended that the origins of government expenditure

Moritz Piatti-Fünfkirchen is a senior economist and James Brumby is a senior adviser at the World Bank. Ali Hashim is an independent consultant.

microdata be periodically reviewed for data provenance and integrity, comprehensiveness, usefulness, consistency, and stability. It is recommended that such work be publicly disclosed, to the extent possible. This can be used as a baseline upon which a reform program can be built to address deficiencies.

- Take a microdata-driven approach to expenditure analysis. An analysis of microlevel expenditure data can offer data-driven and objective insights into expenditure management practices and the impacts of expenditure policy. From this analysis, a government expenditure profile can be derived, which shows where large transactions with high fiduciary risks are taking place, how these compare to low-value transactions at points of service delivery, and where expenditure policy intentions are not being converted into the desired impact. Such analysis can offer operational insights for better expenditure management that serves expenditure control and service delivery objectives.

INTRODUCTION

Public resources are scarce. Increasingly, competing demands on the public purse make prudent and evidence-based expenditure decisions ever more important. This requires, among other things, accurate and timely government expenditure data. Government expenditure data are central to the social contract between society and elected officials. They provide an important basis for accountability, insights into whether resources are being used for budgeted priorities, and assessments of whether spending is sustainable and equitable.

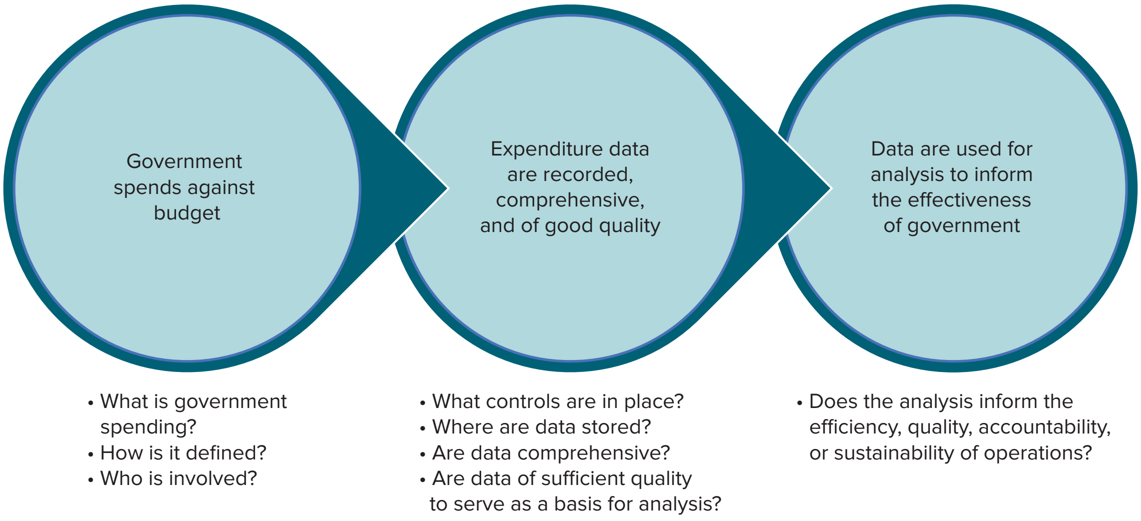
Expenditure data are central to assessing the fiscal health of a country (Burnside 2004, 2005) and necessary for debt sustainability analyses (Baldacci and Fletcher 2004; Di Bella 2008). These are core elements of government operations and often accompany World Bank Public Expenditure Reviews (PERs) or International Monetary Fund (IMF) Article IV reports. A government's commitment to deficit targets can, for example, be measured by identifying whether large expenditure items in the budget are subject to necessary internal controls (Piatti, Hashim, and Wescott 2017).[1]

Expenditure data can be used to assess whether spending is effective and efficient. They thus provide information about the functioning of public administration. Such assessments can be made at the very granular level of a department, unit, or even project. For example, budget data might indicate that a construction project is not disbursing as quickly as projected, limiting its progress. This is indicative of potential problems in expenditure management. Such analysis of government expenditure data is used by the executive branch, audit institutions, the legislative branch, and civil society to offer insights into the quality of the administration.

To get to the stage where expenditure data can effectively inform the functioning of government, a simplified, three-step logframe indicates two preparatory stages (see the figure 11.1; a detailed exposition of the stages in the creation of expenditure data is provided in appendix C). First, there needs to be clarity about what government spending means, who it involves, and what it covers. Second, there is the question of how spending is executed (for example, what controls it is subject to), where data are stored, how comprehensive they are, and what the quality of the data is. Third, high-quality expenditure data with high coverage can lend themselves to analyses that inform the effectiveness of government.

This logframe illustrates that the value of any analysis is a function of the quality of the underlying expenditure data. It is therefore important for practitioners to reflect carefully on how government expenditures are defined and where they come from and to critically assess the quality of government expenditure microdata (or transactions data). While there is a lot of guidance on how to analyze government expenditure data (step 3 in figure 11.1), the literature is relatively silent on how to assess the quality of expenditure data, as well as on how poor data may affect the validity of the conclusions drawn from such analyses (step 2). Despite clear guidance on definitions and coverage (step 1), the term *government expenditure* continues to be used to imply a multitude of different concepts that are frequently not well communicated, leading to confusion among analysts.

FIGURE 11.1 Use of Government Expenditure Data for Government Analytics


Government spends against budget
Expenditure data are recorded, comprehensive, and of good quality
Data are used for analysis to inform the effectiveness of government
• What is government spending?
• How is it defined?
• Who is involved?
• What controls are in place?
• Where are data stored?
• Are data comprehensive?
• Are data of sufficient quality to serve as a basis for analysis?
• Does the analysis inform the efficiency, quality, accountability, or sustainability of operations?

Source: Original figure for this publication.

This chapter walks through each of these steps as follows. It starts by discussing issues related to defining government expenditure data (step 1). It then reviews the attributes of good government expenditure data and makes observations about how data can be strengthened and made more reliable and useful for analysis. The chapter highlights the importance of data provenance and integrity, comprehensiveness, usefulness, consistency, and stability as critical attributes (step 2). Examples of how to pursue these characteristics are provided and illustrated through case studies. These case studies indicate that deficiencies in any of these characteristics constitute a risk to the ability of analysts to use these data to inform an understanding of government functioning (step 3).

WHAT ARE GOVERNMENT EXPENDITURE DATA?

Despite the centrality of government expenditure, definitional issues remain. The term *government expenditure* is often used with liberty among practitioners and analysts. For example, *budget*, *commitment*, and *expenditure data* are sometimes used interchangeably. Further, there is often insufficient differentiation between cash and accrual concepts. Suffice it to say, it is important to be clear and precise when using the term *expenditure* to allow for an effective dialogue and comparability over time and across countries.

Expenditure is defined by the Organisation for Economic Co-operation and Development (OECD) as "the cost of goods and services acquired, regardless of the timing of related payments." Expenditures are, therefore, different from cash payments. Instead, "expenditures on goods and services occur at the times when buyers incur liabilities to sellers, i.e. when either (a) the ownership of the goods and services concerned is transferred from the seller to the new owner; or (b) when delivery of the goods and services is completed to the satisfaction of the consumer." Conversely, the term *expense* "defines the set of transaction flows that reduce net worth over the accounting period" (Allen and Tommasi 2001, 452). This distinction reveals that while an *expenditure* may result in the acquisition of a capital item, an *expense* will apply to the use (depreciation) or care (maintenance) of the item.

Governments spend money as a result of a series of economic relationships. The main ones are as follows:

- To pay wages, salaries, and other emoluments for labor
- To purchase goods and services that are then used in the production of government outputs
- To purchase assets
- To transfer resources (unrequited) to other levels of government, households, or firms
- To meet the cost of servicing debts
- For various other purposes, such as meeting legal claims.

Expenses can be incurred for events that do not involve a same-time transaction—for instance, changes in the estimate of unfunded liabilities associated with government pensions or the impairment of an asset through its use (depreciation). The distinction considers an *expenditure* to acquire goods, with the *expense* occurring when the goods are used.

All expenditure transactions that are routed through a *financial management information system* (FMIS) are reflected in the government's accounts, or *general ledger*, without exception, providing a comprehensive data source for analysis. Each transaction originates from a spending unit within the government, ensuring that each transaction can be mapped to a particular office. Because these transactions must be executed against the index of allowed payments agreed upon in the budget, or *chart of accounts* (COA), and must specify the amount, the details of the payee (including the recipient's account number and the time of the transaction) are a natural component of expenditure data. Depending on the level of detail of the COA, the transaction may capture the source of funds, the organizational code, the purpose of the expenditure (economic classification or line item), the jurisdiction in which the transaction happened, and the program or subprogram it related to. The format that the data structure of financial transactions in an FMIS typically takes is given in table 11.1.

Transactions may also be processed manually, outside the FMIS, and then posted manually to the general ledger. These transactions are thus not automatically subject to the same set of FMIS internal controls, and the same level of transaction detail may not be available. Furthermore, these transactions may be aggregated and posted in bulk, making the desired analysis of microdata difficult.

ATTRIBUTES OF GOOD-QUALITY EXPENDITURE DATA

Understanding definitional nuances and assessing the quality and credibility of the underlying microdata both benefit from an understanding of the government information system's architecture. There are multiple functions, processes, agencies, and associated systems at play. These include processes and systems for

TABLE 11.1 Example of Expenditure Data, by Transactions

Transaction ID	Time stamp (date)	Chart of accounts segment					Amount	Payee (and account number)
		Source of funds	Organization code	Purpose code (line item)	Location code	Program/ subprogram code		
Transaction 1								
Transaction 2								
...								
Transaction *n*								

Source: Hashim et al. 2019.

macroeconomic forecasting; budget preparation systems; treasury systems; establishment control, payroll, and pension systems; tax and customs systems; debt management systems; and auditing systems. Together, these systems represent the information architecture for government fiscal management, underpinning government expenditure management, and are the basis for government expenditure data. A detailed account of these systems is provided by Allen and Tommasi (2001), Hashim (2014), and Schiavo-Campo (2017). Carefully designed, functional processes supported by adequate systems, and the good utilization of those systems, will yield good-quality government expenditure data that can be analyzed to inform policy. Weaknesses in any one of these processes, by contrast, will undermine the quality of expenditure data.

Spending, and the production of expenditure data, follows a process. Once the budget is authorized and apportioned to spending units, commitments can be made. The receipt of goods or services then needs to be verified before a payment order can be initiated. Bills are paid upon the receipt of the payment order. This is then accounted for against the full COA and provides the basis for expenditure data (figure 11.2). A full account of these processes, including differentiation by colonial history, is offered by Potter and Diamond (1999) and Shah (2007). Further details are provided in appendix C.

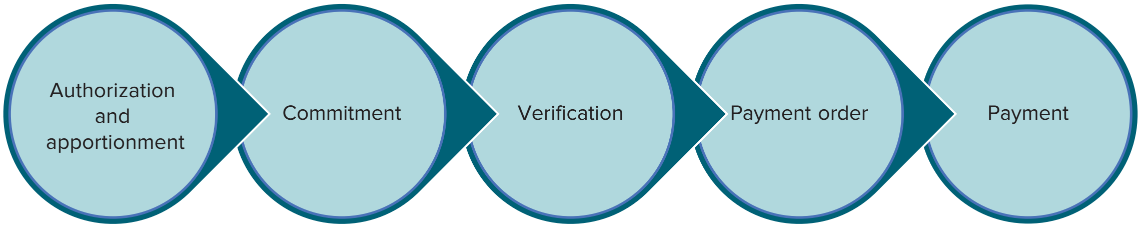
There are numerous agencies and processes involved in the production of government expenditure data. The quality and credibility of these data depend on how well the data production process is implemented across these agencies and processes. This chapter identifies five principles in the data production process that can help assess the adequacy of the data for further analysis. Each adds to the likelihood that expenditure is reliable. Unlike personnel data, discussed in chapter 10 of this *Handbook*, it is more difficult to present a "ladder" of quality for expenditure data. These five principles interact to determine the utility of the resulting data. For example, an incomplete data set that focuses on the 50 percent largest expenditure items is likely to cover a substantial portion of total expenditure. These principles should be seen as underpinning a high-quality expenditure-data-generating system, and what will yield the greatest improvement in overall quality will be specific to contexts and almost to data points.

Data Provenance and Integrity

Expenditure data are useful for analysis if there is confidence in their integrity. *Data provenance*—the documentation of where data come from and the processes by which they were produced—is necessary to have this confidence. There should be a clear sense of what systems data have come from, who was involved in the production of the data, and where the data are stored. Internal controls for systems should ensure data provenance and integrity. If systems are used, controls are applied, and data are immutable (or there is a clear trail in any changes), there can be confidence in data integrity. The use of an FMIS, for example, should guarantee data provenance and integrity—if transactions were executed through the system and, therefore, were subject to FMIS internal controls.

If expenditures are not routed through the dedicated government system, data provenance and integrity are more difficult to guarantee (Chami, Espinoza, and Montiel 2021; Milante and Woolcock 2021).

FIGURE 11.2 Stages in the Execution Process That Create Government Expenditure Data



Sources: Adapted from Potter and Diamond 1999; Shah 2007.

As evidenced in the literature, in many lower- and middle-income countries, such as Ghana, Pakistan, and Zambia, FMIS coverage remains limited (European Commission and IEG 2017; Hashim, Farooq, and Piatti-Fünfkirchen 2020; Hashim and Piatti-Fünfkirchen 2018; Piatti-Fünfkrichen 2016). In some instances, it may be for good reason that systems are not used. There may, for example, be information and communication technology limitations, a lack of access to banking services, or human capacity constraints in remote areas. In other instances, not using systems may be a purposeful choice to avoid said controls and, thus, clear data provenance. In either case, transactions posted manually to the general ledger are more susceptible to manipulation. In these cases, confidence that the reported expenditure reflects actual spending is likely to require a costly ex post audit. An example of how FMIS utilization helped resolve a major corruption episode in Malawi is illustrated in box 11.1.

Mixing good-quality data with questionable data calls into question the credibility of the entire data set because the provenance is not accurately tracked for each and every transaction. It is therefore important to understand which transactions were processed by the FMIS and where transactions that were *not* processed by the FMIS come from, as well as whether their integrity can be assured (Hashim and Piatti-Fünfkirchen 2018).

Comprehensiveness

Comprehensiveness is defined with respect to the reporting entity. If the desire is to review expenditure performance across the entire government sector, then this requires data to be comprehensive across levels of government, sources, and, preferably, time. Data comprehensiveness is complicated by mismatches between the institutional setup of a government (consolidated fund or public account) and the definition of *government*, which may include bodies that are outside the consolidated fund or a jurisdictional structure that clearly separates the levels of government, as is the case with federations. What is important for the

BOX 11.1 How Utilization of a Financial Management Information System in Malawi Supported Data Provenance and Helped Resolve a Major Corruption Episode

Adequate utilization of the financial management information system (FMIS) in Malawi helped ensure that most spending was transacted through the system and that expenditure data were recorded and stored on the general ledger. During a major corruption episode, data provenance—ensured through the FMIS—enabled the tracing of the transactions and events (Baker Tilly Business Services Limited 2014; Bridges and Woolcock 2017; Hashim and Piatti-Fünfkirchen 2018; World Bank 2016a).[a] This, consequently, allowed authorities to follow up, identify collusion, and prosecute. In an environment where transactions are posted manually, it is easier to tamper with records, which undermines the integrity of the data and, thereby, the ability of authorities to ensure accountability. The increasing penetration of banking innovations, such as mobile money or smart cards, offers governments the ability to make electronic transfers or payments even in remote areas (where access to conventional banking services is unavailable), which leave a digital footprint. Even if the FMIS does not execute the transaction, posting these onto the ledger would strengthen data provenance, transparency, and accountability (Piatti-Fünfkirchen, Hashim, and Farooq 2019). This practice has been widely applied for cash transfers in some countries, such as Kenya, Rwanda, Uganda, and Zambia.

a. There was a misconception that the FMIS was at fault for not preventing the misappropriation of funds. Collusion among main stakeholders was a human, not a system, error. The FMIS should be credited with ensuring data provenance and supporting prosecution in due course (World Bank 2016a).

integrity of the analysis is that the comprehensiveness of the reporting entity can be established. If transaction coverage for a reporting entity is not comprehensive, the findings will reflect this and may fall short of their intended purpose.

Guidance on how the *public sector* is defined is available in the IMF's *Government Finance Statistics Manual* (IMF 2014) and in the *Handbook of International Public Sector Accounting Pronouncements* (IFAC 2022). However, the application of this guidance can vary across countries and institutions, making meaningful cross-country comparisons for reporting purposes difficult (Barton 2011; Challen and Jeffery 2003; Chan 2006). In some cases, the public sector may be narrowly defined, with asymmetrical representation of the general government and public corporations. Reporting on the public sector may be partial—even at the aggregate level—in sensitive sectors, such as defense, the police force, or space programs, or it may be partial due to the funding source. The budget sector (and within it, the various forms of annual and standing appropriations), the public account, the general government sector, and the broader public sector may all justifiably be the entity of interest for some analyses; what is important is to understand *why* a given entity is the entity in question and what activity it excludes relative to a more relevant entity. For example, when seeking to communicate restraint, a central government may highlight its total spending, including transfers to lower levels of government, whereas a subnational government may wish to distinguish the central government's spending for its own purposes and programs from the funds it transfers to lower levels of government. This distinction may reveal that a large portion of the central government's "restraint" comes from cuts to others' programs rather than restraint in the delivery of the central government's own work.

The comprehensiveness of spending can suffer from a lack of transparency on debt. As countries are increasingly indebted, this factor becomes increasingly important. Drawing from new databases and surveys, Rivetti (2021) finds that nearly 40 percent of low-income developing countries (LIDCs) have never published debt data on their websites or have not updated their data in the past two years. When debt data are available, they tend to be limited to central government loans and securities, excluding other public sector components and debt instruments. For some LIDCs, debt data disclosed across various sources show variations equivalent to as much as 30 percent of a country's gross domestic product—often because of differing definitions and standards and recording errors. Data in the debt management system should comprehensively reflect all loans and liabilities and actual debt servicing requirements. Actual spending should be comprehensively reflected in the FMIS. Even here, it is important that expenditure controls apply in order to avoid expensive short-term borrowing that has not been budgeted for (Hashim and Piatti-Fünfkirchen 2018).

Comprehensiveness is equally important for sector expenditure analysis. Health spending, for example, is frequently benchmarked against the Abuja Declaration target of 15 percent of the government budget (African Union 2001). However, how well one can proxy this indicator depends on a country's ability to credibly populate the numerator (health spending) and the denominator (general government spending), and the literature has shown this to be difficult (Piatti-Fünfkirchen, Lindelow, and Yoo 2018). Estimating health spending typically goes beyond just one reporting entity. Thus, reporting comprehensively on all health spending, including relevant off-budgetary funds, the use of internally generated funds (for example, user fees), development partners (Piatti-Funfkirchen, Hashim, et al. 2021), and the use of tax expenditures (Lowry 2016) becomes important in a consideration of the resources dedicated to the sector.[2] Estimating the comprehensiveness of the denominator, then, is complicated by all the factors outlined above.

Comprehensiveness also requires comprehensive reporting over time. A timing mismatch between receiving a good or service and the payment of cash can lead to the creation of *payment arrears*—a liability that is past due. Accurate reporting on such arrears is important for comprehensiveness. The 2020 Public Expenditure and Financial Accountability (PEFA) global report notes that countries' stock of expenditure arrears was around 10 percent of total expenditure, well above the 2 percent considered good practice (PEFA 2020).[3] If these are not adequately reported, any expenditure analysis will be inaccurate. The PEFA indicator for expenditure arrears (PI-22) is, however, one of the poorest-rated indicators in the framework (PEFA 2022). This is despite the fact that adequate expenditure controls tend to be in place, suggesting that these are frequently bypassed, leading to the aforementioned data provenance and integrity concerns.

Finally, many aspects of government expenditure are driven by trends that extend beyond the annual time cycle that generally applies to budgets. For example, changing demographics mean that societal needs for services such as education and health care change over time. Similarly, differences in timing between the creation of an obligation (such as a pension) and the payment of that obligation mean that it is important to consider the multiannual nature of spending to get a more complete picture. Spending (or not spending) today may create important obligations over time. Consumption- and investment-related spending are fundamentally different and need to be recognized as such. Yet annual expenditure reporting requirements tend to take a short-term perspective regardless of the nature of spending. Further, what is captured as expenditure may be influenced by what is not captured, which may nevertheless impact what is left to be performed by functions requiring expenditure—for example, regulation and its associated compliance and tax expenditures. If wider resource use is a concern, rather than narrow expenditure, then the analytical net should also be cast much wider (see, for example, the methods in Stokey and Zeckhauser [1978]).

Usefulness

In order to analyze and interpret findings in a way that meaningfully informs government administration, government budget data also need to be structured in a meaningful way. Budget and expenditure data are generally presented by administrative, economic, programmatic, and functional segments (see table 11.1). The purpose of the administrative segment is clear: it allows the government to allocate, monitor, and hold to account spending within its administrative structures. The purpose of the economic classification is also clear. It classifies the expenditure according to what inputs it has been spent on, which is necessary for accountability. Countries with a program structure require program classification in the COA because appropriations happen accordingly. Functional classification is appealing because it allows decision-makers to readily identify how much has been allocated and spent according to specific functions, such as primary education and basic health care. If expenditure items can be clearly mapped to functions, this type of classification offers substantial analytical possibilities. An example of a classification of the functions of government (COFOG) pertaining to the health sector is offered in table 11.2. Together, these set of classifiers should let analysts cross-tabulate expenditure data in many meaningful ways.

Business intelligence strategies and technologies can then be used for the analysis of the information stored in the data warehouse. Appropriate tagging and data structure allow for automated reporting and analytical processing following business needs. Dashboards can be developed to provide information to management in government agencies on issues such as budget execution, cash position, and audit, allowing for real-time, evidence-based decision-making (Negash and Gray 2008).

TABLE 11.2 Example of Classification of Functions of Government from the Health Sector

First level	Second level
Health	Medical products, appliances, and equipment
	Outpatient services
	Hospital services
	Public health services
	R&D health
	Health n.e.c.

Source: Eurostat 2019, 37.
Note: n.e.c. = not elsewhere classified; R&D = research and development.

However, classifying these functions may not be trivial. With reference to the health sector, the following issues may arise:

- **Classifying by some functions may not always be possible.** In the health sector, a hospital generally offers both inpatient and outpatient services. Unless it has dedicated departments drawing on distinct cost centers, it may not be possible to differentiate between these services. It may be possible to understand total hospital spending but not necessarily the functions to which spending was dedicated within the hospital. Furthermore, health staff may provide both inpatient and outpatient services, and it would be difficult to apportion wages without a robust time-recording system. Similarly, in countries where the region- or district-level administration is also the lowest spending unit, it can be difficult to apportion specific functions because district authorities (in health) generally need to offer primary and secondary care as well as public health services. Therefore, if the spending unit does not have a clear mandate that maps directly to the COFOG functions, it is necessary to make assumptions, and these may not always be helpful or appropriate. In case there is no clear fit, it may be more accurate to simply report by administrative segment than to fit a square peg into a round hole. A COA reform process can be pursued over time to make spending more meaningful from a functional perspective.
- **Reporting by function requires a discrete choice.** Spending can be classified as *either* health or education spending—but not *both*. However, there are teaching hospitals that could be classified as either. There may also be medical facilities managed by the defense sector where allocation could be contested. Further, it is unclear whether subsidies for enrolling the poor in health insurance should be considered a health or a social protection function.
- **Not all functions, per COFOG categories, can be clearly understood as government functions.** For example, in the health sector, the COFOG category of *medical products, appliances, and equipment* may more appropriately be classified as *inputs* in the economic classification rather than *functions*. This also raises the question of inconsistencies within classifications because these medical products also serve other functions in the COFOG, such as hospital care or outpatient services.
- **There may be an overlap between functional and program classifications** because programs are output oriented and should serve specific government functions. There can still be added value for having both, but this needs to be clarified.

Reporting by functional classification is useful as long as it can be done credibly. Whether and how this exercise is done should reflect local context, demand, and capacity. Recommendations to shift spending toward some functions will remain unhelpful if these cannot clearly be traced back to the government's administrative structures. For example, it may be appealing to recommend more spending on outpatient services (which tend to be more efficient than hospital services), but as long as the government cannot clearly differentiate between spending on inpatient and outpatient services at the hospital level, such recommendations will remain unhelpful. Furthermore, as long as functional classification remains subjective, based on the assumptions of the analyst, any recommendations to adjust spending will lack credibility. This problem was recognized by the *Rwanda Nutrition Expenditure and Institutional Review 2020*, which cautions that extensive allocative efficiency analysis will remain futile as long as it cannot be clearly mapped back to the budget (Piatti-Fünfkirchen et al. 2020). Instead, a COA reform process may be more meaningful to improve the functional classification toward what is needed for the easier interpretation of expenditure data (see box 11.2).

Data presented in a useful format with integrity will likely foster demand for analysis. To make data more useful to analysts, there are ongoing initiatives by development partners that systematically clean, process, and categorize FMIS data (see, for example, the World Bank's BOOST initiative).[4] There has been a lot of demand for these initiatives because they support the various other analytical products that require government expenditure data in an interpretable format. However, as long as this work is not produced domestically through domestic systems, it is unlikely to be sustainable and will not undergo the required domestic checks and balances. This is an essential task of government data management—data storage in an adequate format in the business warehouse, from which a business intelligence system can pull meaningful reports—possibly requiring investments in institutional, systems, and human capacity.

BOX 11.2 How the Government of Rwanda Uses Budget Tagging to Implement a High-Priority, Cross-Cutting Agenda

The budget of the government of Rwanda, like in many countries, is organized vertically by ministry, department, and agency. This lends itself well to oversight and accountability. For some issues, such as climate change, gender, or nutrition, where implementation cuts across sectors and agencies, it can be difficult to identify what relevant activities were budgeted for and implemented. The government therefore introduced an upstream tagging process, in which ministries identify what interventions in the budget are related to these issues. This has provided a crucial management function because the financial management information system can now produce functional budget execution reports that reflect spending on these issues. At any point in time, it provides insight into what activities have been implemented, which activities remain to be implemented, and what the remaining cash flow requirements are. It thereby uses the budget as a tool for oversight, coordination, and mutual accountability for implementing a high-priority, cross-cutting agenda.

Consistency

Consistency in data management enables the production of data that can interface across systems and over space and time to allow for meaningful analysis. The COA's definition and use in government systems are influenced by different public financial management (PFM) traditions. PFM traditions can leave countries with the application of different COAs across levels of decentralization (Cooper and Pattanayak 2011). As long as this is the case, it is difficult to have a unified data set that allows for the analysis of expenditure information across the country, which complicates management and decision-making (PEMPAL 2014). One example of such a case is Indonesia, where a long-standing reform process has aimed to unify the COA across the country.

Consistency is also required across the system landscape in a country. This means that the same COA should be used throughout the FMIS and that taxes, debt management, and payroll should be classified according to the codes in the COA. Without unified use of the COA, adequate integration across systems to conduct government analytics will not be possible. For example, understanding the full fiscal health of an organization requires an integrated data set on the budget, debt, and payroll. If development partners are an important source of revenue, they should be encouraged to use the same classification structure so that comprehensive expenditure reports can be produced (Piatti-Funfkirchen, Hashim, et al. 2021).

It is equally important that the COA is used as intended. If activities are posted as line items, or vice versa, this creates problems for the quality of expenditure data and, subsequently, for analysis (Farooq and Schaeffer 2017). Similarly, it is important not to confuse programs with projects. A *program* is a set of activities that contribute to the same set of specific objectives, an *activity* is a subdivision of a program into homogenous categories, and a *project* is a single, indivisible activity with a fixed time schedule and a dedicated budget or activities. In some instances, development partners are responsible for the introduction of line-item codes into the COA in order to give a certain engagement more visibility or allow for the allocation of resources to one specific engagement area. This can come at the cost of coherence and consistency. For example, in Zimbabwe's health sector, there is a line item called *results-based financing*, one called *hygiene and sanitation*, and one called *malaria control*. All of these are important engagement areas but not inputs. They should be reflected as such in the COA. Similarly, in Rwanda, there is a line item called *maternal and child health*, which is also not a reflection of inputs but rather a target group.

Finally, it is important to be clear about nomenclature. Mixing budget data, release data, commitment data, and actual expenditure data within the same data set will lead to inconsistencies and problems.

Stability

The comparability of data over time is assisted by having a stable process to produce them and a stable classification system to parse and present them. But perfect stability does not occur: some degree of variation is natural and to be expected as conditions change, knowledge advances, and governments address evolving needs. Changes in reporting may be consequential, through the introduction of new spending agencies or the shift from input budgets to program structures. Stability does not require a static reporting structure, which would be unrealistic and unhelpful. It does, however, require the government to be able to connect current expenditure information to the past to be able to make use of trend data. This can be done by designing a coding scheme that can accommodate older and newer codes; by taking a sequenced, incremental approach to reforms; or by at least maintaining tables that form a bridge between data series to allow for reasonable consistency between the past, the present, and the future.

If such mitigation measures are not taken, change can be disruptive. For example, in Zimbabwe, the program structure in the health sector was substantially revised at both the program and the subprogram levels to accommodate an additional focus on research by the Ministry of Health and Child Care. A program and four subprograms were added, and four subprograms were removed. This meant that 35 percent of the approved 2020 budget had been allocated to programs that no longer existed in the 2021 budget. Instability in the classification of the program structure over time without adequate mitigation measures (for example, bridge tables) raises the question of what kind of actual reallocations accompanied these shifts. The possibility of multiyear analysis for costing or value for money remains severely limited in such scenarios. Similarly, the changes mean that performance targeting may be disrupted, as it was in the Zimbabwe health case, in which none of the 17 program outcome indicators in the 2020 budget remained available in the 2021 budget (World Bank 2022).

EXPLORING MICROLEVEL GOVERNMENT EXPENDITURE DATA

Developing comprehensive, appropriately structured, consistent, and stable data with a clear provenance provides a foundation for effective analytics. Though a large literature on the analysis of expenditure data exists (see, for example, Robinson [2000], Tanzi and Schuknecht [2000], and some discussion in appendix C), there is less discussion of how these data might be used to understand the functioning of government itself.

There are many examples of how government expenditure data can be used to inform the efficiency of government spending and better understand how a government is functioning. Expenditure information is necessary for an administration to explore opportunities for reducing the cost of the resources used for an activity or for increasing the output for a given input while maintaining quality (McAfee and McMillan 1989). The National Audit Office in the United Kingdom assesses how well the administration makes use of resources to achieve intended outcomes (NAO 2020). In Kenya, "data envelope analysis" is used to compare the efficient utilization of resources across counties (Kirigia, Emrouznejad, and Sambo 2002; Moses et al. 2021).

Information on differences in the amounts paid for goods between the public and private sectors is also frequently used to measure inefficiencies and can point to deep-rooted problems in the quality of an administration (see chapter 12). In Zambia, for example, such an analysis found that the rapid accumulation of payment arrears led to suppliers' building in a risk premium and, consequently, to the government's paying higher prices and suffering an unnecessary efficiency loss (World Bank 2016b). Generally, efficiency analyses are a central component of many analytical products of governments and development partners, such as Public Expenditure Reviews (PERs), and guidance on how to conduct these is widely available (Coelli et al. 2005; Greene 2008; Pradhan 1996; Shah 2005).

Government expenditure data can also be used to inform allocative choices, determining which groups, geographic regions, or sectors receive the most resources. Equity analysis allows for reorienting spending to better follow needs if resources are not flowing to the areas identified as requiring the most resources. In the health sector, benefit and expenditure incidence analyses are commonplace (Binyaruka et al. 2021; Mills et al. 2012; Mtei et al. 2012; Wagstaff 2012) and often accompany PERs. They provide insight into who pays for services and, separately, who utilizes services. They can thus offer concrete recommendations about how to restructure spending to be more equitable. More broadly, Commitment to Equity Assessments offer a methodology to estimate the impact of fiscal policy on inequality and poverty (Lustig 2011, 2018).

Government expenditure data are used as a foundation for accountability (Ball, Grubnic, and Birchall 2014; Griffin et al. 2010; Morozumi and Veiga 2016). If government expenditure data can be made publicly accessible for analytical purposes, this extends the benefits further. Groups across society can use published data to undertake their own assessments of government functioning and the distribution of public resources. A growing body of research tests the notion that transparency facilitates accountability and leads to a host of developmental outcomes. Using the frequency of the publication of economic indicators, including those related to government expenditure, Islam (2003) finds that countries with better information flows have better-quality governance. Hameed (2005) analyzes indexes of fiscal transparency based on IMF fiscal Reports on the Observance of Standards and Codes (ROSCs) and shows, after controlling for other socioeconomic variables, that more-transparent countries tend to have better credit ratings, better fiscal discipline, and less corruption. Similarly, an analysis of the Open Budget index shows that more-transparent countries tend to have higher credit ratings (Hameed 2011). Looking at each of the six PFM pillars covered by the current PEFA framework, de Renzio and Cho (2020) find that the "transparency of public finances" and "accounting and reporting" have the most direct effect on budget credibility. The authors stipulate that this may be because more information and timely reporting allow for more direct, real-time control of how public resources are being used.[5]

To provide practical details of this type of data analysis, this chapter now focuses on some of the most basic but useful analyses of expenditure data that can assist in understanding the functioning of government administration, with particular reference to a case study in Indonesia.

Basic Descriptives from Government Expenditure Microdata

First, to gain a sense of the completeness of the data being used for analysis, analysts may wish to estimate the budget coverage, which requires the summation of the value of all expenditure transactions routed through the data source (usually an FMIS) in a given fiscal year and, subsequently, the division of this value by the total approved budget reported by the government. This is presented in equation 11.1, where *t* represents the fiscal year and *i* the individual transaction:

$$\frac{\sum_{t,i}(trans_{1,1} + trans_{1,2} + trans_{2,2} + ... + trans_{t,i})}{Total\ approved\ budget}. \tag{11.1}$$

Equation 11.1, in turn, provides inputs to a table of the form of table 11.3.

The FMIS budget coverage statistics can be calculated for the general government, subagencies, and provinces or other subnational levels of government separately. These calculations then give an idea of the

TABLE 11.3 Sample Output of FMIS Budget Coverage Estimation

	2019	2020	2021
Total approved budget			
Total volume processed through the FMIS			
Percentage processed through the FMIS			

Source: Original table for this publication.
Note: FMIS = financial management information system.

agencywide and geographic spread in the coverage of the FMIS, allowing analysts to assess what percentage of the approved budget is processed by the FMIS.

Second, budget expenditure data can be used to identify trends and patterns in *budget execution rates*: the proportion of intended expenditures that have been undertaken within a specific time period. Budget execution data are a basic but important representation of how an organization is using resources and, when coupled with other information, how well it is working. If it is spending well but producing no outputs or not spending despite important upcoming commitments, these are signals of problems within the administration. Execution analysis also serves as a foundation for accountability because it can shed light on whether funds have been used for their intended purpose.

The analysis of budget execution rates can be conducted for the government as a whole or for specific sectors, spending units, line items, or programs. The type of analysis done will depend on how analysts want to assess the effectiveness of the administration. The aggregate budget execution rate alone—say, at the agency level—only informs analysts of whether resources are being used in line with authorized amounts and spending within the budget. Such aggregate analysis can hide important details, such as overspending on some items and underspending on others.[6] Disaggregation in the analysis frequently leads to insights. For example, overspending on the wage bill in the health sector is often associated with expenditure cuts on goods and supplies or capital expenditures (Piatti-Fünfkirchen, Barroy, et al. 2021). This undermines the quality of the health services provided.

Third, a *transactions profile* can be developed as a useful way to map out expenditure patterns and management (Hashim et al. 2019). The transactions profile is a measure that gauges how government expenditure transactions are distributed by size. The actual pattern of financial transactions can have significant implications for how activities are actually being executed and, hence, can be useful for understanding what is driving effective government functioning. To do this, analysts can calculate the number of transactions, the percentage of transactions, the cumulative share of the number of transactions, and the cumulative share of the amount processed through the FMIS for specific sets of transaction types. Table 11.4 provides a sample template.

TABLE 11.4 Template for a Government Expenditure Transactions Profile

Range (US$ equivalent)	Number of transactions	Share of transactions (%)	Cumulative share (%)	Total amount of transactions (US$)	Share of amount processed through FMIS (%)	Cumulative share of amount processed through FMIS (%)
<100						
100–200						
200–500						
500–1k						
1k–5k						
5k–10k						
10k–25k						
25k–100k						
100k–500k						
500k–1,000k						
1,000k–50,000k						
>50,000k						
Total						

Source: Hashim et al. 2019.
Note: FMIS = financial management information system.

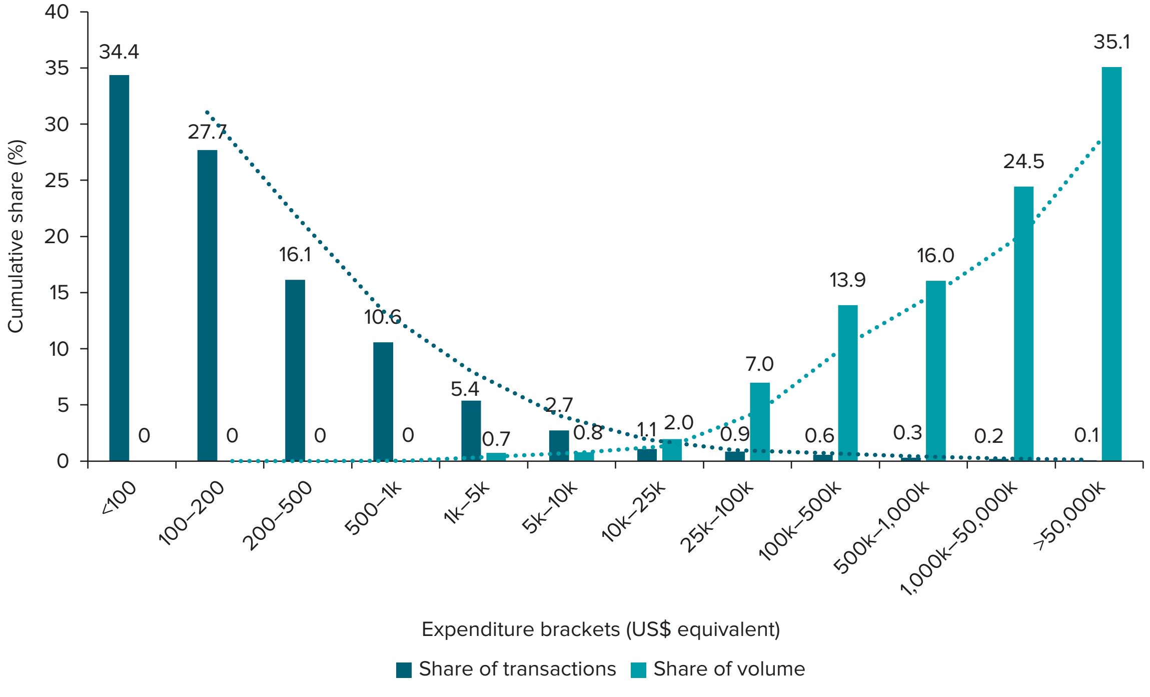
The transactions profile can then be displayed graphically (figure 11.3 provides an example from Bangladesh), where expenditure brackets are plotted against the cumulative share of the number of transactions and value of transactions. Typically, a larger percentage of transactions are small value transactions and even in sum cover only a small share of total spending. At the same time, high value transactions tend to be few in number but make up a large share of the total volume of spending.

Assessing the Attributes of Expenditure Data

As well as providing useful descriptions of basic patterns in the expenditure data, budget execution data, FMIS coverage data, and the transactions profile offer useful information for analysts on the expenditure data's attributes (see the section above on Attributes of Good-Quality Expenditure Data). Specifically, analysts may further probe the data in the following ways.

To assess *integrity and data provenance*, analysts may first wish to get clarity on how various transactions are processed and what kinds of controls they are subject to. For example, how are debt payments, subsidies, wage payments, or payments for goods and services handled, and is this equal across all expenditure items? A useful starting point may be to document which transactions are processed through the FMIS and which ones are not. Follow-up questions may then relate to whether the current process for how various transactions are treated is adequate from an integrity and provenance perspective. Does it suffice to merely post some transactions, such as wage payments or debt payments, to the general ledger? This also opens an important political-economic dimension because it may show the revealed preferences of governments that wish to control spending on certain line items (for example, not using the FMIS would make it easier to adjust spending by the executive without legislative approval). Therefore, discussing this openly and bringing transparency into the process would be a useful first step. Second, analysts may wish to identify technical challenges in routing certain transactions through the FMIS and then explore how advancements in

FIGURE 11.3 Expenditure Transactions Profile, Bangladesh



Source: Hashim et al. 2019.

maturing technologies (for example, financial technology innovations or the use of blockchain technology) could help strengthen the process.

As part of assessing the *comprehensiveness* of government expenditure data, analysts may wish to critically review how the government and the broader public sector are defined within the data. This should be followed by an assessment of whether these are appropriately reported across agencies. Identifying potential shortcomings in comprehensiveness, such as a lack of reporting on sensitive sectors or expenditure arrears, is another red flag for the FMIS data, as may be reporting against various select appropriation types. Such checks will minimize the risk of misinterpreting the findings and establishing poor indicators and targets that are poor representations of true spending patterns. These red flags are an opportunity for improvements in the comprehensiveness of expenditure reporting.

To assess the *usefulness* of government expenditure data, analysts may wish to explore what elements are captured and how they relate to government priorities. Do the data allow analysts to identify who spent funds, what they spent them on, and whether this usefully informs progress against the priorities set out in the agenda? On the question of who spends, it would be useful for the data to have sufficient detail in the administrative classification. Is it possible, for example, to know which hospital, health clinic, or school received what budget? What they spent it on should then be clear from the line item or activity segment. What purpose they spent it on (for example, malaria, primary education, and so on) can potentially be derived from the functional classification, but it can be difficult to establish this well. If the government has a functional classification, it may be useful to review how the mapping is generated and how well it serves its purpose. Given all of the above, the overarching questions for analysts will then be how well the established classification of expenditure data can be used to inform government priorities and what can be done to improve it.

To assess the *consistency* of the data, analysts can check whether there is consistency in the application of the COA across levels of decentralization and information systems across the government to allow for adequate integration. Analysts may also check for quality in the application of data entry to ensure the COA has been used as intended. Inconsistencies in the actual application can lead to problems in analysis and interpretation. Finally, in environments where development partners are an important source of revenue, analysts can review whether they have followed the same basis for accounting as the government to allow for the integration of expenditure data and comprehensive reporting.

Finally, to assess the *stability* of the data, analysts can review major changes in the expenditure data structure over time. If these are evident, analysts may explore whether tools to compare spending over time have been developed to give policy makers a multiyear perspective on important expenditure areas. With a solid understanding of the strengths and weaknesses of the underlying data, analysts can then use this expenditure and budget execution data to pursue efficiency, equity, or sustainability analyses to inform the effectiveness of government.

Case Study: Investigating Ineffective Capital Expenditure in Indonesia

At the request of Indonesia's Ministry of Finance, the World Bank conducted an institutional diagnostic to understand the causes of "low and slow" capital budget expenditure execution (World Bank 2020). The study is an example of the value of drilling down deep on expenditure data, with information from 11,589 spending units and survey responses from nearly 2,000 spending units. By matching spending data and survey responses, the study identified that over 80 percent of capital budget allocations were directed to only 3 percent of spending units, and 78 percent were directed to four ministries, all of which had lower execution rates than others.[7]

The survey indicated that line ministries found planning difficult because they were not provided with predictable indicative budget ceilings for the next three years. They therefore prepared capital projects to align with annual budgets. Only 6 percent of spending units used multiyear contracts. The rest split their projects across annual contracts, leading to inefficiencies in contract implementation that contributed to low budget execution. For example, in 2019, disbursements were bunched at the end of the year, with 44 percent being made in the fourth quarter.

Compounding this, annual budgets tended to be very rigid, with expenditure authority lapsing at the end of the year. This led to a stop-start approach to projects due to the annual cessation of appropriation approval, limiting the administrative ability of agencies to implement the capital works program, given the multiyear nature of many projects.

The analysis also allowed World Bank staff to assess whether preexisting reforms to confront these problems were working. They did not seem to be. The spending units of only one ministry—the Ministry of Public Works and Housing—made use of early procurement, which was supported by a ministerial decree. While there was a government regulation that enabled spending units to begin the procurement process in the preceding year, 60 percent of spending units prepared their procurement plans after the start of the new fiscal year, thereby introducing bunching and delays in the execution of the program.

At least part of the root cause came from the supplier side. Half of all spending units faced difficulties in ensuring that vendors submitted invoices within five days of finishing work. Further, 73 percent reported that incomplete proof in vendors' invoices was the main cause for delays in preparing payment requests. The analysis also identified other areas of concern. Some 42 percent of spending units reported that difficulties in obtaining land approvals delayed contract implementation. A particular blockage occurred in cases where the land value, determined in a quasi-judicial proceeding for land acquisition, was greater than the budget. There was also a concern that fiduciary (audit) control discouraged spending units' performance in project implementation. Some 14 percent of spending units said that auditors created a delay in implementation, and 32 percent of respondents preferred splitting activities into multiple contracts to avoid the audit of large contracts.

Overall, this detailed diagnostic enabled specific, practical recommendations for improved government management. It was only made possible by triangulating microlevel expenditure data at the spending unit with survey data.

CONCLUSION

Government expenditure data can assist our understanding of the functioning of government agencies, acting as a basis for conducting broader efficiency, equity, or productivity analyses.[8] Such analyses can be valuable and informative for policy and for improving the quality of the administration. However, expenditure data are only useful for these ends if they also have the right attributes.

All technical solutions require an enabling environment of government commitment, actionable political economy, and resilience to shocks. It is important that strong systems for government expenditure data are in place and protected during times of adversity. Governments are encouraged to put in place processes that identify deficiencies in routines to allow for strengthening over time. The root causes of distortions may take considerable effort to uncover. Political leadership and a willingness to embrace transparency in the identification process are key.

This chapter has provided health warnings that should be considered when using expenditure data and has identified the following five attributes of good-quality expenditure data:

- Data provenance and integrity
- Comprehensive across space and over time
- Usefulness
- Consistency
- Stability.

How well government expenditure data meet the above attributes is rarely emphasized in analytical work or considered directly in its underlying methodologies. Instead, expenditure data are often taken at face value, with the implicit assumption that the above conditions are met. If they are not, it can render the analysis incorrect and misleading.

This chapter suggests a periodic and data-driven review of these issues in all budgeting systems. For example, expenditure data can be used to estimate FMIS budget coverage. Such statistics provide insight into whether budget managers have incentives to avoid FMIS internal controls. This chapter advocates for estimating budget coverage periodically and making it publicly available in an effort to deepen the understanding of the incentives and the underlying political economy of budget controls. A step beyond this is to assess how variation in expenditure management relates to government effectiveness.

Budget coverage statistics could accompany analytical products that draw on these data to offer cautions in the interpretation of the data. Audit institutions can report on why FMIS coverage may be low and what can be done to strengthen it in their management letters and reports to the legislature.[9] Alongside this indicator, a transactions profile can be mapped to identify where risks in current expenditure management may lie and what types of reform may be warranted to improve expenditure control and service delivery objectives.

High-quality government expenditure microdata can be used by analysts to provide insight into expenditure management practices, functional effectiveness, and the related pursuit of public policy. A basic analysis simply assesses how capable expenditure units are at absorbing and spending funds.

The analysis of expenditure data benefits from triangulation with performance information on spending units to guide a dialogue on public sector effectiveness. Just as reviewing the calories one takes in without considering the activities undertaken may shed little light on the fitness and workings of one's metabolism, so, too, is the consideration of expenditure data limited if not aligned with the impacts of the activities being funded.

The strongest analysis frames the discussion of expenditure in terms of a logframe of expenditure (figure 11.1): where do expenditure data come from and how is expenditure defined, what are their quality and comprehensiveness, and how do they impact government effectiveness? Framing the discussion within government in terms of these steps is important because it facilitates noticing and learning (Hanna, Mullainathan, and Schwartzstein 2012). The "failure to notice" systemic problems may be a key binding constraint in reaching the production frontier if practitioners only excel at one aspect of the logframe—in this case, the analysis of data without sufficient regard to their origins and quality.[10]

It almost goes without saying that expenditure data may not be everything in the pursuit of government effectiveness. Some organizations spend very little but have very important public mandates, such as a policy, coordination, or regulatory function. However, for some of the most important government functions—such as the building of large capital projects—expenditure data can be a critical lens for understanding government functioning.

NOTES

1. Expenditure data can also capture governments' responses to shocks through reallocation and adjustments to their revealed preferences (Brumby and Verhoeven 2010). After the global financial crisis, expenditure analysis showed that countries were temporarily expanding safety nets, protecting social sector spending through loans, redirecting funding to retain social spending, and harnessing the crisis to achieve major reforms to improve efficiency and quality.
2. Lowry (2016) estimates that health-related tax expenditures in the United States involved almost US$300 billion in 2019.
3. The PEFA program provides a framework for assessing and reporting on the strengths and weaknesses of public financial management (PFM), using quantitative indicators to measure performance. PEFA is designed to provide a snapshot of PFM performance at specific points in time using a methodology that can be replicated in successive assessments, giving a summary of changes over time.
4. More information about the BOOST initiative is available on the World Bank's website at https://www.worldbank.org/en/programs/boost-portal.

5. More broadly, Kaufmann and Bellver (2005) find that transparency is associated with better socioeconomic and human development indicators, higher competitiveness, and reduced corruption. They show that for countries with the same level of income, a country with a more transparent environment tends to have more-effective government agencies. Glennerster and Shin (2008) find that countries experience statistically significant declines in borrowing costs as they become more transparent.
6. PEFA assessments can offer valuable information on budget execution rates. Not spending as intended and spending more than intended are considered equally problematic. A 15 percentage point deviation from the original appropriation is considered poor practice by the PEFA because, at that point, it likely renders the budget not credible or effective.
7. Over 64 percent of the capital budget was allocated to spending units in Jawa.
8. Beyond governments, these data are also used by international organizations for Public Expenditure Reviews (PERs), Public Expenditure Tracking Surveys, Commitment to Equity Assessments, and Article IV agreements.
9. As blockchain technology matures, it may also offer a pathway to the immutability of records, making them less susceptible to manipulation.
10. The "learning through noticing" approach alters the standard intuition that experience guarantees effective technology use (see, for example, Foster and Rosenzweig 2010; Nelson and Phelps 1966; Schultz 1975).

REFERENCES

African Union. 2001. *Abuja Declaration on HIV/AIDS, Tuberculosis and Other Related Infectious Diseases*. African Summit on HIV/AIDS, Tuberculosis, and Other Related Infectious Diseases, Abuja, Nigeria, April 24–27, 2001. OAU/SPS/ABUJA/3. https://au.int/sites/default/files/pages/32894-file-2001-abuja-declaration.pdf.

Allen, Richard, and Daniel Tommasi, eds. 2001. *Managing Public Expenditure: A Reference Book for Transition Countries*. Paris: OECD Publishing.

Baker Tilly Business Services Limited. 2014. *National Audit Office Malawi: Report on Fraud and Mismanagement of Malawi Government Finances*. Report to the Auditor General of the Government of Malawi. London: Baker Tilly Business Services Limited.

Baldacci, Emanuele, and Kevin Fletcher. 2004. "A Framework for Fiscal Debt Sustainability Analysis in Low-Income Countries." In *Helping Countries Develop: The Role of Fiscal Policy*, edited by Sanjeev Gupta, Benedict Clements, and Gabriele Inchauste, 130–61. Washington, DC: International Monetary Fund.

Ball, Amanda, Suzana Grubnic, and Jeff Birchall. 2014. "Sustainability Accounting and Accountability in the Public Sector." In *Sustainability Accounting and Accountability*, 2nd ed., edited by Jan Bebbington, Jeffrey Unerman, and Brendan O'Dwyer, 176–96. London: Routledge.

Barton, Allan. 2011. "Why Governments Should Use the Government Finance Statistics Accounting System." *Abacus* 47 (4): 411–45.

Binyaruka, Peter, August Kuwawenaruwa, Mariam Ally, Moritz Piatti, and Gemini Mtei. 2021. "Assessment of Equity in Healthcare Financing and Benefits Distribution in Tanzania: A Cross-Sectional Study Protocol." *BMJ Open* 11 (9): e045807. http://doi.org/10.1136/bmjopen-2020-045807.

Bridges, Kate, and Michael Woolcock. 2017. "How (Not) to Fix Problems That Matter: Assessing and Responding to Malawi's History of Institutional Reform." Policy Research Working Paper 8289, World Bank, Washington, DC.

Brumby, Jim, and Marijn Verhoeven. 2010. "Public Expenditure after the Global Financial Crisis." In *The Day after Tomorrow: A Handbook on the Future of Economic Policy in the Developing* World, edited by Otaviano Canuto and Marcelo Giugale, 193–206. Washington, DC: World Bank.

Burnside, Craig. 2004. "Assessing New Approaches to Fiscal Sustainability Analysis." Working paper, Economic Policy and Debt Department, World Bank, Washington, DC.

Burnside, Craig, ed. 2005. *Fiscal Sustainability in Theory and Practice: A Handbook*. Washington, DC: World Bank.

Challen, Don, and Craig Jeffery. 2003. "Harmonisation of Government Finance Statistics and Generally Accepted Accounting Principles." *Australian Accounting Review* 13 (30): 48–53.

Chami, Ralph, Raphael Espinoza, and Peter Montiel, eds. 2021. *Macroeconomic Policy in Fragile States*. Oxford, UK: Oxford University Press.

Chan, James L. 2006. "IPSAS and Government Accounting Reform in Developing Countries." In *Accounting Reform in the Public Sector: Mimicry, Fad or Necessity*, edited by Evelyne Lande and Jean-Claude Scheid, 31–42. Paris: Expert Comptable Média.

Coelli, Timothy J., D. S. Prasada Rao, Christopher J. O'Donnell, and George Edward Battese. 2005. *An Introduction to Efficiency and Productivity Analysis*. New York: Springer.

Cooper, Julie, and Sailendra Pattanayak. 2011. *Chart of Accounts: A Critical Element of the Public Financial Management Framework*. Washington, DC: International Monetary Fund.

de Renzio, Paolo, and Chloe Cho. 2020. "Exploring the Determinants of Budget Credibility." Working paper, International Budget Partnership, Washington, DC.

Di Bella, Gabriel. 2008. "A Stochastic Framework for Public Debt Sustainability Analysis." IMF Working Paper WP/08/58, International Monetary Fund, Washington, DC.

European Commission and IEG (Independent Evaluation Group, World Bank). 2017. *Joint Evaluation of Budget Support to Ghana (2005–2015): Final Report*. Brussels, Belgium: European Commission.

Eurostat. 2019. *Manual on Sources and Methods for the Compilation of COFOG Statistics: Classification of the Functions of Government (COFOG)*. Luxembourg: Publications Office of the European Union.

Farooq, Khuram, and Michael Schaeffer. 2017. "Simplify Program Budgeting: Is There a Place for 'Activities' in a Program Classification?" *IMF Public Financial Management Blog*, October 30, 2017. International Monetary Fund. https://blog-pfm.imf.org/en/pfmblog/2017/10/simplify-program-budgeting-is-there-a-place-for-activities-in-a-program-classifi.

Foster, Andrew D., and Mark R. Rosenzweig. 2010. "Microeconomics of Technology Adoption." *Annual Review of Economics* 2: 395–424.

Glennerster, Rachel, and Yongseok Shin. 2008. "Does Transparency Pay?" *IMF Staff Papers* 55 (1): 183–209.

Greene, William H. 2008. "The Econometric Approach to Efficiency Analysis." In *The Measurement of Productive Efficiency and Productivity Growth*, edited by Harold O. Fried, C. A. Knox Lovell, and Shelton S. Schmidt, 92–250. New York: Oxford University Press.

Griffin, Charles C., David de Ferranti, Courtney Tolmie, Justin Jacinto, Graeme Ramshaw, and Chinyere Bun. 2010. *Lives in the Balance: Improving Accountability for Public Spending in Developing Countries*. Washington, DC: Brookings Institution Press.

Hameed, Farhan. 2005. "Fiscal Transparency and Economic Outcomes." IMF Working Paper WP/05/225, International Monetary Fund, Washington, DC.

Hameed, Farhan. 2011. "Budget Transparency and Financial Markets." Working paper, International Budget Partnership, Washington, DC.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2012. "Learning through Noticing: Theory and Experimental Evidence in Farming." NBER Working Paper 18401, National Bureau of Economic Research, Cambridge, MA.

Hashim, Ali. 2014. *A Handbook on Financial Management Information Systems for Government: A Practitioners Guide for Setting Reform Priorities, Systems Design and Implementation*. Africa Operations Services Series. Washington, DC: World Bank.

Hashim, Ali, Khuram Farooq, and Moritz Piatti-Fünfkirchen. 2020. *Ensuring Better PFM Outcomes with FMIS Investments: An Operational Guidance Note for FMIS Project Teams Designing and Implementing FMIS Solutions*. Guidance Note, Governance Global Practice. Washington, DC: World Bank.

Hashim, Ali, and Moritz Piatti-Fünfkirchen. 2018. "Lessons from Reforming Financial Management Information Systems: A Review of the Evidence." Policy Research Working Paper 8312, World Bank, Washington, DC.

Hashim, Ali, Moritz Piatti-Fünfkirchen, Winston Cole, Ammar Naqvi, Akmal Minallah, Maun Prathna, and Sokbunthoeun So. 2019. "The Use of Data Analytics Techniques to Assess the Functioning of a Government's Financial Management Information System: An Application to Pakistan and Cambodia." Policy Research Working Paper 8689, World Bank, Washington, DC.

IFAC (International Federation of Accountants). 2022. *Handbook of International Public Sector Accounting Pronouncements*. New York: IFAC.

IMF (International Monetary Fund). 2014. *Government Finance Statistics Manual 2014*. Washington, DC: IMF.

Islam, Roumeen. 2003. "Do More Transparent Governments Govern Better?" Policy Research Working Paper 3077, World Bank, Washington, DC.

Kaufmann, Daniel, and Ana Bellver. 2005. "Transparenting Transparency: Initial Empirics and Policy Applications." Draft discussion paper presented at the International Monetary Fund conference on transparency and integrity, July 6–7, 2005, World Bank, Washington, DC.

Kirigia, Joses M., Ali Emrouznejad, and Luis G. Sambo. 2002. "Measurement of Technical Efficiency of Public Hospitals in Kenya: Using Data Envelopment Analysis." *Journal of Medical Systems* 26 (1): 39–45.

Lowry, Sean. 2016. *Health-Related Tax Expenditures: Overview and Analysis*. CRS Report Prepared for Members and Committees of Congress. Washington, DC: Congressional Research Service.

Lustig, Nora. 2011. "Commitment to Equity Assessment (CEQ): A Diagnostic Framework to Assess Governments' Fiscal Policies Handbook." Tulane Economics Working Paper 1122, Department of Economics, Tulane University, New Orleans, LA.

Lustig, Nora, ed. 2018. *Commitment to Equity Handbook: Estimating the Impact of Fiscal Policy on Inequality and Poverty*. Washington, DC: Brookings Institution Press.

McAfee, R. Preston, and John McMillan. 1989. "Government Procurement and International Trade." *Journal of International Economics* 26 (3–4): 291–308.

Milante, Gary, and Michael Woolcock. 2021. "Fiscal Policy in Fragile Situations: Flying in Fog with Limited Instrumentation." In *Macroeconomic Policy in Fragile States*, edited by Ralph Chami, Raphael Espinoza, and Peter Montiel, 271–96. Oxford, UK: Oxford University Press.

Mills, Anne, John E. Ataguba, James Akazili, Jo Borghi, Bertha Garshong, Suzan Makawia, Gemini Mtei, et al. 2012. "Equity in Financing and Use of Health Care in Ghana, South Africa, and Tanzania: Implications for Paths to Universal Coverage." *The Lancet* 380 (9837): 126–33.

Morozumi, Atsuyoshi, and Francisco José Veiga. 2016. "Public Spending and Growth: The Role of Government Accountability." *European Economic Review* 89: 148–71.

Moses, Mark W., Julius Korir, Wu Zeng, Anita Musiega, Joyce Oyasi, Ruoyan Lu, Jane Chuma, and Laura Di Giorgio. 2021. "Performance Assessment of the County Healthcare Systems in Kenya: A Mixed-Methods Analysis." *BMJ Global Health* 6 (6): e004707.

Mtei, Gemini, Suzan Makawia, Mariam Ally, August Kuwawenaruwa, Filip Meheus, and Josephine Borghi. 2012. "Who Pays and Who Benefits from Health Care? An Assessment of Equity in Health Care Financing and Benefit Distribution in Tanzania." *Health Policy and Planning* 27 (suppl. 1): i23–i34.

NAO (National Audit Office, UK). 2020. "Assessing Value for Money." *Successful Commissioning Toolkit*. United Kingdom Government. https://www.nao.org.uk/successful-commissioning/general-principles/value-for-money/assessing-value-for-money/.

Negash, Solomon, and Paul Gray. 2008. "Business Intelligence." In *Handbook on Decision Support Systems*, edited by Frada Burstein and Clyde W. Holsapple, 2: 175–93. Berlin: Springer.

Nelson, Richard R., and Edmund S. Phelps. 1966. "Investment in Humans, Technological Diffusion, and Economic Growth." *American Economic Review* 56 (1/2): 69–75.

PEFA (Public Expenditure and Financial Accountability). 2020. *Global Report on Public Financial Management*. Washington, DC: PEFA Secretariat.

PEFA (Public Expenditure and Financial Accountability). 2022. *Global Report on Public Financial Management*. Washington, DC: PEFA Secretariat.

PEMPAL (Public Expenditure Management Peer Assisted Learning). 2014. "Integration of the Budget Classification and Chart of Accounts: Good Practice among Treasury Community of Practice Member Countries." PEMPAL.

Piatti, Moritz, Ali Hashim, and Clay G. Wescott. 2017. "Using Financial Management Information Systems (FMIS) for Fiscal Control: Applying a Risk-Based Approach for Early Results in the Reform Process." Paper presented at the IPMN Conference on Reform, Innovation, and Governance: Improving Performance and Accountability in the Changing Times, School of International and Public Affairs, the China Institute of Urban Governance, and the Center for Reform, Innovation, and Governance of Shanghai Jiao Tong University, Shanghai, China, August 17–18, 2017. http://dx.doi.org/10.2139/ssrn.3090673.

Piatti-Fünfkrichen, Moritz. 2016. "What Can We Learn from a Decade of Public Financial Management and Civil Service Reform in Malawi?" IEG Project Lessons, World Bank, Washington, DC, April 22, 2016. https://ieg.worldbankgroup.org/news/what-can-we-learn-decade-public-financial-management-and-civil-service-reform-malawi.

Piatti-Fünfkirchen, Moritz, Helene Barroy, Fedja Pivodic, and Federica Margini. 2021. *Budget Execution in Health: Concepts, Trends and Policy Issues*. Washington, DC: World Bank.

Piatti-Fünfkirchen, Moritz, Ali Hashim, Sarah Alkenbrack, and Srinivas Gurazada. 2021. *Following the Government Playbook? Channeling Development Assistance for Health through Country Systems*. Washington, DC: World Bank.

Piatti-Fünfkirchen, Moritz, Ali Hashim, and Khuram Farooq. 2019. "Balancing Control and Flexibility in Public Expenditure Management: Using Banking Sector Innovations for Improved Expenditure Control and Effective Service Delivery." Policy Research Working Paper 9029, World Bank, Washington, DC.

Piatti-Fünfkirchen, Moritz, Liying Liang, Jonathan Kweku Akuoku, and Patrice Mwitende. 2020. *Rwanda Nutrition Expenditure and Institutional Review 2020*. Washington, DC: World Bank.

Piatti-Fünfkirchen, Moritz, Magnus Lindelow, and Katelyn Yoo. 2018. "What Are Governments Spending on Health in East and Southern Africa?" *Health Systems and Reform* 4 (4): 284–99.

Potter, Barry H., and Jack Diamond. 1999. *Guidelines for Public Expenditure Management*. Washington, DC: International Monetary Fund.

Pradhan, Sanjay. 1996. "Evaluating Public Spending: A Framework for Public Expenditure Reviews." World Bank Discussion Paper 323, World Bank, Washington, DC.

Rivetti, Diego. 2021. *Debt Transparency in Developing Economies*. Washington, DC: World Bank.

Robinson, Marc. 2000. "Contract Budgeting." *Public Administration* 78 (1): 75–90.

Schiavo-Campo, Salvatore. 2017. *Government Budgeting and Expenditure Management: Principles and International Practice*. New York: Taylor & Francis.

Schultz, Theodore W. 1975. "The Value of the Ability to Deal with Disequilibria." *Journal of Economic Literature* 13 (3): 827–46.

Shah, Anwar, ed. 2005. *Public Expenditure Analysis*. Public Sector Governance and Accountability Series. Washington, DC: World Bank.

Shah, Anwar, ed. 2007. *Budgeting and Budgetary Institutions*. Public Sector Governance and Accountability Series. Washington, DC: World Bank.

Stokey, Edith, and Richard Zeckhauser. 1978. *A Primer for Policy Analysis*. New York: Norton.

Tanzi, Vito, and Ludger Schuknecht. 2000. *Public Spending in the 20th Century: A Global Perspective*. Cambridge, UK: Cambridge University Press.

Wagstaff, Adam. 2012. "Benefit-Incidence Analysis: Are Government Health Expenditures More Pro-Rich Than We Think?" *Health Economics* 21 (4): 351–66.

World Bank. 2016a. *Evaluation of the Malawi Financial Management, Transparency, and Accountability Project*. Project Performance Assessment Report, Report 103060. Washington, DC: World Bank.

World Bank. 2016b. *Zambia Public Sector Management Program Support Project*. Project Performance Assessment Report, Report 106280. Washington, DC: World Bank.

World Bank. 2020. *Indonesia Revenue and Budget Management: Institutional Diagnostic of Low and Slow Central Government Capital Budget Execution.* Report AUS0001636. Washington, DC: World Bank.

World Bank. 2022. *Zimbabwe Health Public Expenditure Review*. Washington, DC: World Bank.

CHAPTER 12

Government Analytics Using Procurement Data

Serena Cocciolo, Sushmita Samaddar, and Mihaly Fazekas

SUMMARY

The digitalization of national public procurement systems across the world has opened enormous opportunities to measure and analyze procurement data. The use of data analytics on public procurement data allows governments to strategically monitor procurement markets and trends, to improve the procurement and contracting process through data-driven policy making, and to assess the potential trade-offs of distinct procurement strategies or reforms. This chapter provides insights into conducting research and data analysis on public procurement using administrative data. It provides an overview of indicators and data sources typically available on public procurement and how they can be used for data-driven decision-making, the necessary data infrastructure and capacity for optimizing the benefits from procurement data analytics, and the added value of combining public procurement data with other data sources. Governments can take various steps to create the conditions for effectively using data for decision-making in the area of public procurement, such as centralizing public procurement data, periodically assessing their quality and completeness, and building statistical capacity and data analytics skills in procurement authorities and contracting entities.

ANALYTICS IN PRACTICE

- The increasing availability of public procurement administrative microdata should be exploited for evidence-based decision-making. The digitalization of national public procurement systems across the world has opened enormous opportunities to measure procurement outcomes through the analysis of administrative data now available in machine-readable formats on electronic government procurement (e-GP) systems. The full potential of e-GP reforms can be realized when data analytical tools are systematically applied at scale for the monitoring and evaluation of public procurement.

Serena Cocciolo is an economist at the World Bank. Sushmita Samaddar is a researcher at the University of Kansas. Mihaly Fazekas is an assistant professor at the Central European University and scientific director at the Government Transparency Institute.

- Procurement data analytics can be used for monitoring and characterizing public procurement. Public procurement data can be used to characterize national public procurement spending; describe time trends; compare procurement performance across procuring entities, regions, and types of contract, as well as across types of procedure, sector, or supplier; and identify performance and compliance gaps in the national public procurement system. Interactive dashboards are increasingly widespread tools for monitoring public procurement through descriptive analysis because they enable procurement authorities to track, analyze, and display key performance indicators through customizable and user-friendly visualizations.

- Procurement data analytics can be used for data-driven policy making. The analysis of public procurement data can enable procurement agencies to develop key procurement policies or refine and assess existing regulations. First, data analytics allows agencies to assess existing efficiency gaps and understand the drivers of performance; these empirical insights are useful to identify and prioritize potential areas for interventions and reform efforts. Second, data analytics allows agencies to monitor the consequences of new policies, assess whether they are delivering the expected outcomes, and understand potential trade-offs. Especially in cases where an e-GP system already exists at the time of piloting and implementing new strategies, public procurement can also be a rich space for research and impact evaluations because the necessary data for tracking key outcome indicators are readily available from the existing e-GP system.

- Appropriate data infrastructure and capacity are necessary for effectively using public procurement data for decision-making. First, procurement data should be homogeneously collected and maintained across procuring entities and connected to a centralized platform. Second, data generated from different stages of the procurement cycle (for example, tendering process, bidding process, bid evaluation, contract award, and contract signing) should be consistently organized and connected through key identifiers. Third, the availability of data should be expanded to cover the full public procurement and contract management cycle, including parts of the process that are not typically included in procurement data, such as data on public procurement planning and budgeting, tender preparation data, contract execution data, and complaints data. Fourth, data quality and completeness should be improved through relatively simple and practical steps by the government, such as automated data quality checks in the e-GP system and periodic data audits. Finally, the necessary capacity for statistical analysis should be built in the public procurement authority, potentially including the creation of a dedicated statistical unit.

- A "whole-of-government" approach should be adopted in procurement data analytics. Public procurement is multidimensional and critically interconnected with other functions of the public sector and public administration. Yet the integration of e-procurement systems into other e-government systems is not yet a common practice. Data innovations should enable the integration of public procurement data with administrative microdata from other parts of the public sector, such as justice, firm registries, and tax administration. This would provide a comprehensive picture of the procurement function, holistically explore the environment within which procurement is conducted, and enable the government to develop innovative and impactful procurement strategies.

- Procurement data analytics should move beyond traditional public procurement indicators and data sources. While there is widespread consensus about the measurement framework for some dimensions of public procurement, including costs, price efficiency, integrity risks, transparency, and competition, other relevant aspects of public procurement, such as the inclusiveness and sustainability of public procurement and the quality of contract implementation, currently lack well-defined and commonly used indicators. Using nontraditional public procurement data can contribute to the development of new measures and expand the scope of public procurement data analytics, such as survey data with firms or procurement officers.

INTRODUCTION

While it is difficult to measure the size of public procurement transactions in each country, a global exercise by Bosio et al. (2022) estimates that around 12 percent of the global gross domestic product is spent on public procurement—the process by which governments purchase goods, services, and works from the private sector. Given this massive scale, public procurement has the potential to become a strategic policy tool in three crucial ways.

First, improved public procurement can generate sizeable savings and create additional fiscal space by reducing the price of purchases and increasing the efficiency of the procurement process (Bandiera, Prat, and Valletti 2009; Best, Hjort, and Szakonyi 2019; Singer et al. 2009).[1] Second, public procurement can support national socioeconomic and environmental aspirations by encouraging the participation of local small firms in the public contract market, promoting green and sustainable procurement, and creating jobs through large public works (Ferraz, Finan, and Szerman 2015; Krasnokutskaya and Seim 2011). Finally, efficient public procurement can improve the quality of public services through several channels, such as the selection of higher-quality goods, more timely delivery of goods and completion of public infrastructure, and better planning of purchases and stock management. Given these strategic functions, efficient and effective public procurement can contribute to the achievement of the development goals of ending poverty and promoting shared prosperity.[2]

Data and evidence are necessary to monitor public procurement spending and identify the optimal policies and strategies for efficient, inclusive, and sustainable procurement. The use of data can contribute to a problem-driven, iterative approach to strengthening and modernizing national public procurement systems through the identification of efficiency and integrity gaps, analysis of the trade-offs associated with alternative procurement strategies, the development of data tools for monitoring the public procurement function, and the generation of knowledge and evidence on the impact of certain policies.

The digitalization of national public procurement systems across the world has opened enormous opportunities to measure procurement outcomes through the analysis of administrative data now available in machine-readable formats on electronic government procurement (e-GP) systems. E-procurement refers to the integration of digital technologies to replace or redesign paper-based procedures throughout the procurement cycle (OECD 2021). While countries are increasingly digitalizing public procurement processes, the functionalities covered by e-GP systems vary widely across countries (box 12.1), and this has implications for the accessibility and quality of procurement and contract data for analysis and research. Map 12.1 shows advancements in e-GP adoption globally and highlights the different degrees of sophistication of national e-GP systems, depending on the extent to which various procurement stages—advertisement, bid submission, bid opening, evaluation, contract signing, contract management, and payment—can be implemented electronically.[3]

Governments can take various steps to create the conditions for effectively using data for decision-making in the area of public procurement, such as centralizing public procurement data, periodically assessing their quality and completeness, creating the data infrastructure for integrating data from various stages of the procurement cycle and from other e-government systems, measuring the socioeconomic and environmental dimensions of government purchases, integrating procurement data and systems into other e-government data and systems, and building statistical capacity and data analytics skills in procurement authorities and contracting entities.

This chapter provides insights and lessons on how to leverage administrative microdata for efficient and strategic public procurement. The chapter provides an overview of indicators and data sources typically available on public procurement and how they can be used for data-driven decision-making (section 2), the necessary data infrastructure and capacity for optimizing the benefits from procurement data analytics (section 3), and the added value of combining public procurement data with other data sources (section 4).

BOX 12.1 Types of Digitalization of Public Procurement Systems

The degree to which the procurement process is digitalized and integrated with other functions of government plays an important role in determining the accessibility and quality of administrative procurement microdata and how they can be used for conducting data analysis and research on public procurement.

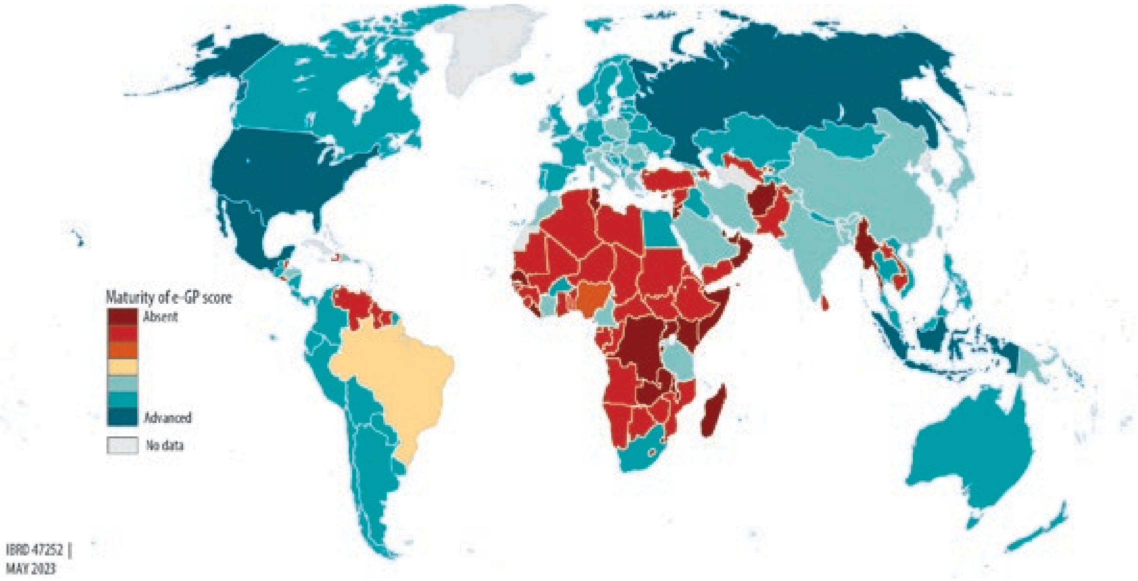
The digitalization of public procurement systems has been implemented in different ways across the world, with implications for data availability and quality. Most commonly, electronic government procurements (e-GP) systems are used to publish and store public procurement information. For example, in Pakistan and Tanzania, the online procurement system allows for the upload of tender and contract documents as scanned copies or PDFs.[a] In these cases, data would first need to be scraped from PDF documents and organized in a structured manner before any kind of data analysis could be performed. In fewer countries, the e-GP system includes functionalities related to the transactional aspects of public procurement, such as e-tendering, electronic submission of bids, e-evaluation, e-awarding, and, in the most advanced cases, electronic submission of invoices, e-catalogs, and contract management. In these cases, the e-GP system generates data in machine-readable formats, readily available for analysis. For example, in Colombia, a data management system has been implemented following the Open Contracting Data Standard guidelines on data transparency, so the data from the e-GP system can be downloaded in the form of Excel files and readily used for analysis.[b]

There are variations in the quality and completeness of data generated from e-GP systems, as well as in how well the data from different parts of the procurement process can be integrated or merged for a holistic view of government purchases. The integration of e-procurement systems into other e-government systems is not yet a common practice, and further work is needed to promote this "whole-of-government" approach from a data perspective.

a. For Pakistan, see World Bank (2017). For Tanzania, see the example of an invitation for bids from the Tanzania National Roads Agency (Tender AE/001/2020-21/HQ/G/79) available at https://www.afdb.org/sites/default/files/documents/project-related-procurement/invitation_for_tenders_-_edms.pdf.

b. For more information about the Open Contracting Data Standard, see the project website at https://standard.open-contracting.org/latest/en/.

MAP 12.1 Use of Electronic Government Procurements across the World, 2020



Source: World Bank, based on Doing Business 2020 Contracting with the Government database, https://archive.doingbusiness.org/en/data/exploretopics/contracting-with-the-government#data.

Note: The Maturity of e-GP score was calculated based on the number of features existing in the electronic government procurement (e-GP) system portal, as reported in the World Bank's Contracting with the Government database.

HOW DO WE USE PUBLIC PROCUREMENT DATA FOR DECISION-MAKING?

Procurement Indicators Used for Data Analysis

Based on the perspective that public procurement is a strategic function contributing to efficient public spending, as well as to the achievement of national socioeconomic and environmental objectives, this chapter provides a holistic view of public procurement. While the application of data analytical tools is often associated with the use of corruption flags to uncover malpractice, this focus risks discouraging governments from using and opening public procurement data. Data analytical tools' main purpose is strengthening the efficiency of public procurement and government spending in achieving national objectives, and a stronger focus on these more comprehensive goals could help reduce resistance from governments to adopting them.[4] Following this view, in this section, we present a broad set of procurement indicators and uses of procurement data analytics corresponding to a wide range of objectives, including (but not only) anticorruption goals. Table 12.1 provides an example of public procurement indicators that can be used to measure the performance of the public procurement system along the dimensions described in the following paragraphs: economy and efficiency, transparency and integrity, competition, inclusiveness, and sustainability.

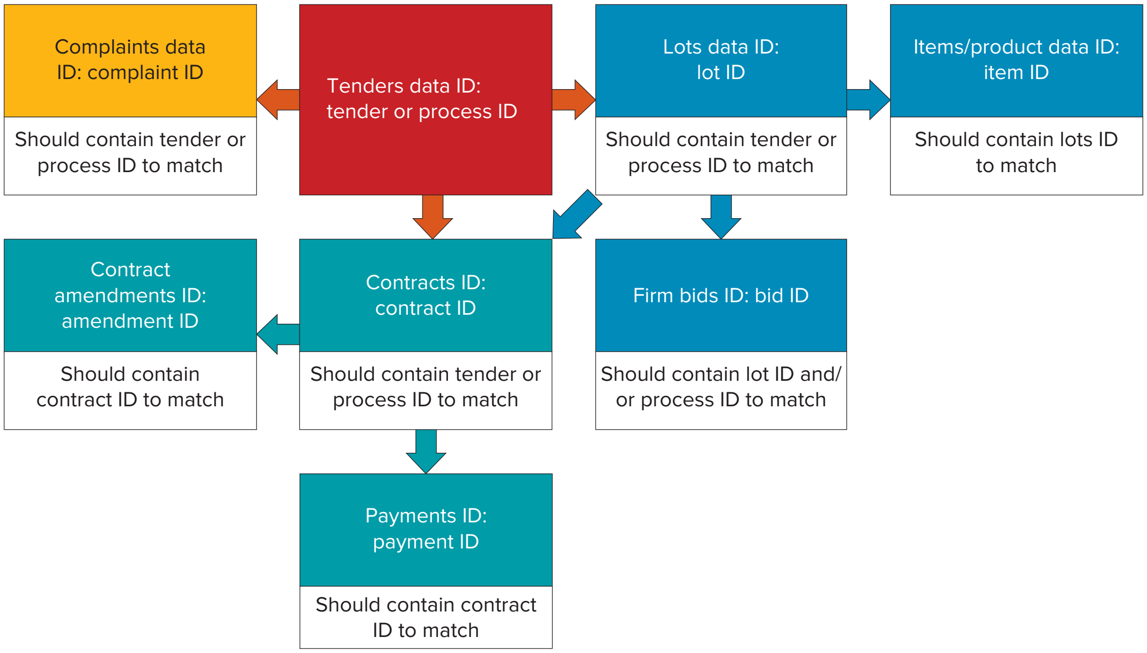
The procurement and contracting cycle refers to a sequence of related activities, from needs assessment through competition and award to payment and contract management, as well as any subsequent monitoring or auditing (OECD 2021). It is typically divided into the following stages: (1) budget planning and tender preparation; (2) tendering process, bidding process, and bid evaluation; (3) contract award and contract signing; and (4) contract execution and monitoring. Traditional public procurement data often cover only stages (2) and (3) because the other stages are typically managed by other units (budget and financial management) and therefore recorded in separate systems. These data can be organized at the tender, lot, item (product), bid, and contract levels. Figure 12.1 provides a visual representation of how the different levels of public procurement data connect. Specifically, tenders can be divided into lots, and each lot can specify different product items. Firms submit bids to specific tenders or lots and can submit for specific tenders; tenders result in one or more contracts, which are then linked to contract amendments and payments. Understanding the structure of public procurement data and the links between different stages is the first step for effectively using and analyzing it. For example, the e-GP systems for Brazil, Romania, Croatia, and Honduras organize open procurement data at the tender, lot, contract, and bid levels, allowing users to connect these different stages of the process through unique identifiers for each data set.

TABLE 12.1 Examples of Public Procurement Indicators

Economy and efficiency	Transparency and integrity	Competition	Inclusiveness and sustainability
Tender and bidding process			
• Total processing time • Evaluation time • Contracting time	• Time for bid preparation • Single-bidder tender	• Open procedure • Number of bidders • Share of new bidders	• Share of SME bidders • Share of WOE bidders
Assessment and contracting			
• Awarded unit price • Final unit price after renegotiation	• Share of excluded bids	• Number of bidders • New bidders	• Share of SME bidders • Share of WOE bidders
Contract implementation			
• Final unit price after renegotiation • Time overrun	• Variation orders • Renegotiations		

Source: Original table for this publication.
Note: SME = small and medium enterprise; WOE = women-owned enterprise.

FIGURE 12.1 Data Map of Traditional Public Procurement Data


Complaints data ID: complaint ID
Should contain tender or process ID to match
Tenders data ID: tender or process ID
Lots data ID: lot ID
Should contain tender or process ID to match
Items/product data ID: item ID
Should contain lots ID to match
Contract amendments ID: amendment ID
Should contain contract ID to match
Contracts ID: contract ID
Should contain tender or process ID to match
Firm bids ID: bid ID
Should contain lot ID and/or process ID to match
Payments ID: payment ID
Should contain contract ID to match

Source: Original figure for this publication.

The academic literature and practitioners in the field have identified a common set of indicators that are typically used to measure the efficiency, effectiveness, and integrity of the public procurement function. These indicators cover dimensions of public procurement related to methods and procedures (for example, use of open methods), transparency and integrity (for example, time for bid submission), competition (for example, number of bidders), processing time (for example, time for bid evaluation), price (for example, unit prices), and contract implementation (for example, time overrun). (A full list of indicators is provided in appendix D.) In addition to performance indicators, public procurement microdata can also be used for the construction of red flags for corruption or collusion risk. The richness of the data available on public tenders has allowed economists, anticorruption authorities, and competition agencies to develop different screening techniques and has offered the opportunity to test them empirically. Red flags can be useful to identify unusual patterns in certain markets, but these patterns are not sufficient evidence of misbehavior. Rather, red flags can be used as the starting point for further investigation and as sufficient evidence for courts to authorize inspections of dawn raids (OECD 2013). One reason why red flags cannot provide sufficient proof of corruption or collusion is that by design, these data-driven methods can produce false positives (by flagging cases that do not merit further scrutiny) and false negatives (by failing to identify cases that do merit further scrutiny). Given that corruption risk indicators and cartel screens do not directly point to illegal activities, establishing their validity is of central importance.[5] Boxes 12.2 and 12.3 present the existing literature on corruption risk indicators and cartel screens and some recent advances in these techniques thanks to novel machine-learning applications.

Beyond a transactional view of public procurement, there is increasing interest in measuring dimensions of public procurement related to the strategic role it can play to promote inclusive and sustainable growth and the achievement of socioeconomic and environmental objectives. Recent studies and research on these topics have focused both on the development of new procurement indicators (for example, on green procurement and socially responsible procurement) and on linking public procurement data with other data sources to promote a holistic approach to data analytics (for example, firm registry microdata). These topics are discussed in more detail in section 5.

An area that would require further development and research is the measurement of contract implementation quality. Various approaches have been experimented with in the literature, but there is no agreed-upon strategy yet, and this is a dimension where data constraints are particularly binding. One option would be to use audits data, but the limitations are that audits often focus on compliance with procurement regulations rather than on actual project implementation and that audits data are not typically integrated with public procurement data. Contract supervision data and project management reports could also be used to generate information on contract implementation. The potential for integrating data from various stages of the public procurement cycle and from other functions of the state is discussed further in section 4. Ad hoc data collection could also be considered for specific sectors—for example, through engineering assessments of the material used for the construction of infrastructure projects (Olken 2007) or through visits to hospitals to verify the availability of medicines and their quality. With respect to the construction sector, recent advances in technology (for example, drones and satellite images) can monitor the progress—but not necessarily the quality—of construction work, while information on quality can be obtained from citizen monitoring. More work is needed to assess the pros and cons of different measurement strategies, particularly in terms of the objectivity of different measurement approaches and their scalability.

BOX 12.2 What We Know about Corruption Risk Indicators

The starting point for measuring any corrupt phenomenon is to define the particular behaviors of interest (Mungiu-Pippidi and Fazekas 2020). In public procurement, one definition widely used in both academia and policy considers corruption to be the violation of impartial access to public contracts—that is, a deliberate restriction of open competition to the benefit of a connected firm or firms (Fazekas and Kocsis 2020).

Corruption risk indicators identify the factors and traces of corrupt transactions defined by deliberate competition restrictions favoring connected bidders. Widely used corruption risk indicators in public procurement include single bidding in competitive markets, restricted and closed procedure types, or the lack of publication of the call for tenders (Fazekas, Cingolani, and Tóth 2018). These risk indicators have been shown to correlate with already established indexes of corruption, such as the Control of Corruption scores in the Worldwide Governance Indicators (Fazekas and Kocsis 2020), as well as with other markers of corruption, such as the price of auctions (Fazekas and Tóth 2018), the political connections of bidding firms (Titl and Geys 2019), and proven cases of corruption (Decarolis et al. 2020).

Novel machine-learning applications have been used to advance the measurement of corruption risks. For example, machine-learning approaches have been used on carefully curated data sets of proven corrupt and noncorrupt cases to train algorithms predicting corruption risks (Decarolis and Giorgiantonio 2022; Fazekas, Sberna, and Vannucci 2021). Advanced network science methods have also been increasingly used to detect high-corruption-risk groups of organizations (Wachs, Fazekas, and Kertész 2021).

Corruption risk indicators have been used in numerous civil society and journalistic applications, as well as by multilateral banks and national authorities for policy design and implementation. For example, the European Commission and Organisation for Economic Co-operation and Development's (OECD) Support for Improvement in Governance and Management (SIGMA) initiative (OECD and SIGMA 2019) has regularly monitored some risk indicators, such as single bidding and the publication of calls for tenders. The International Monetary Fund (IMF) has endorsed corruption risk indicators and models predicting the price impacts of such risks as valuable inputs to addressing macrocritical risks. The European Investment Bank uses public procurement risk indicators, combined with internal financial risk assessments, to select projects for prior integrity reviews (Fazekas, Ugale, and Zhao 2019), an approach highlighted as good practice by the European Court of Auditors (Adam and Fazekas 2019).

BOX 12.3 What We Know about Collusion and Cartel Screens

The characteristics of collusive behavior in public procurement markets are similar to those in conventional markets: companies coordinate their behavior regarding price, quantity, quality, or geographic presence to increase market prices.

Cartel screens are defined according to two key competition and economy principles. First, it is expected that in competitive tendering processes, bids will be submitted independently; therefore, signs of coordination between bidders can be interpreted as signs of collusion. Second, bids submitted by independent competitors should appropriately reflect the costs of each bidder in a competitive market. Based on these two criteria, various elementary collusion risk indicators have been developed for the early detection of collusive bidding, such as the submission of identical bids, high correlation between bids, lack of correlation between the costs and the bid submitted by each bidder, the relative difference between the lowest and the second lowest bid price per tender, the relative standard deviation of bid prices per tender, and the range of submitted bid prices per tender.

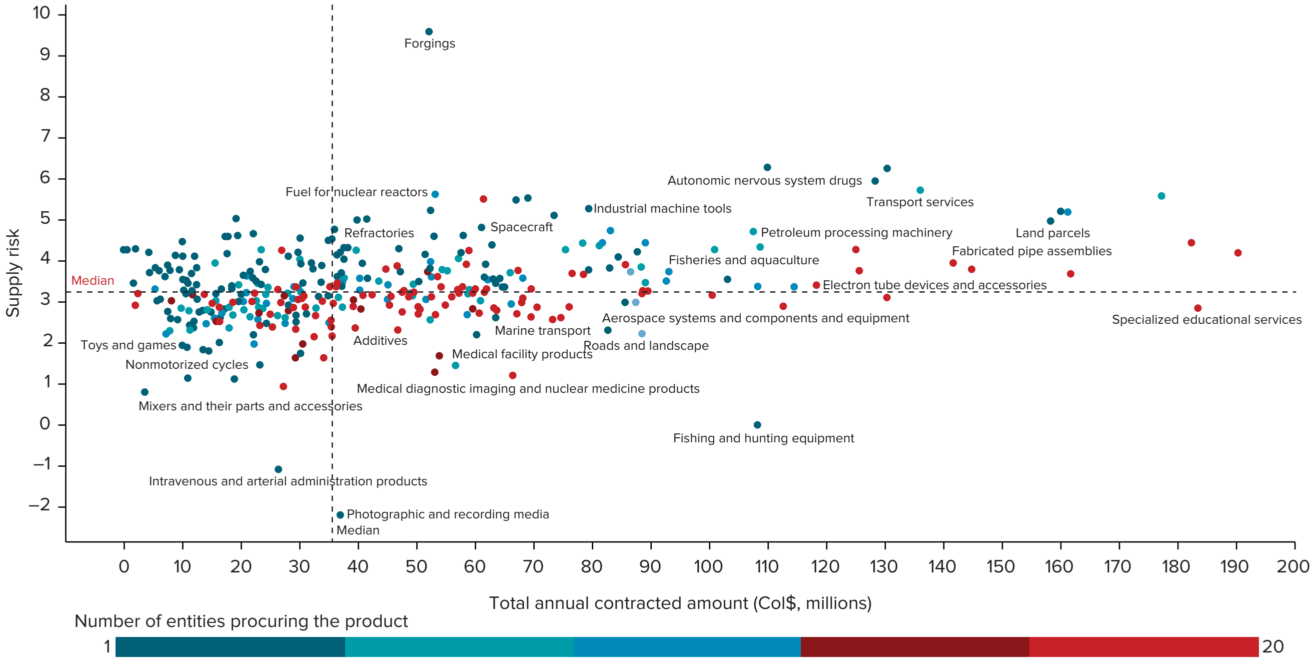
Increasingly, more advanced statistical techniques have been used to define cartel screens as well as develop algorithms that minimize the probability of both false positives and false negatives. For example, Conley and Decarolis (2016) have developed statistical tests of coordination based on randomization inference methods. These tests identify unexpected firm behaviors conditional on their characteristics—for example, the unexpected joint entry of firms within a group of bidders given observed firm and process characteristics. Huber and Imhof (2019) study the performance of different screening algorithms, specifically a lasso logit regression and a weighted average of predictions based on bagged regression trees, random forests, and neural networks. Most interestingly, these recent examples use machine-learning techniques to identify optimal algorithms thanks to the combination of public procurement data and judicial and auctions data for validation.

Government Monitoring of Public Procurement

With the increasing use of e-GP systems and access to structured procurement data, public procurement authorities are often using the common procurement indicators discussed in appendix D to monitor the performance of their own public procurement systems. These public procurement authorities use the available procurement data to characterize national public procurement spending and identify performance and compliance gaps in the national public procurement system. This descriptive analysis can include time trends or comparisons of performance indicators across procuring entities, regions, and types of contract, as well as types of procedure, sector, or supplier. In some cases, this exercise may be mandated by international treaties or organizations, or as a prerequisite to access financing from multilateral development banks.[6] The results of this monitoring are often reported in the form of annual reports on the functioning of the procurement system, and they can be used for informing and guiding reform efforts and the development of new strategies and policies in public procurement. For example, in Poland, the Public Procurement Office (PPO) prepares the annual report on the functioning of the procurement system, which is posted on the PPO website following approval by the Council of Ministers.[7]

Public procurement agencies may use certain tools or mechanisms to describe their procurement data and trends. For example, spend analysis is a widespread approach for monitoring and assessing public procurement, consisting of various tools (for example, the Kraljic matrix; see figure 12.2) that provide a first overview of the procurement market and, specifically, what is being purchased, by whom, and from which suppliers. This analysis is used to identify the areas (products, entities, and suppliers) for which the improvement of efficiencies is expected to have the largest budget implications, to define procurement strategies, and to adapt relationship management for different types of suppliers.

FIGURE 12.2 Kraljic Matrix for Colombia

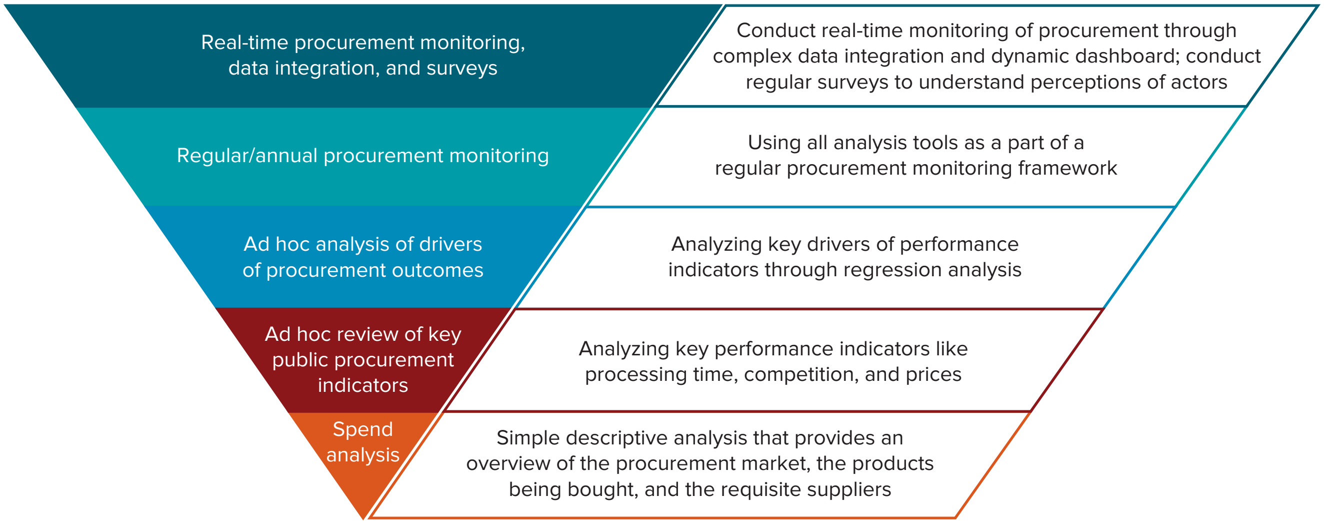


Source: Original figure for this publication based on Colombia's open public procurement data.
Note: Col$ = Colombian peso.

The analysis of performance and compliance offers another set of tools typically used by public procurement authorities and audit authorities to monitor the national public procurement system. This monitoring may include the compliance of procurement regulations as reported by procurement agencies.[8] This type of descriptive analysis explores efficiency indicators, like competition, price, and processing time, as well as the extent to which procurement regulations (for example, regulations on contract thresholds, the use of open participation methods, or the use of centralized procurement methods) are met. This type of analysis is useful to describe the efficiency gaps that exist in the public procurement system and to help prioritize audit activities. For example, Best, Hjort, and Szakonyi (2019) show that in Russia, individuals and organizations of the bureaucracy together account for more than 40 percent of the variation in prices paid and that moving the worst-performing quartile of procurers to 75th percentile effectiveness would reduce procurement expenditures by around 11 percent, or US$13 billion each year.

As illustrated in figure 12.3, these descriptive analysis tools are the least complex uses of public procurement administrative data. Figure 12.3 shows a ladder for analysis tools in procurement monitoring, in which each step of the ladder represents analytical tools conducted on procurement at different levels of complexity. Beyond descriptive analytics, diagnostic analysis (for example, regression analysis) can be used to identify the drivers of performance and therefore inform the government of potential strategies to improve efficiency and integrity. The following section discusses in detail diagnostic analysis tools for data-driven policy making. However, descriptive analysis tools can still be among the most advanced uses of public procurement when they are systematically embedded in the public procurement monitoring and reporting function—for example, for the preparation of annual reports or through interactive dashboards, which typically require institutional reorganization and the acquisition of necessary skills in the public procurement authority.

FIGURE 12.3 Complexity Ladder for Analysis Tools in Procurement Monitoring and Evaluation


Real-time procurement monitoring, data integration, and surveys
Conduct real-time monitoring of procurement through complex data integration and dynamic dashboard; conduct regular surveys to understand perceptions of actors
Regular/annual procurement monitoring
Using all analysis tools as a part of a regular procurement monitoring framework
Ad hoc analysis of drivers of procurement outcomes
Analyzing key drivers of performance indicators through regression analysis
Ad hoc review of key public procurement indicators
Analyzing key performance indicators like processing time, competition, and prices
Spend analysis
Simple descriptive analysis that provides an overview of the procurement market, the products being bought, and the requisite suppliers

Source: Original figure for this publication.

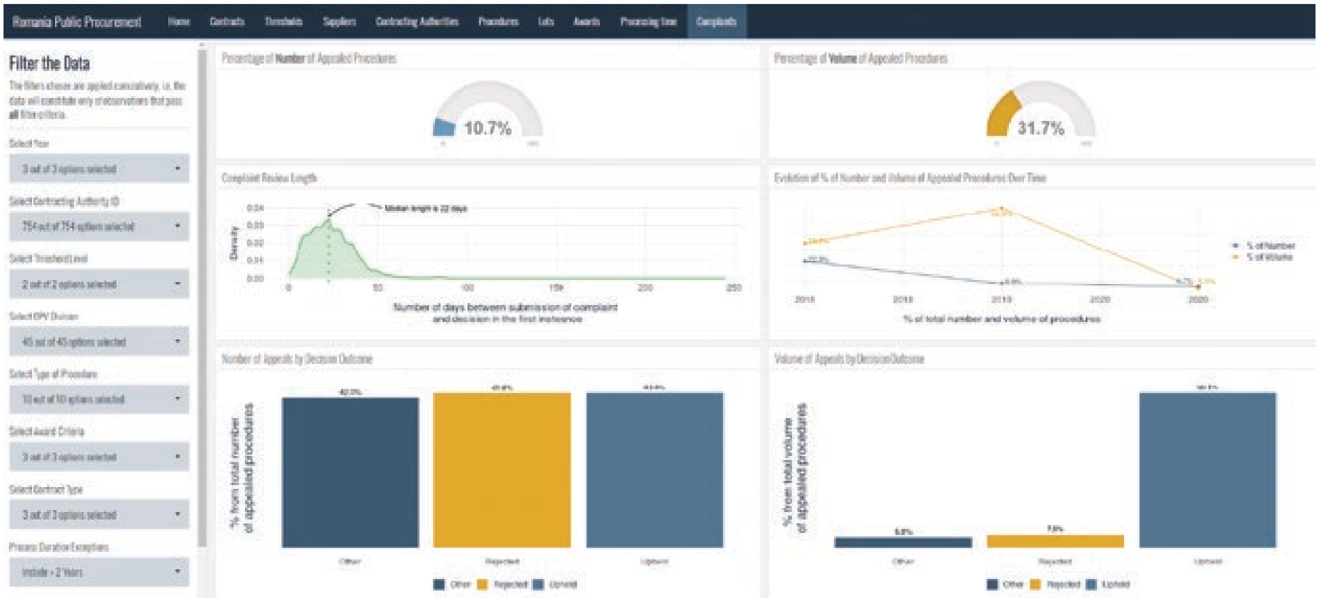
Interactive dashboards are increasingly widespread tools for monitoring public procurement through descriptive analysis because they enable procurement authorities to track, analyze, and display key performance indicators through customizable and user-friendly visualizations. One of the great advantages of these dashboards is that they allow users to focus their analysis at different levels of government or in specific markets. Depending on how the public procurement system is set up, these interactive dashboards can be connected directly with the underlying e-GP system or can be regularly updated. These dashboards can be built for the use of the national public procurement authorities and individual procuring entities for monitoring their procurement activity, or they can be made public for greater accountability of the public procurement system.

For example, between 2020 and 2021, the World Bank worked with the National Agency for Public Procurement (ANAP) in Romania to develop a monitoring mechanism, in the form of a dashboard that would enable the public procurement agency to track its own key performance indicators and would enable easy reporting to the EU (World Bank 2019). The dashboard (figure 12.4) was developed in close collaboration with the ANAP to ensure that the most relevant indicators were captured. Regular data analysis workshops conducted by the World Bank ensured that staff in the ANAP had the capacity and training to replicate and add to the dashboard to ensure its sustainability in the long run.

Data-Driven Policy Making

The analysis of public procurement data can enable procurement agencies to develop key procurement policies or refine and assess existing regulations. Data analytics allows agencies to assess existing efficiency gaps and understand the drivers of performance, and these empirical insights are useful to identify and prioritize potential areas for interventions and reform efforts. For example, in 2019, the World Bank conducted a complete and comprehensive analysis of Uruguay's public procurement data that generated discussion and space for policy recommendations to improve the performance of the procurement system. This analysis identified demand consolidation as the most significant potential source of savings, with framework agreements being the most effective instrument to implement the strategy. Based on these empirical insights, in 2021, the World Bank worked with the Regulatory Agency for Public Procurement and the Central Procurement Unit within the Ministry of Economy and Finance to implement these recommendations, specifically building capacity in the generation and management of framework agreements and supporting the development of pilot framework agreements for goods and services with the greatest savings potential.

FIGURE 12.4 National Agency for Public Procurement (ANAP) Dashboard, Romania


Filter the Data

The filters chosen are applied cumulatively, i.e. the data will constitute only of observations that pass **all** filter criteria.

Select Year: 3 out of 3 options selected

Select Contracting Authority ID: 754 out of 754 options selected

Select Threshold Level: 2 out of 2 options selected

Select CPV Division: 45 out of 45 options selected

Select Type of Procedure: 10 out of 10 options selected

Select Award Criteria: 3 out of 3 options selected

Select Contract Type: 3 out of 3 options selected

Process Duration Exceptions: Include > 2 Years

Romania Public Procurement
Home
Contracts
Thresholds
Suppliers
Contracting Authorities
Procedures
Lots
Awards
Processing time
Complaints
Percentage of Number of Appealed Procedures
10.7%
Percentage of Volume of Appealed Procedures
31.7%
Complaint Review Length
Median length is 22 days
Density
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
Number of days between submission of complaint and decision in the first instance
Evolution of % of Number and Volume of Appealed Procedures Over Time
% of Number
% of Volume
2018
2018
2019
2020
2020
% of total number and volume of procedures
Number of Appeals by Decision Outcome
% from total number of appealed procedures
Other
Rejected
Upheld
Volume of Appeals by Decision Outcome
% from total volume of appealed procedures
5.8%
7.8%
Other
Rejected
Upheld

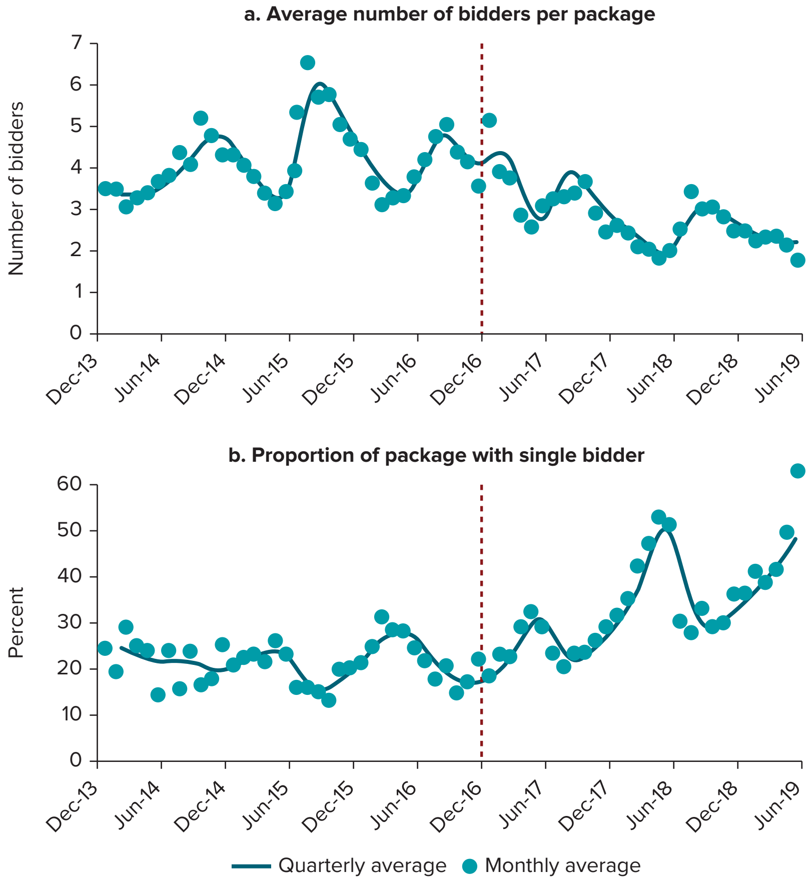
Source: Screenshot of the ANAP dashboard, World Bank 2019.

Data analytics is also a useful tool to monitor the consequences of new policies, assess whether they are delivering the expected outcomes, and understand potential trade-offs. For example, in 2020, the World Bank worked on an assessment of the national public procurement system in Bangladesh, the objectives of which were to identify its strengths and weaknesses, formulate appropriate mitigation measures for identified gaps, and develop an action plan for future system development (World Bank 2020). The assessment was built on various data-driven components, such as an analysis of the so-called 10 percent rule (rejecting a tender that is 10 percent below or above the estimated cost) introduced by the government of Bangladesh in December 2016 for the procurement of works using the open tendering method. The primary objective of this policy was to improve the quality of construction works and reduce the risk of cost overruns by restricting bidders from quoting very low prices. However, procuring entity officers largely expressed the opinion that the quality of works had not improved after the introduction of the 10 percent rule, and quantitative analysis of time trends also revealed that this rule had produced undesired consequences, such as decreasing competition (figure 12.5). These empirical insights were instrumental in providing fact-based recommendations to the government about reevaluating the 10 percent rule.

With respect to understanding potential trade-offs, increasing attention toward the multidimensional nature of public procurement implies that policies and strategies should be assessed based on a variety of considerations, including efficiency, integrity, value for money, and socioeconomic and environmental aspects. There are many trade-offs associated with the public procurement function in connection to the private sector and public service delivery, and a comprehensive approach to procurement data analytics allows agencies to correctly assess the potential trade-offs associated with procurement policies and provide complete and accurate policy recommendations. For example, a 2021 World Bank report on the use of framework agreements (World Bank 2021a) shows that in Brazil, the use of framework agreements could reduce unit prices and avoid repetitive processes, but it could also discourage participation by small and medium enterprises (SMEs) and their likelihood of being awarded a contract (table 12.2).[9]

These examples show that quantitative analysis can be quite powerful in identifying key procurement trends in a country and can be foundational in developing and evaluating procurement policies. Given the

FIGURE 12.5 Assessment of Bangladesh's 10 Percent Rule



Source: World Bank 2020.

TABLE 12.2 Regression Analysis of Framework Agreements versus Other Open Methods, Brazil

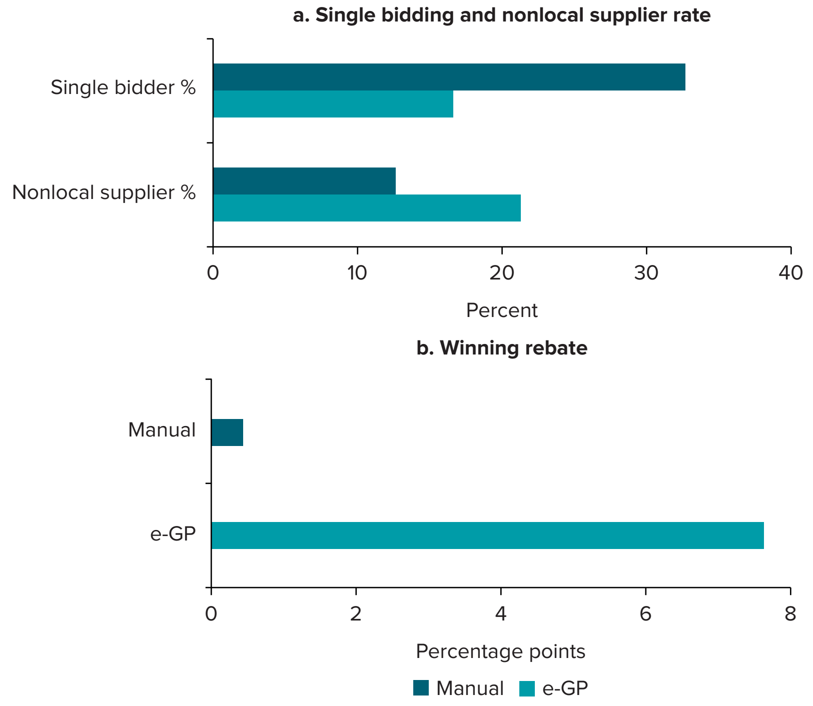
Outcome of interest	Unit price (log)	SME winner
Framework agreements vs. other open methods	−0.0919** (0.0407)	−0.0198*** (0.00582)
Observations	172,605	166,399
R-squared	0.910	0.566

Source: World Bank 2021a.
Note: Model specifications: Comparing FAs and non-FA open methods for the purchase of the same product by the same entity (product—entity FE), with year and quarter FEs. FA = framework agreement; FE = fixed effect; SME = small and medium enterprise. Robust standard errors in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

abundance of microdata in countries with e-GP systems, public procurement can also be an ideal space for implementing impact evaluations of specific procurement policies. Impact evaluations represent one of the most reliable forms of policy assessment because they allow agencies to contrast actual outcomes with counterfactual scenarios by comparing units subjected to a given policy intervention to otherwise similar units that have not yet been treated.

For example, starting in 2011, the government of Bangladesh began the rollout of a comprehensive e-GP system, and the World Bank worked with the Central Procurement Technical Unit to evaluate the impact of the new system on procurement outcomes.[10] The evaluation revealed that the implementation of the e-GP system had led to an improvement in public procurement performance, as demonstrated, for example, by an increase in winning rebates, a decrease in single bidding, and an increase in contracts awarded to nonlocal suppliers (figure 12.6). During the piloting stage, the preliminary results from the evaluation

FIGURE 12.6 Procurement Outcomes under Manual versus Electronic Government Procurement Systems, Bangladesh



Source: Turkewitz, Fazekas, and Islam 2020.
Note: e-GP = electronic government procurement.

helped demonstrate that the new e-GP system was having a good impact on efficiency, transparency, and competition, and this was extremely useful to build consensus and political support around this difficult reform (Turkewitz, Fazekas, and Islam 2020). This example demonstrates the feasibility and usefulness of impact evaluations for navigating the political economy of reforms. Specifically for public procurement, the abundance of data generated by the e-GP system creates a rich space for research because the data already available from the existing e-GP system at the time of piloting and implementing new strategies allow for the tracking of procurement outcomes, from baseline to endline, with no additional costs for data collection.

Monitoring of Public Procurement by the Public

Public procurement is one of the public administration functions with a prevalence of publicly available data. With the increase in the use of e-GP systems, there is greater potential for increasing transparency and accountability in public procurement processes through monitoring by the public. Open data can be used by civil society, the media, and citizens to acquire information on specific contracts, buyers, or products.[11] Increased transparency and accountability can enable citizen engagement in monitoring public procurement and, therefore, enhance trust between citizens and the state, strengthen the social contract, and improve the quality of contract execution. For example, a citizen engagement project was implemented by the World Bank in collaboration with the Central Procurement Technical Unit in Bangladesh to enable and support the monitoring of civil works projects by local community groups (Hassan 2017). Through the project, "citizen committee" members frequently visited project offices and reported anomalies in civil works projects to the engineer's office. The project reduced information gaps and increased trust between government officials and local community leaders on civil works projects in their areas, and it also reduced monitoring-related transaction costs through higher citizen engagement.

Making public procurement data available to the public has great potential to increase transparency and accountability. However, even when data are publicly available online, it may be challenging to extract useful and actionable information from open public procurement data, which requires connecting data sources from different stages of the procurement cycle, constructing relevant indicators, and analyzing microdata for large numbers of tenders and contracts. In light of these challenges, various international organizations have been developing dashboards to ease access to public procurement data for citizens, civil society, and researchers.

For example, in 2021, the World Bank, in collaboration with the Government Transparency Institute, launched a prototype of a global dashboard that provides access to open data from national electronic procurement systems from 46 countries, as well as open data on World Bank– and Inter-American Development Bank–financed contracts for over 100 countries.[12] Similarly, the Opentender dashboard provides access to tender data from Tenders Electronic Daily (TED), which covers 33 EU jurisdictions, including 28 EU member states, Norway, the EU institutions, Iceland, Switzerland, and Georgia.[13] The creation and maintenance of these global public goods, and the use of open public procurement data in general, would be simplified by the adoption of common data standards globally, and the Open Contracting Data Standard from the Open Contracting Partnership provides a promising starting point.[14]

Some governments are also creating public dashboards using their own public procurement data. For example, an intuitive and simple dashboard prepared by the National Informatics Centre in India and hosted on the Central Public Procurement Portal allows users to get some key performance and compliance indicators on public procurement in India.[15] This dashboard not only allows the government to monitor key procurement trends but also reports these indicators to the public for greater transparency and accountability. The public procurement authority in Ukraine has also designed and published a public dashboard to increase transparency and accountability in public procurement.[16] The COVID-19 crisis prompted more countries to increase transparency and enable public scrutiny of purchases made during the health emergency, such as Brazil, Moldova, Lithuania, and South Africa.[17]

SETTING UP DATA INFRASTRUCTURES AND CAPACITY FOR MEASUREMENT ON PROCUREMENT

Centralize Public Procurement Data

To ensure that public procurement data are used effectively for decision-making, it is necessary that data are homogeneously collected and maintained across procuring entities and connected to a centralized platform. This is necessary both for external users and researchers as well as for the national agency or central procurement authority. Public procurement data are often decentralized or housed by different institutions responsible for managing different parts of the procurement process, and this may complicate the process of data harmonization and centralization, especially in countries without an e-GP system and where reporting of procurement activities to a central authority is not mandatory or audited. For example, in 2021, a World Bank team conducted a data collection exercise of paper-based procurement records in St. Lucia, Dominica, St. Vincent and the Grenadines, and Grenada to assess public procurement systems in the four countries. The key constraint on data collection was the decentralization of the information among multiple institutions; ultimately, the data had to be collected by enumerators from different procuring entities and national authorities.

Enabling factors for the centralization of public procurement data include legislation, administrative structure, and data infrastructure. Simple mechanisms like an annual data collection exercise at the central level, in which procuring entities send Excel files to a central authority (which audits a sample for data accuracy), can help slowly transfer local data storage mechanisms to more efficient, centralized data management systems. For example, this was recommended in the case of St. Lucia, Dominica, St. Vincent and the Grenadines, and Grenada, with the additional recommendation to conduct regular audits of the quality and accuracy of the data provided by each procuring entity to the central authority. This step can be a key foundation on which an appetite for data literacy and digitalization can be created among governments. In contrast to data sitting in physical files in different procuring entities, a centralized data collection mechanism can allow for easy access to procurement data even in cases where an e-GP system has not yet been implemented.

Integrate Data from Various Stages of the Public Procurement and Contract Management Cycle

Data integration can be an important step in exploring all the stages of the public procurement and contract management cycle. Data integration can be accomplished through two related steps: (1) matching data from various procurement stages and (2) expanding the availability of data to study procurement more holistically. With respect to the first step, public procurement data typically cover the following stages: tendering process, bidding process, bid evaluation, contract award, and contract signing. To meaningfully use this data, it is necessary that the tenders data, lots data, bids data, and contracts data are consistently organized and can be connected (see figure 12.1).

The second step in data integration is expanding the availability of data to cover the full public procurement and contract management cycle, including parts of the process that are not typically included in procurement data, such as data on public procurement planning and budgeting, tender preparation data, contract execution data (for example, data on subcontracting and payments to vendors), complaints data, and proprietor information and beneficial ownership data.

There is great scope for using these additional data sources for procurement data analytics, and some countries are taking steps in this direction. The development of integrated data systems requires the close engagement and partnership of multiple government institutions that house different parts of the procurement and contract management cycle. For example, as part of the design and development of the monitoring mechanism delivered to the ANAP in Romania (see above), the World Bank was able to add data on complaints registered in public procurement processes to the dashboard by leveraging existing data-sharing agreements between the ANAP and the National Council for Solving Complaints. While the establishment of streamlined data management systems is a necessary technical requirement for a data integration process, the most significant constraints often lie in the administrative and bureaucratic structures that may complicate collaboration and data-sharing agreements between different institutions.

Data Quality and Completeness

Data quality and completeness are crucial determinants of the quality of empirical analysis that can be performed on public procurement data. Common issues in public procurement data are noted across countries, both in the data obtained from open sources as well as in the data obtained from governments. Some of these common issues, which are listed in box 12.4, range from missing data to incorrect or ambiguous data structures that restrict or hinder comprehensive empirical analysis.

Some data quality and completeness issues can be mitigated through relatively simple and practical steps by the government. The e-GP system can include automated data quality checks during data entry by procuring entity officers—for example, checking that the procurement process dates follow a logical order and that the contract amounts are within reasonably expected ranges. Detailed audits of the data entered by procuring entity officers may also be conducted regularly to ensure that the official tender and contract documents reflect the data entered into the system. The central procurement authority can also review the data maintained in the procurement system to assess their completeness, especially in light of the compliance and performance indicators the government is interested in monitoring. Last, implementing a fully transactional system that manages the entire procurement process from start to finish and allows multiple government agencies and ministries to engage with different parts of the procurement process may allow for the ideal data integration environment to holistically analyze the full procurement process and all related parts in public administration.

When planning for the public disclosure of procurement data, the same principles of data quality and completeness apply to ensure data transparency and accessibility. In addition, in this case, it is important that the raw data entered into the system are made public, not only the indicators and measures constructed from the administrative microdata. Observations across several countries also show that open data and good policies for data openness and transparency do not necessarily correlate with data quality

BOX 12.4 Examples of Common Issues with Data Quality and Completeness

- **Missing observations or variables and data errors that pertain to important aspects of the procurement process.** In most countries, electronic government procurement (e-GP) data are not created directly from digitized tenders and contracts but are separately inputted by procuring entity officers. In these cases, the procuring entity officers may still have the option to leave certain data fields blank. This generates gaps in the data and can also indicate strategic behavior by procuring entity officers, who may systematically choose to leave more sensitive data fields blank. Data quality and completeness should be systematically reviewed by a central authority, including for data disclosed to the public.
- **Ambiguity in the level of observation for the data.** Data from different stages of the public procurement cycle (figure 12.1) should be meaningfully connected for analysis through unique identifiers, such as the tender ID or the entity ID. The absence of unique identifiers creates ambiguity in the interpretation of the data and hinders comprehensive empirical analysis. For example, in the case of framework agreements, there can be multiple contracts, buyers, and suppliers under a tendering process, and multiple orders can be associated with the same umbrella contracts. Having a clear and unambiguous data structure is necessary to correctly represent framework agreement processes and enable accurate analysis.
- **Correction of incorrect entries in the e-GP system by entering the entire tendering process again.** This issue is observed in countries where e-GP data are not created directly from digitized tenders and contracts but are separately inputted by procuring entity officers. Possibly because of integrity concerns, some e-GP systems do not allow officers to correct information already entered into the system in the event of data-entry errors. In these cases, the officers' only option is to create a new entry, but the system does not record which entry is correct and which entry is wrong.
- **Poor data integration during transitions from one e-GP system to another.** Throughout the digitalization of public procurement, countries may shift from one e-GP platform to another. For example, Romania transitioned from a platform called SEAP (Sistemul Electronic de Achizitii Publice) to an upgraded platform called SICAP (Sistemului Electronic Colaborativ de Achizitii Publice), and Colombia from a platform called SECOP I to an upgraded platform called SECOP II (Sistema Electrónico para la Contratación Pública). In cases of transition between e-GP systems, it is necessary to ensure that data from both platforms can be integrated and that procuring entity officers cannot enter data for procurement processes in both platforms during the transition.

and completeness. For example, the Open Contracting Data Standard provides guidelines on the effective disclosure of public procurement data to the public, with the ultimate goal of increasing transparency in procurement and allowing analysis of procurement data by a wide range of users. While an increasing number of e-GP systems follow the Open Contracting Data Standard for the public disclosure of procurement data, how well disclosure is implemented largely depends on the quality and completeness of the data made publicly available.

Building Capacity for Statistical Analysis and a Culture of Data-Driven Policy Making

The adoption of e-GP systems has created a great wealth of data, but it is not obvious that their use and impact are currently being maximized by governments. The development of the capacity for statistical analysis and a culture of data-driven decision-making can help maximize the potential of the microdata available through e-GP platforms. This may include the creation or strengthening of a dedicated statistical office within the public procurement authority.

For example, as part of the design and development of the monitoring mechanism delivered to the ANAP in Romania (see above), the entire monitoring mechanism was created in close collaboration with ANAP staff through weekly capacity-building workshops and meetings to discuss the operational workflow of the monitoring mechanism. This close collaboration and cocreation of the interactive dashboard for visualizing key procurement indicators allowed the government to engage with the data-cleaning and visualization process and built an appetite for data analysis. ANAP staff were provided with the necessary skills and knowledge to edit and develop the code that was used to create the interactive dashboard. Engagements like this allow products like an interactive dashboard to be hosted in a data-curious and analytical environment that builds long-term sustainability through the empowerment of its users.

Beyond statistical capacity and data analytics skills, the proactive use of data and evidence to drive policy-making decisions also requires the necessary organizational culture, institutional arrangements, and incentive systems. For example, data and empirical evidence can be used to improve the performance of procuring entities. This requires the necessary skills and tools to exploit the potential of data analytics, but it also depends on other systemic factors, such as whether and how the performance of procuring entities is evaluated, whether there are consequences of performance evaluations, whether procuring entity officers are incentivized to improve their efficiency and effectiveness, and whether procuring entity officers have space to make discretional decisions or instead are expected to merely execute regulations. These considerations are related to a broader discussion on management practices in public administration and specifically in procuring entities, and the following section provides more detail on how some of these aspects can be studied empirically.

A WHOLE-OF-GOVERNMENT APPROACH: STRATEGIC COMPLEMENTARITIES TO PUBLIC PROCUREMENT DATA

Measuring the Socioeconomic and Environmental Dimensions of Public Procurement

Increasingly, governments consider using public procurement as a strategic tool to sustain the private sector, especially groups of firms that are traditionally underrepresented in public procurement, such as SMEs and women-owned enterprises (WOEs). Similarly, governments are increasingly adopting green public procurement (GPP) strategies, such as green evaluation criteria, green eligibility criteria, or life-cycle approaches to costing (box 12.5).[18]

However, there is no clear evidence of the best public procurement strategies and policies to achieve these socioeconomic and environmental outcomes. For example, from a theoretical point of view, it is not clear how to incentivize the participation of SMEs in public procurement effectively and efficiently. While this might be achieved through targeted policies (for example, preference policies or set-aside quotas), these policy tools might be distortionary (Medvedev et al. 2021; OECD 2018) or suffer from poor implementation and compliance. Relying on untargeted policies can be an alternative, but it is perhaps a less impactful approach. Two studies conducted on the same preferential treatment program for small firms in California elucidate these potential trade-offs, with Marion (2007) finding that procurement costs are 3.8 percent higher on auctions using preferential treatment and Krasnokutskaya and Seim (2011) finding that those distortionary effects are not huge in comparison to benefits to firm growth. With limited evidence on the impact and trade-offs of these different policy options, there are no clear guidelines on the best strategies to involve SMEs and other underrepresented groups in public procurement.

As another example, some public procurement laws mandate the application of green criteria for bid evaluation, especially in sectors such as transport (for example, types of vehicle and emissions) and construction (for example, construction materials) (Palmujoki, Parikka-Alhola, and Ekroos 2010), but it is unclear what the direct and indirect cost implications of these requirements are. By design, GPP introduces additional laws and regulations, requirements for firms, and more complex criteria for bid evaluation. Therefore, it is natural that there might be concerns about whether GPP compromises the efficiency of public procurement procedures and reduces the attractiveness of public procurement contracts for firms. Providing robust knowledge on the costs and benefits of GPP will support governments in making informed decisions and might remove some of the concerns that prevent broader adoption.

This focus and strategic approach to public procurement requires that public procurement data be expanded to include the necessary information to measure the socioeconomic and environmental dimensions of public procurement, such as by associating an SME tag with bidders and suppliers or by labeling tenders that follow GPP principles. For example, Nissinen, Parikka-Alhola, and Rita (2009) develop a detailed list of environmental indicators to measure GPP, including indicators on product characteristics, policy attached, level of emission of the company, chemistry, and amount of energy used. In practice, across countries, there has been some progress in tagging SME firms—for example, in Croatia, Romania, and Colombia—but very limited progress in GPP (see box 12.5). This impedes advancing the empirical literature on the effectiveness of different policy alternatives, and it also prevents governments and civil society from monitoring the actual use and implementation of GPP legislation.

BOX 12.5 What We Know about Green Public Procurement

Green public procurement (GPP) is defined by the European Commission (2008) as "a process whereby public authorities seek to procure goods, services and works with a reduced environmental impact throughout their life cycle when compared with goods, services and works with the same primary function that would otherwise be procured."

GPP can take different forms, and different measurement options should be considered depending on the GPP approach adopted for each specific tender. A first categorization of GPP approaches is as follows (World Bank 2021b):

- **Contract performance clauses** ensure winning suppliers deliver a contract in an environmentally friendly way and continuously improve their environmental performance throughout the contract duration. Examples of these clauses include the requirement to deliver goods in bulk to reduce packaging, the requirement to optimize delivery schedules, and the requirement to recycle or reuse packaging after delivery.
- **Award criteria** can include optional environmental criteria to encourage and reward bidders that propose solutions with improved environmental performance (for example, a higher percentage of recycled content and functional criteria that allow supplier innovation). This approach requires that procuring entities set weights to evaluate the various dimensions of a proposal, such as environmental criteria and price.
- **Qualification criteria and technical specifications** prescribe core environmental criteria that bidders and/or offers must meet to satisfy the requirements of the tender (for example, minimum recycled content or bans on toxic chemicals).[a] For example, supplier-selection criteria aim to ensure that participating bidders have the technical capabilities, ethics, and management processes in place to deliver on the desired environmental outcome. Examples of these criteria are proof of compliance with environmental laws and regulatory standards, the existence of qualified staff with environmental expertise, and environmental certifications.
- **Life-cycle approaches** consider the total cost of ownership (TCO) of a good, service, or work, an estimate that considers not only its purchase price but also the operational and maintenance costs over its entire life cycle. The life-cycle cost (LCC) goes further than the TCO by also taking into account the cost of environmental externalities that can be monetized (for example, greenhouse gas emissions and pollution fees).

Given the speed of innovations in this field, it may be challenging for procuring entities to define appropriate environmental criteria that correspond to current benchmarks and environmental criteria that can be expected of and met by private sector actors. There are various mechanisms that can help procuring entities determine the "environmental friendliness" of a good, service, work, or firm (World Bank 2021b):

(continues on next page)

BOX 12.5 What We Know about Green Public Procurement *(continued)*

- **Ecolabels** are labels of environmental excellence awarded to products and services meeting high environmental standards throughout their life cycle. Ecolabels can be awarded based on third-party certification, supplier claims of environmental conformity, or third-party validation of an environmental product declaration.
- **"Green" product lists** or online databases of preapproved green goods, works, and services can be created by governments and made available to procurers across the government.
- **Framework agreements** can be set up by central procurement authorities to include GPP approaches, making it easier for all procuring entities to purchase green without entering into difficult processes for market analysis, tender design, and bid evaluation.

a. An example of these criteria is detailed by the European Commission on the EU GPP criteria page of its website: https://ec.europa.eu/environment/gpp/eu_gpp_criteria_en.htm.

Linking Public Procurement Data to Other Dimensions of the Public Sector and Public Administration

Public procurement is multidimensional and critically interconnected with other functions of the public sector and public administration. For example, the participation of small firms in the public procurement market may be influenced by the ease of access to finance or by tax subsidies provided to certain disadvantaged firms. Similarly, the administrative burden of public procurement processes may be influenced by the staffing, training, and resources in the local procuring entities. The incentives of participants in a procurement process may be influenced by several factors. A promising area for advancement in public procurement research would be to collect and integrate data from other parts of the public sector, justice, and tax administration to create novel integrated data sets providing a holistic picture of the procurement function. This would provide governments with comprehensive information to develop innovative and impactful procurement strategies, as well as allow researchers to holistically explore the environment within which procurement is conducted.

Many potential data sets could be used to extend the analysis of public procurement through other dimensions of public administration. One example is linking tax registries and public procurement data. Data on tax filings by firms could be useful to characterize the firms operating in public procurement markets—for example, in terms of size—and the link between public procurement and the growth of firms (Ferraz, Finan, and Szerman 2015), as well as to assess the effectiveness of policies that intend to favor the participation of SMEs in public procurement.

Another potential data set is linking public procurement data with audits data. If properly designed, audits can be an effective tool to disincentivize malpractice in public procurement. However, as demonstrated by Gerardino, Litschig, and Pomeranz (2017), the design and targeting of audits can distort incentives for procurement officers. For example, procurement officers may be less likely to use competitive methods if they expect these procedures will be more likely to be audited due to their complexity, or they may be less likely to comply with regulations that are difficult for auditors to monitor, such as the application of preferential policies for SMEs or the application of green award criteria.[19]

Public procurement data can also be complemented with complaints data and judicial data. Box 12.3 discusses the potential for matching public procurement data with judicial data to validate collusion-screening algorithms. Beyond this type of application, there is also space for further research on how performance in public procurement functions is affected by the efficiencies and performance of the judicial sector. Coviello et al. (2018) have demonstrated, in the context of Italy, the implications of inefficient courts on procurement outcomes, such as longer delays in the delivery of public works, a higher likelihood that contracts are awarded to larger suppliers, and higher shares of payments postponed after delivery. Further studies on the link between public procurement and the justice sector would be necessary to advance our understanding of how these two functions of the state influence each other—for example, whether judicial investigations have an impact on processing and contract execution times, which types of procedures are more likely to result in complaints or investigations, whether the risk of complaints and appeals is a barrier to firm participation, and whether the efficiency of courts has an impact on the propensity of procuring entities to enforce late penalties.

The integration of public procurement into overall public finance management, budgeting, auditing, and service delivery processes has a high potential to lead to better utilization of public resources through better information transmission, standardization, and automation and increased accountability (OECD 2017). Despite this potential, the integration of e-procurement systems into other e-government systems is not yet a common practice. For example, based on a 2016 review of public procurement systems in OECD countries, e-GP systems are most often integrated with business registries (eight countries), tax registries (seven countries), budgeting systems (six countries), and social security databases (six countries) (OECD 2017). Data integration is an area where further work is needed to promote a whole-of-government approach from a data perspective.

Insights on Public Procurement Data from Survey Data

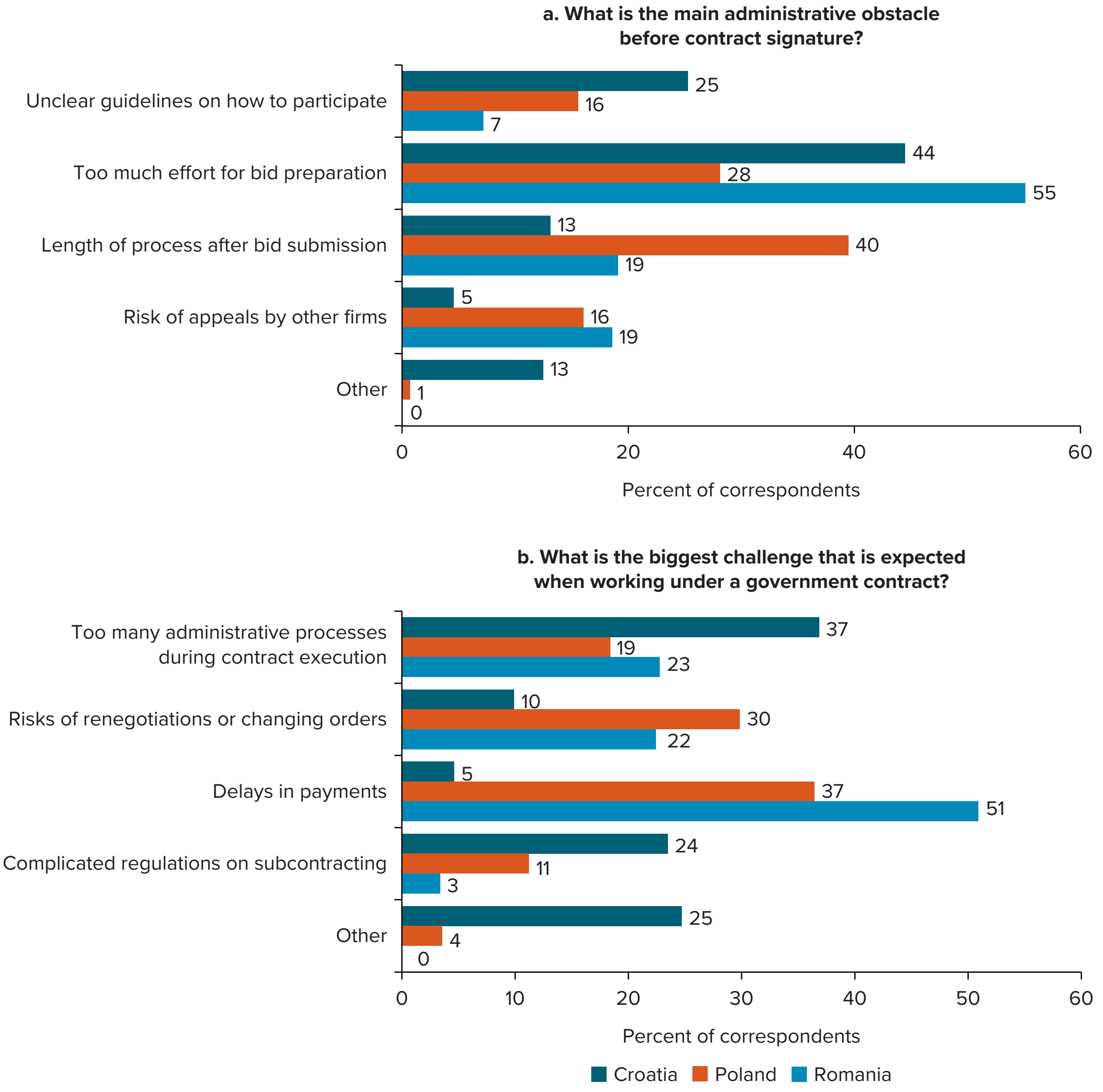
Along with using administrative data on public sector and public procurement, surveys of procuring entity officers and firms provide important context on the environment in which procurement is conducted. Surveys of procuring entity officers can be used to measure procurement-related information otherwise unavailable in the administrative data, such as time for tender preparation, contract execution quality, and firm performance. For example, in an assessment of the public procurement system in Croatia, the World Bank collected survey responses from procuring entity officers on the quality of delivered goods and services by firms and on contract management deadlines, such as the date of delivery and the final payment amount for contracts. These indicators were not available in the publicly available data in Croatia, and this data collection exercise was successful in identifying constraints during the contract management phase.

Surveys of procuring entity officers may also help measure the perceptions and behaviors of procuring entity officers with regard to overall organizational management, the administrative burden of conducting and reporting on procurement regulations, human resource management (HRM), roles, and incentives within their teams. For example, the 2021 World Bank report on the use of framework agreements relied on both administrative microdata and survey data.[20] Procuring entity officers in India and Ethiopia were surveyed about the perceived administrative burden of using framework agreements relative to other public bidding methods. While quantitative analysis revealed some savings in price through the use of framework agreements, the survey provided more context about the burden officers might feel when implementing different types of procurement methods. Similarly, several studies on GPP have been conducted through surveys to understand the incentives of procuring entity officers to adopt GPP criteria in the award process, as well as to map the difficulties and challenges entities face with GPP regulations at different stages of the procurement process.

In addition to surveying procuring entity officers, reaching out to firms participating in the public contract market can also provide complementary information for understanding public procurement from the perspective of private sector actors. For example, Uyarra et al. (2014) find that for firms in the United Kingdom, the main barriers to entry into the public procurement market are a lack of interaction with procuring organizations, the low competency of civil servants, and poor management systems, and Knack, Biletska, and Kacker (2019) find that firms are more likely to participate in public procurement in countries where public procurement systems are more transparent and complaint systems are more effective. Surveys of firms can also be a useful tool to analyze special groups of firms (for example, SMEs and WOEs) in countries where public procurement data do not allow for the identification of bidder and supplier characteristics and where procurement data cannot be linked to other administrative data, such as firm registries. In 2021, the World Bank designed a procurement module as part of the Enterprise Surveys to better understand the barriers and challenges experienced by firms with respect to public procurement, and they piloted this module in Romania, Poland, and Croatia. The survey data reveal that the main administrative obstacle to participation in Poland is the length of the process between bid submission and contract signature, while in Croatia and Romania, it is the fact that too much effort is required for bid preparation (figure 12.7a). The biggest challenge when working under a government contract is payment delays in Poland and Romania and the number of administrative processes during contract execution in Croatia (figure 12.7b).

Using survey data in public procurement is relevant from a policy perspective in order to complement administrative data measurements, but it is also relevant from a research point of view. HRM practices,

FIGURE 12.7 Obstacles and Challenges to Government Contracts, Croatia, Poland, and Romania, 2021



Source: Original figure for this publication based on microdata from the World Bank Enterprise Surveys Follow-Up on COVID-19 2021, Round 3, for Croatia, Poland, and Romania.
Note: Panel a: Weighted results. Only firms that indicated that administrative procedures before contract signature are an obstacle to attempting to secure a government contract. Panel b: Weighted results. Only firms that indicated that expected challenges during contract execution are an obstacle to attempting to secure a government contract.

attitudes, and motivations in public administration are typically measured through surveys of civil servants. Public procurement can be an ideal area to study the link between these dimensions and outcomes, advancing our understanding of the impact of HRM practices, attitudes, and motivations on performance and compliance.

CONCLUSION

This chapter has provided an overview of how public procurement data can be used for monitoring and evaluating public procurement, as well as for informing reform efforts and defining new policies and strategies in public procurement. It has included a description of various data analytical tools that can be applied to public procurement, an account of typical challenges encountered in public procurement data and potential solutions, and a discussion of recent innovations, such as the development of interactive dashboards.

The chapter has included various lessons for practitioners and governments on using and analyzing public procurement administrative data, including centralizing public procurement data, integrating data from different procurement stages and from data systems related to other government functions, ensuring data quality and completeness, and building capacity for statistical analysis, such as by creating a dedicated statistical unit in the public procurement authority.

The chapter has also highlighted various areas where there is a need for further development and research, specifically in measuring the quality of contract implementation, integrating public procurement data with other administrative microdata or survey data, measuring GPP, and, more generally, generating robust empirical evidence on effective ways to improve the efficiency, integrity, inclusiveness, and sustainability of public procurement. For example, the World Bank's Governance Global Practice and the Development Impact Evaluation (DIME) Governance and Institution Building unit have been collaborating on a research agenda about the link between public procurement and private sector growth, which includes a series about research projects and data innovations, such as connecting public procurement data, payment data, and tax registry data.[21]

NOTES

The chapter is based on academic research and operational experience from several World Bank projects that use data analytical tools in the area of public procurement—for example, in Romania (led by Carmen Calin, procurement specialist), Croatia (led by Antonia Viyachka, procurement specialist), and Bangladesh (led by Ishtiak Siddique, senior procurement specialist). The chapter greatly benefited from comments and inputs by Carmen Calin (World Bank, procurement specialist), Maria Arnald Canudo (consultant, Development Impact Evaluation [DIME] Department), Daniel Rogger (senior economist, DIME), and Christian Schuster (professor, University College London). Stephen Shisoka Okiya (consultant, DIME) provided excellent research assistance.

1. A seminal paper by Bandiera, Prat, and Valletti (2009) demonstrates that in Italy, 83 percent of the total estimated waste in public procurement is due to passive waste caused by inefficiencies related to constraints such as lack of skills, lack of incentives, and excessive regulatory burden.
2. For example, with respect to the United Nations Sustainable Development Goals, public procurement can contribute to increasing access to markets for small and medium enterprises (target 9.3), responsible consumption and production through sustainable public procurement (target 12.7), reducing corruption and bribery (target 16.5), developing effective, accountable, and transparent institutions (target 16.6), and ensuring public access to information (target 16.10). More information about the Sustainable Development Goals is available on the United Nations Commission on International Trade Law website at https://uncitral.un.org/en/about/sdg.
3. Further details on the data in figure 12.1 and on the level of e-GP adoption across countries can be found in the World Bank's *Doing Business 2020* data under the topic "Contracting with the Government": https://archive.doingbusiness.org/en/data/exploretopics/contracting-with-the-government.

4. Requirements from international organizations or international treaties could be another strategy to incentivize governments to open public procurement data and adopt transparent monitoring and reporting mechanisms. For example, EU member states are mandated to monitor and report key procurement indicators under Directives 2014/23/EU, 2014/24/EU, and 2014/25/EU.
5. The literature has pointed to three different strategies for measurement validity (Adcock and Collier 2001): content validity (the measurement captures the full content of the definition), convergent validity (alternative measures of the same corrupt phenomenon are correlated), and construct validity (well-established empirical relationships are confirmed by the measurement).
6. As noted above, EU member states are mandated to monitor and report key procurement indicators under Directives 2014/23/EU, 2014/24/EU and 2014/25/EU.
7. The PPO website is available at https://www.uzp.gov.pl/.
8. The Public Procurement Agency in Bulgaria, the PPO in Poland, the Office for Public Procurement in the Slovak Republic, and the National Agency for Public Procurement (ANAP) in Romania are examples of institutions that conduct audits of compliance and performance monitoring.
9. Deliverable under the World Bank project Framework Agreements for Development Impact: Lessons from Selected Countries for Global Adoption (P173392).
10. Report under the project Impact Evaluation of e-Procurement In Bangladesh (P156394).
11. The role of civil society in monitoring public procurement is widely recognized. For example, within the EU project Integrity Pacts—Civil Control Mechanism for Safeguarding EU Funds, "integrity pacts" are established between a contracting authority and economic operators bidding for public contracts, stipulating that parties will abstain from corrupt practices and conduct a transparent procurement process, and a separate contract with a civil society organization entrusts it with the role of monitoring that all parties comply with their commitments. See the Transparency International website at https://www.transparency.org/en/projects/integritypacts.
12. More information about the Government Transparency Institute is available on its website, http://www.govtransparency.eu/. The dashboard prototype is available here: https://www.procurementintegrity.org/.
13. The Opentender dashboard is available here: https://opentender.eu/start.
14. For more information about the Open Contracting Data Standard, see the project website at https://standard.open-contracting.org/latest/en/.
15. The India dashboard is available here: https://eprocure.gov.in/eprocdashboard/KPI.html.
16. The Ukraine dashboard is available here: https://bi.prozorro.org/hub/stream/aaec8d41-5201-43ab-809f-3063750dfafd.
17. On Brazil, see CGU (2020). Moldova's COVID-19 procurement website can be viewed here: https://www.tender.health/. Lithuania's procurement webpage can be viewed on the Public Procurement Office website at https://vpt.lrv.lt/kovai-su-covid-19-sudarytos-sutartys. South Africa's COVID-19 procurement dashboard can be viewed on the National Treasury website at http://ocpo.treasury.gov.za/COVID19/Pages/Reporting-Dashboard-Covid.aspx.
18. Green evaluation criteria can be included in different levels of procurement and in the bidding process by setting technical specifications, specific qualifications, contract requirements, selection criteria, and/or award criteria (Testa et al. 2012).
19. As an example of the former, Gerardino, Litschig, and Pomeranz (2017) investigate the impact of the audit selection process in Chile, using public procurement data from 2011 to 2012. Under the existing audit protocol in that period, open auctions underwent more than twice as many checks as direct contracting. The authors find that, given this protocol, procurement officers shifted toward direct contracting methods and reduced the use of open auctions, especially in procuring entities that experienced more audits and therefore had more opportunities to learn about this targeting design. As an example of the latter, in some countries, procuring entities are required to reserve a given quote of their spending for SMEs, but it is challenging for auditors to monitor compliance with this requirement if public procurement data do not include a tag to identify contracts awarded to SMEs.
20. Deliverable under the World Bank project Framework Agreements for Development Impact: Lessons from Selected Countries for Global Adoption (P173392).
21. See the World Bank project Public Procurement and Firm Behavior (P177551).

REFERENCES

Adam, Isabelle, and Mihály Fazekas. 2019. "Big Data Analytics as a Tool for Auditors to Identify and Prevent Fraud and Corruption in Public Procurement." *European Court of Auditors Journal* 2: 172–80. https://medium.com/ecajournal/big-data-analytics-as-a-tool-for-auditors-to-identify-and-prevent-fraud-and-corruption-in-public-68184529334c.

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46. https://doi.org/10.1017/S0003055401003100.

Bandiera, Oriana, Andrea Prat, and Tommaso Valletti. 2009. "Active and Passive Waste in Government Spending: Evidence from a Policy Experiment." *American Economic Review* 99 (4): 1278–308. https://doi.org/10.1257/aer.99.4.1278.

Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2019. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w23350.

Bosio, Erica, Simeon Djankov, Edward L. Glaeser, and Andrei Shleifer. 2022. "Public Procurement in Law and Practice." *American Economic Review* 112 (4): 1091–117. https://doi.org/10.1257/aer.20200738.

CGU (Controladoria-Geral da União). 2020. "CGU lança painel para dar transparência a contratações relacionadas à Covid-19." Comptroller General of Brazil, March 7, 2020. https://www.gov.br/cgu/pt-br/assuntos/noticias/2020/07/cgu-lanca-painel-para-dar-transparencia-a-contratacoes-relacionadas-a-covid-19.

Conley, Timothy G., and Francesco Decarolis. 2016. "Detecting Bidders Groups in Collusive Auctions." *American Economic Journal: Microeconomics* 8 (2): 1–38. https://doi.org/10.1257/mic.20130254.

Coviello, Decio, Luigi Moretti, Giancarlo Spagnolo, and Paola Valbonesi. 2018. "Court Efficiency and Procurement Performance." *The Scandinavian Journal of Economics* 120 (3): 826–58. https://doi.org/10.1111/sjoe.12225.

Decarolis, Francesco, Raymond Fisman, Paolo Pinotti, and Silvia Vannutelli. 2020. "Rules, Discretion, and Corruption in Procurement: Evidence from Italian Government Contracting." NBER Working Paper 28209, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w28209.

Decarolis, Francesco, and Cristina Giorgiantonio. 2022. "Corruption Red Flags in Public Procurement: New Evidence from Italian Calls for Tenders." *EPJ Data Science* 11: 16. https://doi.org/10.1140/epjds/s13688-022-00325-x.

European Commission. 2008. *Public Procurement for a Better Environment*. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, COM(2008) 400. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0400:FIN:EN:PDF.

Fazekas, Mihály, Luciana Cingolani, and Bence Tóth. 2018. "Innovations in Objectively Measuring Corruption in Public Procurement." Chap. 7 in *Governance Indicators: Approaches, Progress, Promise*, edited by Helmut K. Anheier, Matthias Haber, and Mark A. Kayser. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198817062.003.0007.

Fazekas, Mihály, and Gábor Kocsis. 2020. "Uncovering High-Level Corruption: Cross-National Corruption Proxies Using Public Procurement Data." *British Journal of Political Science* 50 (1): 155–64. https://doi.org/10.1017/S0007123417000461.

Fazekas, Mihály, Salvatore Sberna, and Alberto Vannucci. 2021. "The Extra-Legal Governance of Corruption: Tracing the Organization of Corruption in Public Procurement." *Governance: An International Journal of Policy, Administration, and Institutions* 35 (4): 1139–61. https://doi.org/10.1111/gove.12648.

Fazekas, Mihály, and Bence Tóth. 2018. "The Extent and Cost of Corruption in Transport Infrastructure: New Evidence from Europe." *Transportation Research Part A: Policy and Practice* 113: 35–54. https://doi.org/10.1016/j.tra.2018.03.021.

Fazekas, Mihály, Gavin Ugale, and Angelina Zhao. 2019. *Analytics for Integrity. Data-Driven Approaches for Enhancing Corruption and Fraud Risk Assessments.* Paris: OECD Publishing. https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf.

Ferraz, Claudio, Frederico Finan, and Dimitri Szerman. 2015. "Procuring Firm Growth: The Effects of Government Purchases on Firm Dynamics." NBER Working Paper 21219, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w21219.

Gerardino, Maria Paula, Stephan Litschig, and Dina Pomeranz. 2017. "Distortion by Audit: Evidence from Public Procurement." NBER Working Paper 23978, National Bureau of Economic Research, Cambridge, MA. Revised August 2022. https://doi.org/10.3386/w23978.

Hassan, Mirza. 2017. "Citizen Engagement during Public Procurement Implementation in Bangladesh." South Asia Procurement Innovation Awards 2016, World Bank, Washington, DC. https://wbnpf.procurementinet.org/featured/citizen-engagement-during-public-procurement-implementation-bangladesh.

Huber, Martin, and David Imhof. 2019. "Machine Learning with Screens for Detecting Bid-Rigging Cartels." *International Journal of Industrial Organization* 65: 277–301. https://doi.org/10.1016/j.ijindorg.2019.04.002.

Knack, Stephen, Nataliya Biletska, and Kanishka Kacker. 2019. "Deterring Kickbacks and Encouraging Entry in Public Procurement Markets: Evidence from Firm Surveys in 90 Developing Countries." *World Bank Economic Review* 33 (2): 287–309. http://hdl.handle.net/10986/34863.

Krasnokutskaya, Elena, and Katja Seim. 2011. "Bid Preference Programs and Participation in Highway Procurement Auctions." *American Economic Review* 101 (6): 2653–86. https://doi.org/10.1257/aer.101.6.2653.

Marion, Justin. 2007. "Are Bid Preferences Benign? The Effect of Small Business Subsidies in Highway Procurement Auctions." *Journal of Public Economics* 91 (7–8): 1591–624. https://doi.org/10.1016/j.jpubeco.2006.12.005.

Medvedev, Denis, Ramin N. Aliyev, Miriam Bruhn, Paulo Guilherme Correa, Rodrigo Javier Garcia Ayala, Justin Piers William Hill, Subika Farazi, Jose Ernesto Lopez Cordova, Caio Piza, Alena Sakhonchik, and Morten Seja. 2021. *Strengthening World Bank SME-Support Interventions: Operational Guidance Document*. World Bank Report. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/183521617692963003/Strengthening-World-Bank-SME-Support-Interventions-Operational-Guidance-Document.

Mungiu-Pippidi, Alina, and Mihály Fazekas. 2020. "How to Define and Measure Corruption." In *A Research Agenda for Studies of Corruption*, edited by Alina Mungiu-Pippidi and Paul M. Heywood, 7–26. Cheltenham, UK: Edward Elgar. https://doi.org/10.4337/9781789905007.00008.

Nissinen, Ari, Katriina Parikka-Alhola, and Hannu Rita. 2009. "Environmental Criteria in the Public Purchases above the EU Threshold Values by Three Nordic Countries: 2003 and 2005." *Ecological Economics* 68 (6): 1838–49. https://doi.org/10.1016/j.ecolecon.2008.12.005.

OECD (Organisation for Economic Co-operation and Development). 2013. "Ex Officio Cartel Investigations and the Use of Screens to Detect Cartels." Competition Policy Roundtables DAF/COMP(2013)27, Competition Committee, Directorate for Financial and Enterprise Affairs, OECD, Paris. https://www.oecd.org/daf/competition/exofficio-cartel-investigation-2013.pdf.

OECD (Organisation for Economic Co-operation and Development). 2017. *Government at a Glance 2017.* Paris: OECD Publishing. https://doi.org/10.1787/gov_glance-2017-en.

OECD (Organisation for Economic Co-operation and Development). 2018. *SMEs in Public Procurement: Practices and Strategies for Shared Benefits*. OECD Public Governance Reviews. Paris: OECD Publishing. https://doi.org/10.1787/9789264307476-en.

OECD (Organisation for Economic Co-operation and Development). 2021. *Government at a Glance 2021.* Paris: OECD Publishing. https://doi.org/10.1787/1c258f55-en.

OECD and SIGMA (Support for Improvement in Governance and Management). 2019. *Methodological Framework of the Principles of Public Administration.* Paris: OECD Publishing. https://www.sigmaweb.org/publications/Methodological-Framework-for-the-Principles-of-Public-Administration-May-2019.pdf.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49. https://doi.org/10.1086/517935.

Palmujoki, Antti, Katriina Parikka-Alhola, and Ari Ekroos. 2010. "Green Public Procurement: Analysis on the Use of Environmental Criteria in Contracts." *Review of European Community & International Environmental Law* 19 (2): 250–62. https://doi.org/10.1111/j.1467-9388.2010.00681.x.

Singer, Marcos, Garo Konstantinidis, Eduardo Roubik, and Eduardo Beffermann. 2009. "Does e-Procurement Save the State Money?" *Journal of Public Procurement* 9 (1): 58–78. https://doi.org/10.1108/JOPP-09-01-2009-B002.

Testa, Francesco, Fabio Iraldo, Marco Frey, and Tiberio Daddi. 2012. "What Factors Influence the Uptake of GPP (Green Public Procurement) Practices? New Evidence from an Italian Survey." *Ecological Economics* 82: 88–96. https://doi.org/10.1016/j.ecolecon.2012.07.011.

Titl, Vitezslav, and Benny Geys. 2019. "Political Donations and the Allocation of Public Procurement Contracts." *European Economic Review* 111: 443–58. https://doi.org/10.1016/j.euroecorev.2018.11.004.

Turkewitz, Joel, Mihály Fazekas, and Zafrul Islam. 2020. "Case Study 2: e-Procurement Reform in Bangladesh." In *Enhancing Government Effectiveness and Transparency: The Fight against Corruption*, edited by Rajni Bajpai and C. Bernard Myers, 34–39. World Bank Global Report. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/235541600116631094/Enhancing-Government-Effectiveness-and-Transparency-The-Fight-Against-Corruption.

Uyarra, Elvira, Jakob Edler, Javier Garcia-Estevez, Luke Georghiou, and Jillian Yeow. 2014. "Barriers to Innovation through Public Procurement: A Supplier Perspective." *Technovation* 34 (10): 631–45. https://doi.org/10.1016/j.technovation.2014.04.003.

Wachs, Johannes, Mihály Fazekas, and János Kertész. 2021. "Corruption Risk in Contracting Markets: A Network Science Perspective." *International Journal of Data Science and Analytics* 12: 45–60. https://doi.org/10.1007/s41060-019-00204-1.

World Bank. 2017. *Pakistan—Punjab Land Records Management and Information Systems Project.* ICR00003719. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/632241498842804246/Pakistan-Land-Records-Management-and-Information-Systems-Project.

World Bank. 2019. *Romania—Reimbursable Advisory Services Agreement on Assessment of the Public Procurement System and Further Support to the Implementation of the Public Procurement Strategy: Output 4: Final Version of the Web-Based Guide*. P169141. Washington, DC: World Bank. https://pubdocs.worldbank.org/en/412981574427978384/RO-TOR-Procurement-SME-2019.pdf.

World Bank. 2020. *Assessment of Bangladesh Public Procurement System*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/33882.

World Bank. 2021a. *Econometric Analysis of Framework Agreements in Brazil and Colombia*. Washington, DC: World Bank. https://doi.org/10.1596/36059.

World Bank. 2021b. "Green Public Procurement: An Overview of Green Reforms in Country Procurement Systems." Climate Governance Papers, World Bank, Washington, DC. http://hdl.handle.net/10986/36508.

CHAPTER 13

Government Analytics Using Data on the Quality of Administrative Processes

Jane Adjabeng, Eugenia Adomako-Gyasi, Moses Akrofi, Maxwell Ampofo, Margherita Fornasari, Ignatius Geegbae, Allan Kasapa, Jennifer Ljungqvist, Wilson Metronao Amevor, Felix Nyarko Ampong, Josiah Okyere Gyimah, Daniel Rogger, Nicholas Sampah, and Martin Williams

SUMMARY

This chapter seeks to highlight the value of quantifiable measures of the quality of back-office processes when assessing governments' bureaucratic effectiveness. Conceptually, it defines a framework for understanding administrative *process productivity*. It then presents case studies from the Ghanaian and Liberian civil services, where different measures of internal (within bureaucratic units) versus external (across bureaucratic units) process quality were piloted. Specifically, these pilots sought to assess the feasibility, cost, and scalability of the process measures considered. We explore their correlations with other measures of productivity (for example, financial expenditures and the completion of planned tasks) and the claims and characteristics of civil servants in the surveys we have undertaken.

ANALYTICS IN PRACTICE

- Many of the activities undertaken by public administrators can be characterized as the application of proper processes and procedures to a project, file, or case. As such, the quality of the processing work undertaken by public officials is an important part of the quality of government activity and a determinant of public sector productivity.

Jane Adjabeng, Eugenia Adomako-Gyasi, Moses Akrofi, Maxwell Ampofo, Allan Kasapa, Wilson Metronao Amevor, Felix Nyarko Ampong, Josiah Okyere Gyimah, and Nicholas Sampah are with the Office of the Head of the Civil Service, Ghana. Margherita Fornasari, Jennifer Ljungqvist, and Daniel Rogger (corresponding author) are with the World Bank's Development Impact Evaluation (DIME) Department. Ignatius Geegbae is with the Civil Service Agency, Liberia. Martin Williams is an associate professor in public management at University of Oxford.

- Government analytics systems should include measures of the quality of processing by government officials. This requires a framework for assessing the adherence of administrators' process work to accepted government procedure. An important distinction in the development of a conceptual framework for assessing process quality is whether a process is mainly confined within organizational units (what this chapter calls *internal process productivity*) or is associated with interactions between organizational units (*external process productivity*). Separately, such a framework could be formulated to assess the quality of processes in general within a public service or targeted at domain-specific activities, such as the implementation of an appraisal process.
- By not measuring the quality of process productivity, analytics systems bias measures of the quality of government toward more measurable aspects of government work, such as the activities of frontline officials, and away from the important body of administrative professionals who support them. By taking a narrow approach to defining bureaucratic productivity according to frontline outputs, studies risk missing why a service may be delivered well or poorly.
- Such data can be collected automatically, as part of digitized government work, or manually, by assessors employed to judge process quality in the physical records of projects, files, or cases. Analysts can benchmark the quality of processes evidenced by relevant records in terms of the extent to which they are well organized, logical in flow, and adherent to government procedure; their timeliness with respect to a deadline; or whatever aspect of a process is of interest. To some degree, the digitization of government has supported the improvement of processes—by ensuring that all components of a process are present—but it also facilitates the physical or automated inspection of process quality.
- Measures of process quality in public administration open up three areas of government analytics: assessments of variation in the quality of processes and associated productivity within and across organizations in the same country, comparisons of process quality across countries, and assessments of public sector process quality over time. In this way, government analysts can pinpoint where government procedure is not being adhered to, how different processes relate to public sector productivity, and what dynamics exist across individuals and organizational units.

INTRODUCTION

The effective delivery of government work requires a long chain of processes undertaken by public officials. From policy development, through budget and planning, to monitoring and evaluation, many of the activities undertaken by public administrators can be characterized as the application of proper processes and procedures to a project, file, or case. This is the back-office administration of public policy. In the case of policy development, for example, proper process in most modern governments would include presenting a balance of evidence on the pros and cons of policy content, ensuring broad-based consultation within (and potentially outside) government, and generating a coherent policy that others in government could follow.

The quality of the processing work undertaken by public officials is thus an important part of the quality of government activity. Given how important these processes are in the chain of delivery for government goods and services—such as the budget and procurement processes—process quality is a key component of public sector productivity.

This chapter articulates a framework for measuring the quality of these processes based on the idea of adherence to accepted government procedure. The rationale of adherence to government procedure may be varied: it may include equity considerations (ensuring all cases are dealt with in a similar way), fiduciary concerns (ensuring resources are utilized for the public good), and legal issues (ensuring that actions are in line with existing laws and the rules of the public service). The quality of government processes can be measured along these and other margins as a measure of the nature of government functioning. This chapter outlines specific measures of the quality of government processes and discusses their use through two case studies in Ghana and Liberia.

To date, measures of bureaucratic activity and effectiveness have focused on frontline outputs, as described in chapter 29 of the *Handbook* on service delivery indicators (SDI), financial expenditures or the procurement of goods, as described in the *Handbook* chapters on budget and procurement (chapter 12), and the provision of physical infrastructure, as described in the *Handbook* chapter on task completion (chapter 17). This chapter focuses not on the prices, wages, or services achieved by the government but on the quality of the processes applied toward those ends. For example, a procurement officer may gain low prices for a set of goods procured but do so in a way that breaks public service rules and potentially exposes the public sector to unnecessary reputational risks. Similarly, an officer may complete all assigned tasks within deadline and budget but do so in a way that has negative spillovers on other units in the agency.

This approach to understanding the quality of work in the public sector has parallels to aspects of the SDI outlined in chapter 29. For example, in assessing the quality of education, analysts have assessed the extent to which teachers are subject to classroom observation by an independent assessor and are provided feedback on their teaching. This measure does not score the quality of the teaching itself, or even the quality of the feedback. Rather, it assesses the extent to which a process is in place to provide feedback, assuming that feedback is an important part of a quality teaching environment.[1]

Perhaps the closest approach to an assessment of bureaucratic functioning that attempts to measure the quality of government processes is the approach associated with case analysis. As described in chapter 15 on administrative case data, such analysis assesses the quality of responses by public officials to requests for public services, such as in the social security sector, or the fulfillment of public responsibilities, such as the collection of taxes. However, the data collected are almost universally on outcomes of these activities, such as the volume of cases processed in a particular time frame or, conversely, processing speed, prices paid, and so on. A complement to this analysis characterizes the quality of public sector actors' work processes, from the comprehensiveness of records to the quality of the evidence they provide to back up their assertions. So while measures like those in chapter 15 typically characterize the speed of case completion as a positive outcome, a *process productivity* perspective would assess whether sufficient time had been allotted for consultation (such as for advertising a procurement). Little quantitative work has been undertaken on this margin of government activity.[2]

This paucity of preexisting work stands in contrast to the fact that a substantial portion of the work of public administration is best characterized as processing. Rasul, Rogger, and Williams (2021) find that 73 percent of civil service activities in Ghana can be categorized as "processing tasks." The common conception of government work is frequently back-office process work.

The absence of effective measures of the quality of government processes has skewed the focus of public sector studies toward frontline officials and away from the important body of administrative professionals that support them.

Most civil servants play an important role in facilitating the role of frontline staff, by providing the long chain of supporting activities that are at the core of the effectiveness of government. Processing work is a substantial component of this support.

By taking the narrow approach of defining bureaucratic productivity according to frontline outputs, studies also risk missing why a service may be delivered well or poorly. For example, for a citizen to receive a welfare payment, budgetary officers must ensure sufficient funds are available, contracting officers must ensure effective transfer systems to recipients, and accounting officers must ensure a clear paper trail to reduce the diversion of funds. Wrapping the entirety of these activities into a single indicator of payment disbursement does not allow us to uncover which process creates a bottleneck.

Consequently, this chapter seeks to highlight the value of quantifiable measures of this type of back-office, administrative process productivity when assessing governments' bureaucratic effectiveness. It does so by presenting case studies from the Ghanaian and Liberian civil services, where different measures of *internal* (within bureaucratic units) versus *external* (across bureaucratic units) process quality were piloted.

And it considers both the quality of standard work processes (Ghana) and the implementation of a new set of processes related to staff appraisals (Liberia). These pilots introduce concrete ideas for measuring process quality and showcase their feasibility and scalability to entire public services.

Measures of process quality in public administration open up three areas of government analytics. First, we can use such measures to assess variations in process quality and associated productivity within and across organizations in the same country. For example, by using a common assessment of process quality across organizations, we can identify which organizations are appropriately adhering to government procedure across a government. Second, with appropriate caveats, common measures enable comparisons of process productivity across countries. For example, understanding the time it takes for a social sector ministry to provide inputs to the center of government across countries provides microevidence of the relative quality of governance. Finally, given the relative simplicity of these measures, we could collect productivity data on a regular basis and thus provide a more nuanced assessment of public sector capacity over time.

This chapter continues as follows. It begins with an overview of related measures and then presents concrete applications of these ideas in case studies from Ghana and Liberia. It then showcases the results of measurement in these two settings and discusses what we learn about the nature of process quality in the public service.

CONCEPTUAL FRAMEWORK AND RELATED LITERATURE

Conceptually, the notion that government processes should adhere to particular standards is widespread. Most governments have rules for undertaking (or *processing*) the tasks of public administration that articulate best practices. These best practices almost universally align with themes of completeness, rationality, fairness, and efficiency—all themes extolled by the Weberian school of public administration.

Wilson (1989, 26) argues that understanding public sector productivity, in contrast to that of the private sector, means understanding the processing of tasks. Since the goals of the public sector are too vague to be a useful organizing framework, the public sector must focus on specializing in improved task or process productivity. This reasoning has since been bolstered by a range of authors (for example, Alesina and Tabellini 2007; Dixit 2002).[3]

However, few analysts argue for a coherent notion of government processes as a component of government functioning or of public sector productivity. Yet if government processes mediate the use of inputs to the production function of government, then undertaking them to a high standard would seem to be an output of government work related to functioning and productivity. In relation to ideals of the state, such as the equitable treatment of cases, process may be an end in itself, observed in the capacity of public officials to make coherent decisions.

When public officials make coherent arguments for choosing one policy over another that incorporate relevant information, they improve the quality of government outputs but also characterize government itself. Both of these are public goods of their own type. Similarly, when a manager judges one official eligible for promotion over another using solid evidence of the performance of both officials, the public sector becomes more effective and is characterized as meritocratic. Again, both of these are public goods in distinct ways.

For this reason, whether public officials process government work in the appropriate ways can be studied as a form of public sector productivity: *process productivity*. When the government effectively and efficiently undertakes work according to proper processes, it generates better outputs for the next stage of government work and defines a superior character of government. It is therefore more productive in producing these public goods. Take, for example, a firm that creates parts to sell to other firms that build machines out of those parts: when it does this with a high level of quality and in a reliable way, the parts firm is productive. Likewise, a government organization that undertakes its tasks using proper processes is a more productive institution.

What proper process means will vary by setting and the tasks focused on. However, best practices in government processes often include the clear and complete gathering of evidence and rational decision-making, as well as equity considerations (ensuring all cases are dealt with in a similar way), fiduciary concerns (ensuring resources are utilized for the public good), and legal issues (ensuring that actions are in line with existing laws and the rules of the public service). An example of an approach to assessing the quality of decision-making is the SMART framework, which considers whether relevant elements are specific, measurable, achievable, relevant, and time-bound. This framework will be applied in one of our case studies.

This chapter assesses how an analyst might measure process quality, and thus process productivity, on a large scale (across a substantial portion of units of public administration) using a quantitative approach. Efforts to date to characterize government and its processes have been broad-brush, such as expert assessments of corruption that outline the propensity to circumvent proper processes for personal gain across an administration as a whole.[4] Our focus is on measurement at a granular level, frequently the task, project, or individual level.

This more granular level is the area of measurement for which there is little to no previous work and, as argued in much of the rest of the *Handbook*, where there is the most potential for gains from analysis. For a similar reason, we look for processes that are generally applied across government, rather than a domain-specific set of processes, such as how doctors should treat patients (Bedoya et al. 2017; Daniels et al. 2017; Wafula et al. 2017). However, to provide clarity in the application of our framework to domain-specific settings, our second case study looks at the application of process productivity assessments to a domain-specific activity undertaken by all public servants—performance appraisal.

Empirical assessments of government processes in political science have studied the nature of responses to public information requests (also known as freedom of information requests). By assessing the qualities of government responses, researchers have assessed whether citizens receive a response quickly (Wehner and Poole 2015; Wood and Lewis 2017) and equitably (Berliner et al. 2021; Peisakhin and Pinto 2010). This approach is clearly highly constrained in what it can measure as an external lens with ambiguous links to government functioning.

In the economics literature, Chong et al. (2014) assess the quality of government processes through how quickly misaddressed letters are returned to their original senders. This measure is unrelated to most aspects of government work but can be seen as similar in spirit to the measure we will introduce in this chapter to assess government productivity through how responsive units are to centralized requests for information.

The closest paper to measuring internal process quality in a large-scale, quantitative way is Banerjee et al. (2021), who use retired senior police officers to grade a random set of case files from project police stations. They grade officers on whether scientific techniques were used, the care with which evidence was collected, and so on. Though their focus is explicitly on the clarity of police processes, the approach we elaborate in this chapter is a generalization of their approach.

We distinguish *internal process productivity*, the quality of administrative processes for activities confined within a particular administrative unit, from *external process productivity*, the quality of administrative processes for activities in which units interact. An example of the first is the development of the design of a project in which a unit specializes, while an example of the second is a request for information from one unit by another.

We make this distinction because accountability and professional dynamics vary distinctly between the two cases. Public administration is typically conceptualized around work units organized within a hierarchy. These work units have a degree of flexibility in how they organize their approaches to undertaking government work and implementing process guidelines. However, the head of a unit is responsible for ensuring process quality, as only the head administrator of an organization ensures the organization as a whole adheres to processes. An analogous assessment can be made between organizations and the government as a whole.

Similarly, when communicating within organizational units, different record-keeping formats are required than when communicating between organizations. For this reason, the nature of measurement must vary when analysts are assessing internal versus external conceptions of process productivity.

EMPIRICAL CASE STUDIES

We study the quality of bureaucratic processes in the public administrations of two West African countries: Ghana and Liberia. These are excellent environments for testing new measures of public service productivity. They are governed by clearly defined and well-structured rules for undertaking government processes. However, similar to many developing countries, the productivity of departments and organizations in these settings varies substantially (Rasul and Rogger 2018; Rasul, Rogger, and Williams 2021). There is mounting evidence that this variation in productivity is also prevalent in the public sectors of wealthier nations (Best, Hjort, and Szakonyi 2017; Fenizia 2022), but the variation we analyze likely subsumes this heterogeneity and is representative of a large portion of the world's public administrations.

In Ghana, we study a representative set of administrative tasks under-taken by the core public administration. In Liberia we focus on process quality in relation to a specific administrative activity: the implementation of a staff appraisal system. We split our efforts into understanding the quality of internal processes, by assessing whether the processing of these tasks adheres to government procedures, and external processes, by assessing how promptly units respond to requests from central agencies. Our discussion of the two case studies thus covers the main features of process productivity outlined in the previous section.

Institutional Background

Ghana is a lower-middle-income country home to 28 million people, with a central government bureaucracy that is structured along lines reflecting both its British colonial origins and more presidentialist postindependence reforms. Ghana is one of Africa's most democratic countries.

Liberia is a low-income country of nearly 5 million people, with an agency-based administration similar in design to that of the United States. Years of civil war exacerbated recruitment and rewards based on patronage in the service. The resulting bloated workforce, a lack of established processes and procedures—or the presence of overly bureaucratic processes and procedures—and inadequate office resources have delayed and derailed the processing time for needed administrative procedures in the service. Furthermore, while Liberia is Africa's oldest and first modern republic, with a political system heavily influenced by the US Constitution, it has historically been largely characterized by minoritarianism. Democratic and recognized fair elections only commenced in the 21st century. Ghana and Liberia thus represent polities at two ends of Sub-Saharan Africa's distribution of state fragility.

Ghana's civil service consists of 57 central government ministries and departments that primarily perform the core bureaucratic functions of policy making, administration, and service delivery oversight. Ministries and departments are overseen by the Office of the Head of Civil Service (OHCS), which is responsible for personnel management and performance within the civil service. The OHCS coordinates and decides on all hiring, promotion, transfer, and (in rare circumstances) firing of bureaucrats across the service. Similarly, Liberia's Civil Service Agency (CSA) oversees the strategic leadership and management of the country's civil service, formulating and providing guidance on recruitment, personnel management, standards, and performance in civil service institutions. The Liberian service is made up of 31 ministries and agencies, in addition to the country's numerous public autonomous organizations. The architecture of the administration in the two countries has many commonalities.

Processes under Study

The civil servants we study carry out public administration activities following administrative procedures, which set out guidelines and standards for how to formally proceed with government business. These apply equally across the service and broadly aim to ensure transparency, equity, and efficiency in government business. In both Ghana and Liberia, we seek to assess the efficiency with which civil servants undertake administrative processes. However, the specific processes under study differ.

In Ghana, we focus on an assessment of process quality in core office duties, such as project planning, budgeting, and monitoring. Rasul, Rogger, and Williams (2021) describe the most common types of tasks in Ghana's central government offices. These relate to processing paperwork related to the construction of public infrastructure, such as roads, boreholes, and schools (24 percent of tasks); administrative advocacy (16 percent); and monitoring, review, and auditing (14 percent). The OHCS outlines rules and associated guidelines for Ghanaian civil servants about how to prepare infrastructure or advocacy projects and monitor, review, and audit them according to proper procedures. For this reason, the Ghanaian civil service is characterized by a common set of standards and centrally managed procedures that officials are required to follow when handling administrative files (PRAAD 2007).[5]

In Liberia, we focus on adherence to new processes for performance assessment or "appraisal." Following the end of Liberia's civil war in 2003, the CSA focused on establishing a more meritocratic civil service by, among other policies, developing a performance management system (PMS) policy (CSA and USAID-GEMS 2016; Forte 2010; Friedman 2012; World Bank 2014). Job descriptions were only recently formulated and formalized across all positions in Liberia's civil service, so an appraisal scheme helps embed them as part of the daily work of public servants.

The PMS is similar in structure to most other performance management schemes in public sectors around the world: in collaboration with their manager, employees commit to a set of performance targets at the start of each annual cycle, which are reviewed and assessed over the cycle, typically twice a year. Managers meet with each of their officers at the start of the cycle to agree on their individual performance targets and how they will be assessed, and they record this information in what we call Form 1. They are then supposed to meet again at midyear, to track progress in achieving individual targets, and at the end of the year, to jointly fill in and discuss a performance diagnostic: Form 2, an updated version of Form 1.[6] Processes are governed by detailed guidelines published by the CSA, which also provides training to managers in how to undertake the process correctly. We focus on the proficiency of individuals and their managers in executing the PMS process.

The PMS has given civil servants who use it better insight into their roles and responsibilities and how these feed into their institutions' overall delivery of public services, but measuring, managing, and rewarding performance remains a challenge. Table 13.1 lists some barriers to ministries' and agencies' effective use of the PMS, as observed by CSA officials.[7]

In addition to the quality of a procedural process as implemented within a unit, the extent to which governments can efficiently manage the communication and coordination of processes across work units is another important measure of the quality of government processes. To assess what we have named external process productivity, we examine the extent to which civil service departments respond to external inquiries. The internal management of the many tasks that civil servants carry out depends on external inputs and consequently requires a chain of activities that span organizational units. We therefore implement a common measurement

TABLE 13.1 Reasons for Incomplete Adoption or Nonadoption of the PMS Process, Liberia

Reasons for not adopting the PMS	Reasons for only partly adopting the PMS	Reasons for not filling in the PMS forms properly
HR officers see the PMS as an added burden on their work.	HR officers did not communicate the timeline to staff.	The forms are bulky. The process is paper-based.
The PMS will be used to fire or remove people from their jobs.	Too much of a paper trail.	Some just fill in forms after being coerced and threatened with disciplinary action.
Some institutions struggle to see the benefits of the PMS to them.	Some do not understand what is fully required of them throughout the PMS cycle.	They have not understood the process.
Leadership lack the willpower to adopt the PMS.	Supervisors with more than 10 staff members find the PMS time-consuming.	
The PMS is a CSA-imposed idea.	Staff expect the CSA to provide guidance at every phase of the PMS.	

Source: Original table for this publication.
Note: CSA = Civil Service Agency; HR = human resources; PMS = performance management system.

framework in both Ghana and Liberia that assesses public officials' responsiveness to requests from the central personnel authorities. We track a set of standardized requests relating to annual personnel record updates undertaken by the two institutions of centralized personnel management, the OHCS in Ghana and the CSA in Liberia. Letters requesting information on all officials in an organization were sent to the census of civil service organizations. For example, the central office might request annual updates to the profile of an organization's staff concerning qualifications and training. The aim of such efforts is for the OHCS or the CSA to plan its capacity-building efforts for the next year based on up-to-date information on current capabilities within the public administration. In Ghana, units were asked to provide staff members' names and civil service IDs as well as any training they had received in the past year. In Liberia, units were asked to provide an updated list of the civil service staff currently employed in their team, including staff members' names, payroll IDs, the names of their direct supervisors, and any relevant training undertaken in the past year.

Assessing the Quality of Processes

In both countries, the processes we study are applicable across all organizations and sectors (though the internal measure in Ghana is general and in Liberia is specific to the appraisal process). This allows us to undertake a common analysis of procedure quality within each public service.

Our approach requires a record of public officials' activities that can be assessed by an independent evaluator. The records in both Ghana and Liberia are dominantly paper-based files that record the "treatment" of projects, files, or cases. The vast majority of such physical files are on-site in a government office. Thus, in the case studies we focus on, we were required to build a team of evaluators that could make physical visits to units to review the government files.[8] To some degree, the digitization of government has supported the improvement of process quality by ensuring that all components of a process required by a procedure are present before the case can be completed. It has also facilitated the sort of inspection and assessment outlined here because enumerators can assess process quality remotely by accessing electronic records, which was not possible in our settings. Besides the ability to access records remotely, however, much of the wider approach described here would be the same in the case of digitized records.

Internal Process Productivity

The evaluations of internal process productivity we undertook in both countries focused on the completeness of records, their degree of organization, and evidence of transparent, logical, and equitable decision-making. First, we focused on measuring the level to which the principal components of administrative documents adhere to the general filing rules. Second, we examined whether the argument laid out in those documents was complete and consistent. Such an approach accords with the overarching concern of the public service rules in the countries of focus that decisions or activities be clearly documented and indicate a logical and equitable decision-making process. The public service rules of each country set the baseline for measures of how the files should have been completed. The OHCS guided the process of designing an instrument to assess process quality in Ghana, and the *Performance Management Policy Manual* (CSA and USAID-GEMS 2016), along with guidance from the CSA, informed the corresponding instrument in Liberia.

Completeness is the level to which the principal components of a file adhere to the general filing rules. In Ghana, the assessment tool collected information on whether the file ladder, folios, memos, minutes, letters, and related documents are compiled correctly, following the public service rules. The file ladder is an important element of a file, summarizing file circulation within an organization and expressing how valuable a file is. According to the general procedure, the file ladder should document all file circulation, specifying the date and the documents involved. To guarantee the accessibility of a file, all documents should be numbered consecutively, starting with folios from the opening of the file to the most recent ones. If actions are required, documents and letters should be minuted, dated, and signed, clearly stating from whom the letters are coming and to whom they are directed. The same procedure is applicable to memos and other relevant records in the file. In addition to dates and signatures, incoming and outgoing correspondence requires specific stamps: the organizational (incoming) and dispatch (outgoing) stamp. Once a file has been passed on to other record officers or stored in the records office, it should not contain either duplicated and draft documents or misfiled and miscellaneous items. Thus, *completeness* is a catch-all for the general handling of government files, assessing the completeness of the file ladder; the consecutive organization of folios within a file; the availability of minutes, memos, and other relevant documents; and the proportion of incoming and outgoing correspondence with dates, stamps, and signatures.

For the appraisal process in Liberia, we similarly searched for complete sets of PMS documents, with all three forms expected in the annual cycle, that echoed the above considerations regarding completeness. Specifically, we looked at how much information had been entered on the PMS forms and whether the civil servants' listed work objectives were linked to their performance indicators, their performance reports, and their supervisor's feedback.

Beyond completeness, evaluators assessed the quality of content in terms of the overall clarity of the file subject and the decision process. In Ghana, we assessed files along six margins:

- How clear is the background to issues?
- How clear is the outline of courses of action available or taken?
- Is the file organized in a logical flow?
- Are choices based on evidence in the file?
- Is it clear who should take action?
- What proportion of materials have a clear deadline?

In Liberia, we reviewed the extent to which civil servants' objectives and performance indicators follow the required SMART framework: whether relevant elements were specific, measurable, achievable, relevant, and time-bound. We assessed files along six distinct dimensions:

- Are different/unique categories of objectives presented?
- Are these objectives specific/measurable/time-bound?
- Are there associated performance indicators/measures?
- What is the extent and quality of reporting on each of these measures?
- Did the manager give recommendations as to how to meet the objectives?
- Did the manager identify development needs and how they could be met?

We also made note of the scores given by managers in the appraisals to assess whether they were validated by the evidence presented in the appraisal documents and indicated a true distribution across the unit.

Table E.1 in appendix E presents the instrument used in Ghana to measure the quality of general processes in government files. Files were assessed on the following sets of indicators:

- The comprehensiveness of reporting on the activity across the series of tasks (for example, "Where applicable, are minutes, memos and other necessary records present and complete [including from whom, to whom and signature]?")
- The sufficiency of the evidence and rationale following each of the decisions made (following the government's due process) (for example, "How would you characterise the quality of content you have in the file along the following margins? Choices are based on evidence in file.")
- The overall commitment to effective processes of the unit as reflected in the file (for example, "In general, to what extent does the file organisation adhere to government procedure? [Give an overall score from 0 to 100.]").

Table E.2 in appendix E presents the instrument used in Liberia to measure the quality of implementation of the appraisal process. Files were assessed on the following sets of indicators:

- The comprehensiveness of reporting across the series of appraisal forms (for example, "Which forms have been completed for Employee [Name]?")
- The sufficiency of the evidence and rationale determining each of the appraisal scores given an employee (for example, "Comments are substantive and provide a quality assessment of officer's contributions [even if discussion is that officer had to do work not in key objectives].")
- The overall commitment to an effective appraisal process of the unit as reflected in the package of appraisal documents (for example, "When reviewing the whole set of appraisal forms for a unit/all those filled in by an appraiser, were there any of the following discrepancies in the set of appraisal forms for the unit? Objectives are very similar across forms.").

External Process Productivity

To measure external process productivity, we tracked the timeliness of units' responses to requests from the centralized service management agency (the OHCS or the CSA) and the completeness and quality of the responses. More specifically, we measured the following:

- The time it took for a unit to respond to the request
- The extent to which all officers on the staff roster for that unit were reported on
- The accuracy of the information (through spot checks, where possible).

In Ghana, the research team tracked request letters from three directorates of the OHCS to public service organizations and the date of delivery of their responses, before and after a clear deadline. The three directorates asked organizations to share five different HR documents: promotion registers, training plans and reports, annual performance reports, the chief director's (CD) self-assessment report, and a signed head of department/director's performance agreement. The research team tracked organizations' internal response time in the execution of a request, recording the period when minutes and memos were executed by schedule officers and the final delivery to the OHCS.

In Liberia, over 400 bureaucratic units and divisions from 28 civil service organizations who were participating in an impact evaluation study were asked to submit personnel files to the CSA. This was done by sending a letter with a set of standardized personnel requests to these study units. The research team then looked at whether the units responded to the request and, if so, what their response time was as a measure of process productivity. The survey firm BRAC assisted the CSA in handing out the letter that communicated this file request and in recording when unit representatives submitted their files in response. Personnel listings were submitted either as hard copies in person or via email to the CSA's Management Services Directorate.

Data Collection

The exercise to assess internal process productivity in Ghana started in April 2018 and lasted for six months, including a two-month pilot. In total, 763 files were assessed from 55 organizations. Randomly sampling across the four main administrative directorates and technical units, the research team audited files from 256 divisions and units.[9]

In Liberia, enumerators assessed the quality of PMS files completed in 2017–19 for the same 437 units that had participated in an impact evaluation study at that time. All Liberian civil servants were supposed to use the PMS process to track and improve performance management. The enumerators thus assessed whether all staff in each unit had completed the PMS forms each year and, if so, the quality of those forms. These three assessments each took place after the completion of the annual PMS process cycle in December 2017, 2018, and 2019.[10] In total, civil servants employed in 437 units and divisions across 28 organizations were assessed on whether they had completed,

TABLE 13.2 Completion of PMS Forms, Liberia, 2017–19

PMS form type	PMS in 2017	PMS in 2018	PMS in 2019
Form 1: Employee performance planning and progress review	1,440	1,655	1,232
Form 2: Employee self-assessment form	774	1,110	600
Form 3: Performance appraisal form	1,297	1,197	547
Individuals with at least one form	2,021	1,587	1,202
Individuals with forms 1 and 3	577	1,090	509
Individuals with all three forms	466	948	498

Source: Original table for this publication.
Note: PMS = performance management system.

in full or in part, the PMS process in 2017–19. Survey data were collected for 7,419 bureaucrats across the three years, whereby 4,810 PMS files were found and could be assessed as a census of available documents (see table 13.2).

The exercise to assess external process productivity in Ghana started in February 2018 and ended in May 2019. In total, 750 letters were tracked during the data collection period sent to 31 ministries and departments in 2018 and 30 ministries and departments in 2019, requesting types of data specific to human resource management (HRM) and policy, planning, monitoring, and evaluation (PPME) organizational divisions. In Liberia, the exercise of requesting and tracking the receipt of personnel files to measure units' responsiveness started on February 24, 2020, and concluded on March 24.

RESULTS

Internal Process Productivity in Government

Table 13.3 presents descriptive statistics for the procedural measures of process quality in Ghana, while table 13.4 presents statistics for the quality of the content of assessed files. We can see a substantial number of files were lacking in at least one of our categories, with only 3 percent of files having a complete or near-complete file ladder, 39 percent having close to the required minutes, and 9 percent having sufficient

TABLE 13.3 Procedural Characteristics of Assessed Files, Ghana

	(1) File ladder: Completeness	(2) File ladder: Transparency	(3) Folios	(4) Minutes and memos	(5) Incoming letters	(6) Outgoing letters
Proportion of files (0–19%)	0.40	0.70	0.35	0.04	0.06	0.72
Proportion of files (20–39%)	0.33	0.04	0.07	0.04	0.04	0.03
Proportion of files (40–59%)	0.04	0.04	0.10	0.15	0.13	0.03
Proportion of files (60–79%)	0.05	0.04	0.18	0.36	0.32	0.06
Proportion of files (80–100%)	0.03	0.03	0.27	0.39	0.42	0.09
Not applicable	0.12	0.15	0.00	0.01	0.04	0.07
Observations	763	763	763	763	763	763

Source: Original table for this publication.
Note: The table reports descriptives of the main dimensions of files' procedural quality. Enumerators were asked to assess files on a Likert scale from 1 to 5, where 1 is "0–19%" and 5 is "80–100%," evaluated on the following margins: "How complete is the file ladder?" (column 1), "Does each step in the file ladder have dates?" (column 2), "Are folios within the file organised and numbered consecutively?" (column 3), "Where applicable, are minutes, memos and other necessary records present and complete (including from whom, to whom and signature)?" (column 4), "What proportion of incoming correspondence has an organisational stamp/date/signature?" (column 5), and "What proportion of outgoing correspondence has a despatch stamp/date/signature?" (column 6).

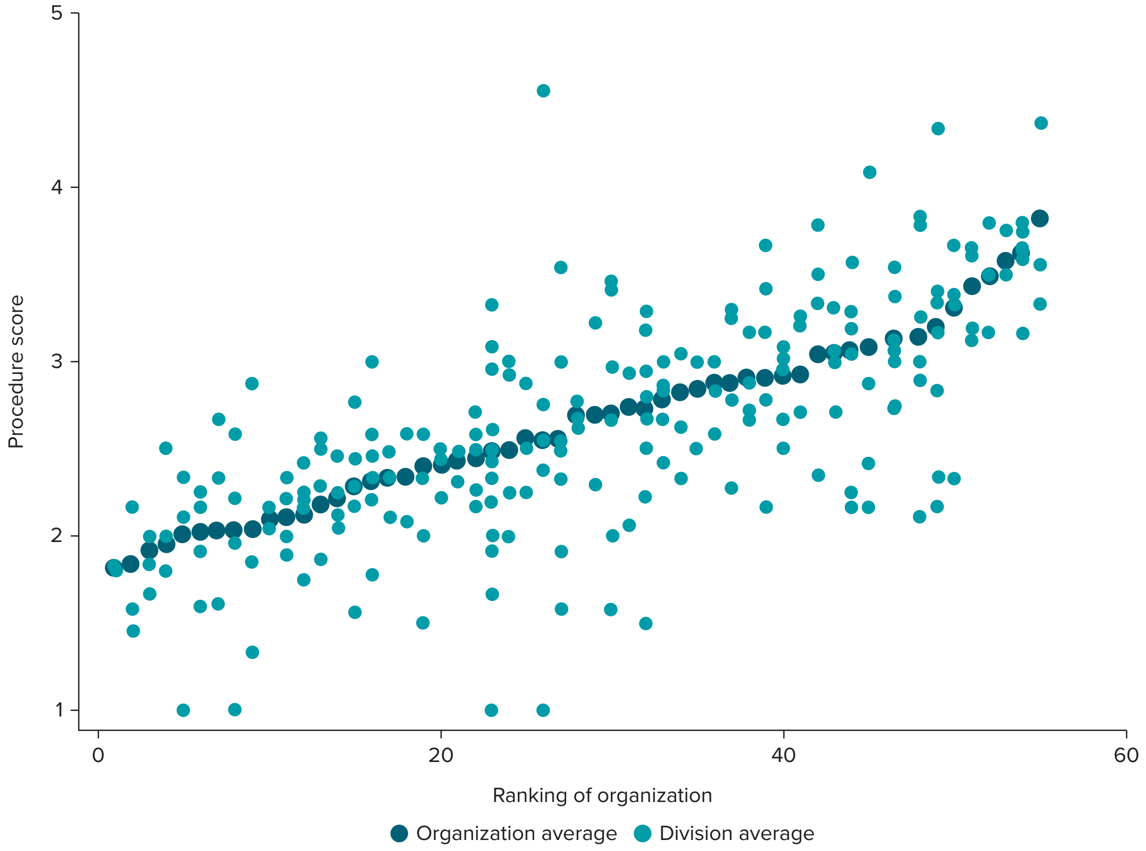
TABLE 13.4 Content Characteristics of Assessed Files, Ghana

	(1) Background to issue	(2) Course action	(3) Logical flow	(4) Choices	(5) Action taken	(6) Clear deadline
Score 1	0.00	0.00	0.02	0.02	0.00	0.08
Score 2	0.08	0.07	0.14	0.17	0.07	0.14
Score 3	0.11	0.09	0.22	0.14	0.09	0.08
Score 4	0.60	0.66	0.47	0.54	0.66	0.03
Score 5	0.19	0.16	0.10	0.11	0.16	0.69
Not applicable	0.01	0.02	0.06	0.03	0.02	0.00
Observations	763	763	763	763	763	763

Source: Original table for this publication.

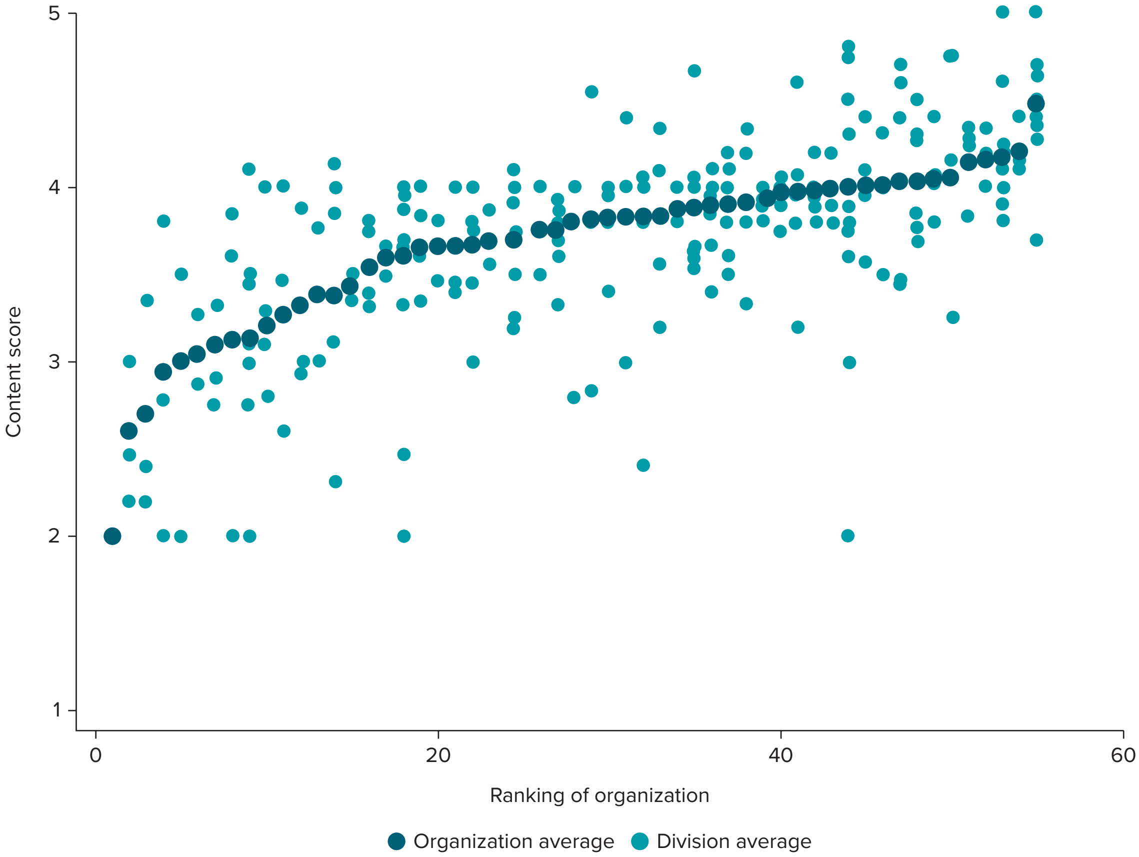
Note: The table reports descriptives on the main dimensions of files' content quality. Enumerators were asked to assess files on a Likert scale from 1 to 5, where 1 is "Very poor" and 5 is "Very good," evaluated on the following margins: "Background to issue" (column 1), "Clearly outlining what courses of action are available or taken" (column 2), "The file is organised in a logical flow" (column 3), "Choices are based on evidence in file" (column 4), and "Clarity on who should take actions at each stage" (column 5). In column 6, enumerators were asked to indicate the proportion of files with a clear deadline.

FIGURE 13.1 Diversity in Level of Procedural Adherence across Organizations and Divisions, Ghana



Source: Original figure for this publication.

FIGURE 13.2 Diversity in Content Scores across Organizations and Divisions, Ghana



Source: Original figure for this publication.

copies of outgoing letters. There is substantial room for improvement. Similarly, only 19 percent of files got the highest score in terms of the background they provided to issues, and 10 percent the highest score for logical flow of the argument. In general, the average level of organizational file adherence to public procedure is poor.

Figure 13.1 showcases how process quality varies across Ghanaian organizations. We average the scores for variables shown in table 13.3 into a single index and plot organizational averages of these scores as dark blue dots. There is a substantial degree of variation in the quality of adherence to processes across organizations. We also plot, stacked vertically at the "rank" of each organization, the scores for individual divisions within those organizations as light blue dots. Thus, the dispersion of the light blue dots around the dark blue dots indicates the degree of variation in process quality within an organization. We take a similar approach to the quality of content in figure 13.2, which summarizes an index created using the measures outlined in table 13.4.

We see a relatively high level of variation across organizations but also within organizations, with those in the middle of the distribution having some units whose process productivity is as bad as the average of the worst-performing organizations. At the same time, there is clearly some degree of correlation between an organization's score and its divisions, indicated by the proximity of the light blue dots to the dark blue ones.

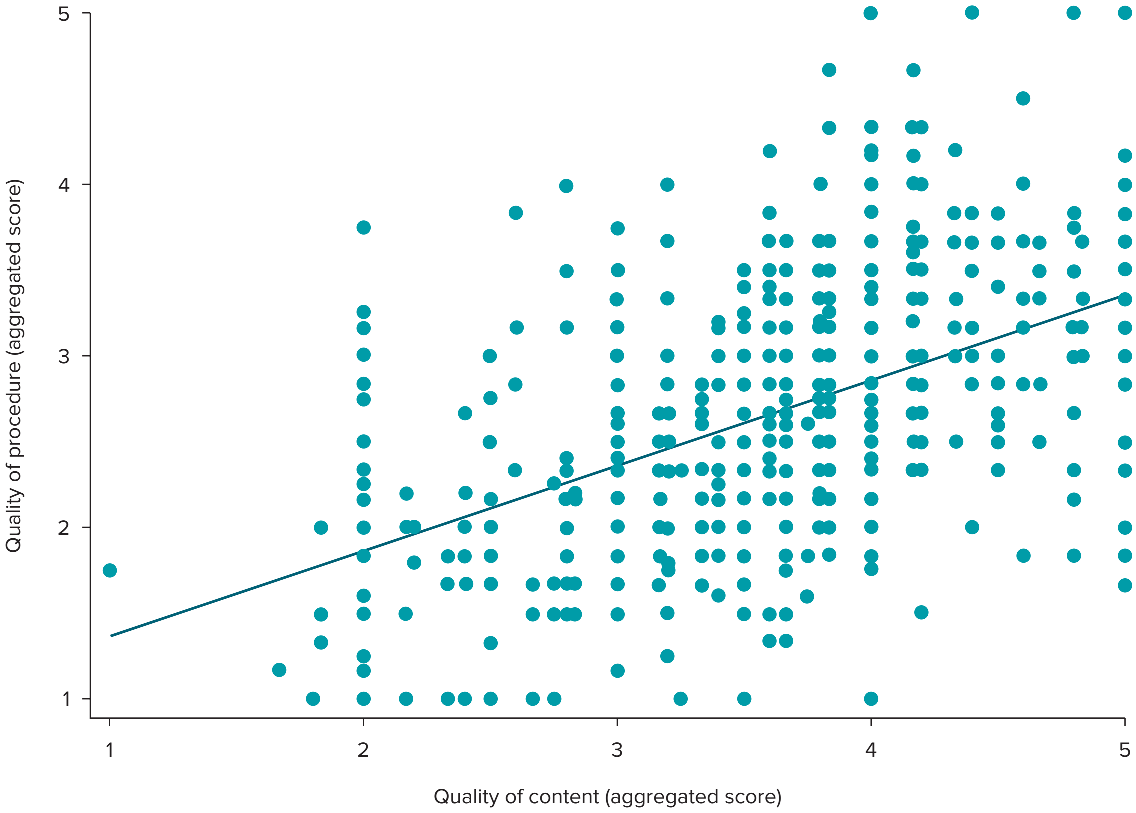
Together, these descriptive statistics tell us that though the general level of process quality is poor, there are some organizations that are able to raise the general standard for processes within their institutions. Though there are still some units that deviate from those practices (either positively or negatively), processes seem to be influenced by organizational practice.

The descriptives also indicate that some areas of process are of higher quality than others. Most of the files presented a blank or not fully complete file ladder, suggesting that organizations were not correctly reporting information on file movement (column 1). About 50 percent of files consecutively numbered folios, while around 40 percent poorly or very poorly organized documents (column 3). A high proportion of memos and minutes on documents were correctly compiled in 75 percent of sampled files (column 4). Looking at correspondence, incoming letters were in general aligned with government procedure, presenting a precise date, a clear signature, and an organizational stamp in 75 percent of cases (column 5). By contrast, outgoing letters were usually poorly compiled: 80 percent of the files show a high percentage of outgoing letters without a dispatch stamp, reflecting an unofficial rule to stamp envelopes rather than letters (column 6).

Likewise, some components of content quality in Ghana fare better than others. About 80 percent of the files had a clear or very clear background to issues (column 1) and clearly outlined what courses of action were available or had been taken (column 2). The files were organized in a logical flow in 57 percent of the cases (column 3), and choices were based on evidence recorded in the documents in 65 percent of the cases (column 4). Documents clearly stated who should take action at each stage in 70 percent of the files in the sample (column 5). On the other hand, the proportion of documents with a very clear deadline is also on the extreme, suggesting either that when documents required a deadline, this was clear, or that some documents did not have a deadline even though required (column 6).[11]

To what extent are those files that adhere to procedures most strongly also those that have a higher quality of content? Figure 13.3 presents a scatterplot (with one dot for each file we study) of the content quality

FIGURE 13.3 Relationship between Adherence to Procedure and Quality of Content, Ghana



Source: Original figure for this publication.
Note: Each dot represents one file that was studied.

score against procedure quality. We see from the trend line that the relationship is positive, and the correlation is 0.48. However, it is also clear from the figure that there is a high degree of variation, with files that are well organized but with weak arguments, and vice versa.

In the Liberian civil service, only a fifth of public officials in the units assessed went through the PMS, reflecting an even weaker adoption of proper procedure in practice in the service.[12] However, when looking at the units that successfully adopted the PMS practice, on average 68 percent of civil servants working in the unit completed at least one of the PMS steps.[13] Table 13.2 illustrates that most staff who utilized the PMS process together with their supervisor completed their initial work plans and midyear assessments (Form 1) and, to a lesser degree, the end-of-year performance appraisal (Form 3). However, less evidence was found of staff assessing their own performance. This is an important piece of the process to ensure that appraisals are fair and the staff are engaged in and buy into the process. Ultimately, 60 percent or less of those who implemented the PMS did so by completing all three required forms. Overall use of the PMS also appears to have peaked in 2018, then fallen in 2019.[14] Hence, completion rates could improve.

Issues around form completeness further hindered enumerators' ability to assess the quality of the content in the forms found. The proportion of forms in which all compulsory sections had been filled in decreased from 84 percent in 2017 to 58 percent in 2018 and 39 percent in 2019. Furthermore, peaking at 50 percent when assessing 2018 forms, enumerators said that they had all the information they needed to assess quality for only 25 percent of the 2019 forms (see table 13.5). On a positive note, the proportion of files stored without a filing system decreased to just 5 percent of all forms found. Even so, table 13.6 shows how a lack of information in the files; poorly organized and at times missing pages; and, to a lesser extent,

TABLE 13.5 Sufficiency of Information for Assessing Quality of PMS Forms, Liberia

Did enumerators have all needed information to assess quality?	PMS in 2017	PMS in 2018	PMS in 2019
Have information needed	743 (37%)	795 (50%)	300 (25%)
Am missing information, but it is not critical to decision on quality	586 (29%)	533 (34%)	521 (43%)
Struggle to make judgment on form quality because of limited information	693 (34%)	259 (16%)	381 (32%)
Observations total	2,027	1,587	1,202

Source: Original table for this publication.
Note: The table shows the number of total observations where true, with the percentage of total observations made in parentheses. Five enumerators refused to answer questions in a survey on the PMS in 2017. PMS = performance management system.

TABLE 13.6 Challenges in Judging the Quality of PMS Forms, Liberia

Form quality issues	PMS in 2017	PMS in 2018	PMS in 2019
No challenges encountered	0.43 (0.50)	0.58 (0.49)	0.36 (0.48)
Little information in file	0.42 (0.49)	0.28 (0.45)	0.49 (0.50)
Poorly organized form	0.15 (0.35)	0.20 (0.40)	0.16 (0.37)
Some pages were missing	0.12 (0.33)	0.11 (0.31)	0.11 (0.31)
Poor level of legibility	0.10 (0.30)	0.11 (0.32)	0.10 (0.30)
Lack of coherence	0.09 (0.28)	0.04 (0.19)	0.08 (0.27)
Subject matter difficult to judge	0.07 (0.25)	0.01 (0.09)	0.02 (0.14)
Total observations	2,027	1,587	1,202

Source: Original table for this publication.
Note: The table shows means, with standard deviation in parentheses. PMS = performance management system.

ineligible, incoherent forms made it difficult to assess the PMS files' quality. This goes to show how a process to map and guide staff performance improvement, such as the PMS, is only as valuable as the level of detail and actionable observations recorded in the PMS forms.

Table 13.7 indicates that civil servants and their supervisors got better, at least initially, at developing SMART objectives and targets, which were then followed up on in midyear and end-of-year progress reports—even if their ability to develop time-bound goals could improve. However, table 13.8 suggests that supervisors could improve at providing practical advice and guidance to their staff through constructive feedback on how they could improve.

Looking at how well staff adopted the SMART objectives as one measure of the quality of the PMS process across organizations and units, figure 13.4 shows substantial variation in adoption. We create an index across different measures of the quality of the performance objectives by averaging the number that meet the relevant criteria. As with the Ghana data, we then take averages of those numbers at the unit and organizational levels. Figure 13.4 illustrates how effectively different organizations have implemented the appraisal process, with some organizations articulating their staff's entire work plan in a single objective. Within these organizations, we see substantial variation, dwarfing the variation across organizations. In the case of the Liberian PMS, process productivity seems to be highly influenced by unit staff.

Drilling down into two of the specific features of SMART indicators—the extent to which they are relevant and measurable—we repeat our analysis but restrict it to the proportion of indicators that were deemed relevant and measurable by our assessors. Figure 13.5 shows that there is once again significant variation across organizations but a similar scale of variation across units within organizations. Thus, again we see evidence that factors at the unit level substantially influence the quality of the PMS process.

TABLE 13.7 Formulating and Reporting on Objectives and Targets, Liberia

SMART objectives and targets	PMS in 2017	PMS in 2018	PMS in 2019
Percent of objectives that are specific	0.92 (0.21)	0.97 (0.10)	0.94 (0.16)
Percent of objectives that are measurable	0.65 (0.42)	0.74 (0.38)	0.56 (0.43)
Percent of objectives that are timebound	0.32 (0.40)	0.34 (0.41)	0.23 (0.36)
Percent of objectives with progress report (midyear)	0.88 (0.30)	0.95 (0.19)	0.81 (0.38)
Percent of objectives that were met/achieved (midyear)	0.81 (0.34)	0.72 (0.41)	0.61 (0.44)
Percent of targets that relate to objectives	0.92 (0.19)	0.93 (0.18)	0.94 (0.17)
Percent of targets that are measurable	0.41 (0.47)	0.66 (0.43)	0.54 (0.45)
Total observations range	924–1,358	1,394–1,545	1,010–1,165

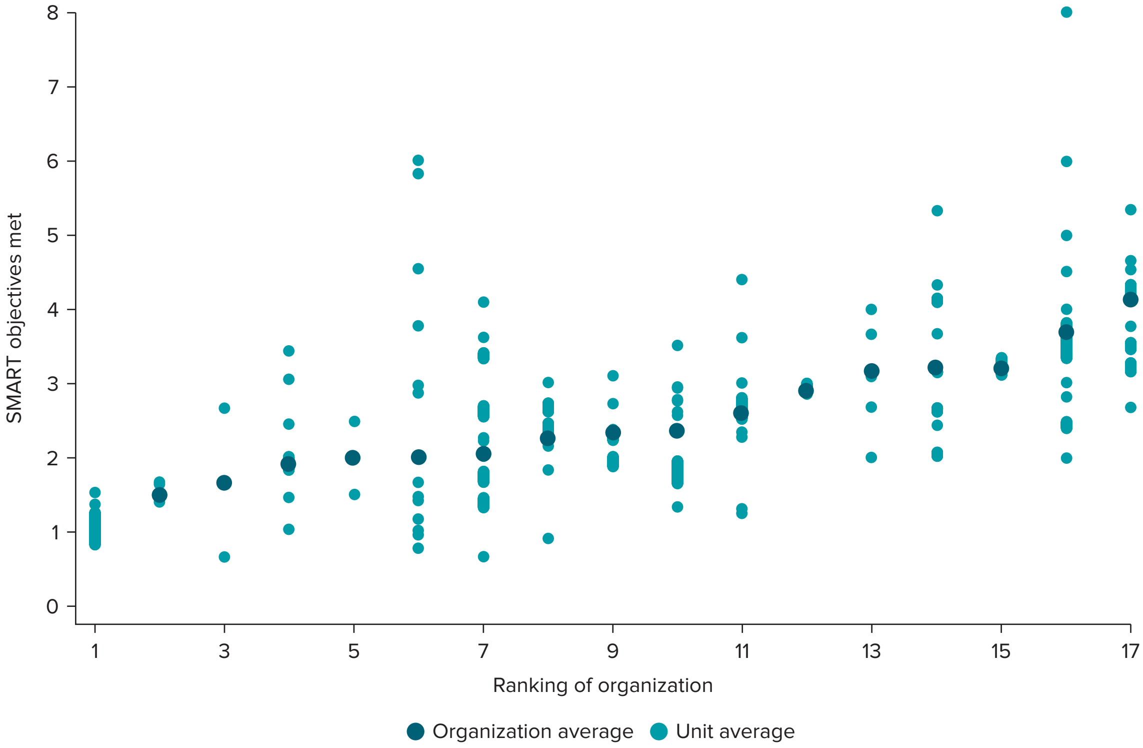
Source: Original table for this publication.
Note: The table shows the means and standard deviation in parentheses. PMS = performance management system.

TABLE 13.8 Quality of Supervisors' Feedback, Liberia

Quality of feedback	PMS in 2017	PMS in 2018	PMS in 2019
Supervisor gave recommendations on how to meet objective	0.66 (0.48)	0.58 (0.49)	0.40 (0.49)
Supervisor identified development needs of the employee	0.48 (0.50)	0.43 (0.50)	0.41 (0.49)
Supervisor recommended activities to build employee's capacity	0.44 (0.50)	0.41 (0.49)	0.36 (0.48)
All objectives are reported on	0.64 (0.48)	0.72 (0.45)	0.23 (0.42)
All comments are substantive	0.38 (0.49)	0.41 (0.49)	0.27 (0.44)
All comments are constructive	0.25 (0.43)	0.31 (0.46)	0.04 (0.19)
Total observations range	943–1,353	1,069–1,544	510–1,010

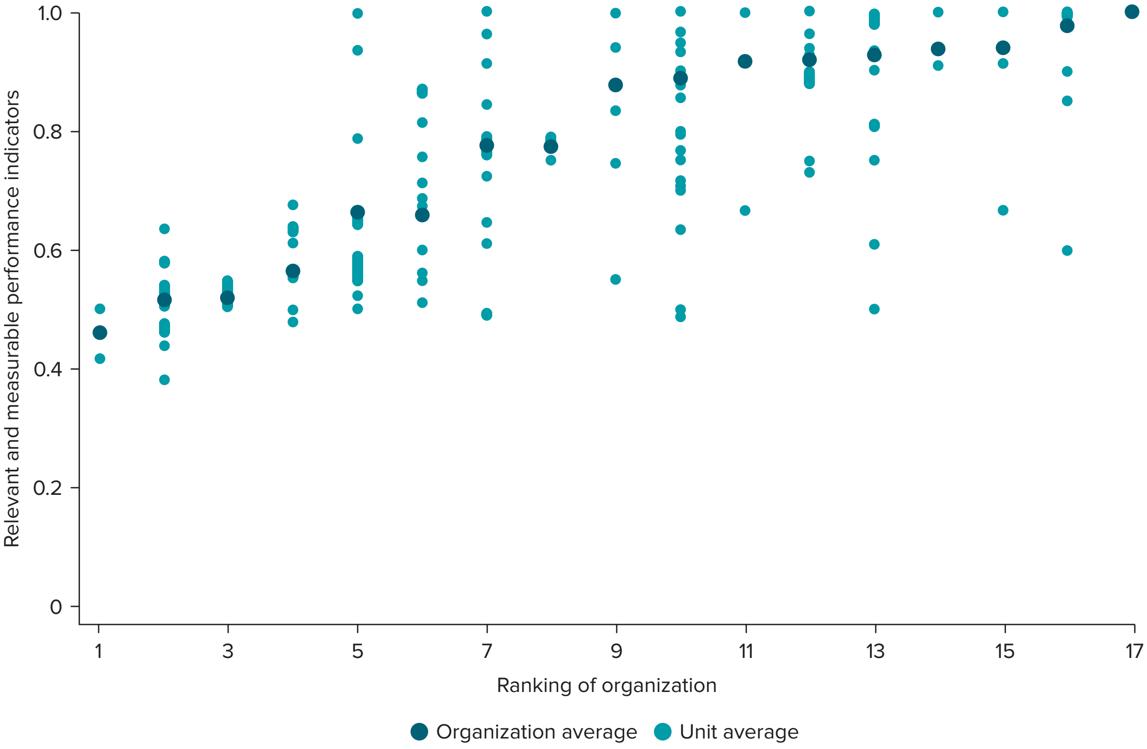
Source: Original table for this publication.
Note: The table shows the means and standard deviation in parentheses. PMS = performance management system.

FIGURE 13.4 Average Number of SMART Objectives Identified in Appraisal Forms, Liberia



Source: Original figure for this publication.

FIGURE 13.5 Average Number of Relevant and Measurable Indicators Identified in Appraisal Forms, Liberia



Source: Original figure for this publication.

External Process Productivity in Government

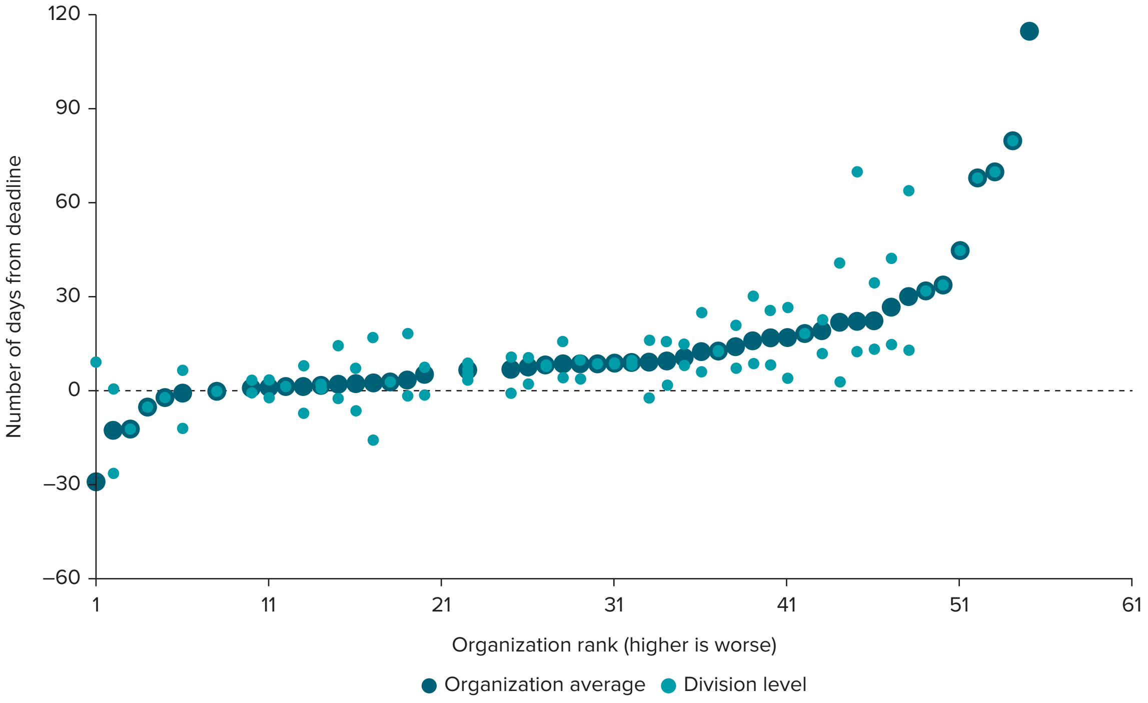
We now turn to the results of our assessments of external process productivity. Figure 13.6 shows the average number of days it took an organization (dark blue dots) to submit a response to the various requests made by the OHCS in 2017 and 2018 (*y* axis), with the organizations ranked by overall speed (*x* axis). Once again, we also present averages for the units within those organizations (light blue dots stacked vertically at their organization's ranking), but given that many such requests must be sent by the centralized dispatch office of the organization, we see a lot of clustering in the unit averages. A negative number on the *y* axis implies that the submission was received before the deadline (represented by 0 on the *y* axis).

Perhaps a third of organizations in Ghana's public service who eventually responded met the deadlines set by centralized entities. A minority of organizations were fully unresponsive and thus are not displayed in figure 13.6. However, even among those who eventually responded, perhaps a quarter did so a month or more late. Such delays impact the ability of central organizations to continue activities for which they require external information.

Turning to quality, figure 13.7 shows the completeness of the submissions received by the OHCS. The *y* axis displays the proportion of requests for which an organization (dark blue dot) or unit (light blue dot) submitted the required information. Here, organizational and unit averages are less closely related, since central dispatch offices will rarely mediate the quality of submissions. A few ministries, departments, and agencies submitted more than 80 percent of the data requested by the OHCS, and some units submitted all the information, while others submitted less than 20 percent of the information requested. The average level of quality is rather low, with the median organization submitting just over 60 percent of the information requested. All of this has knock-on effects on the capacity of the OHCS to undertake its work.

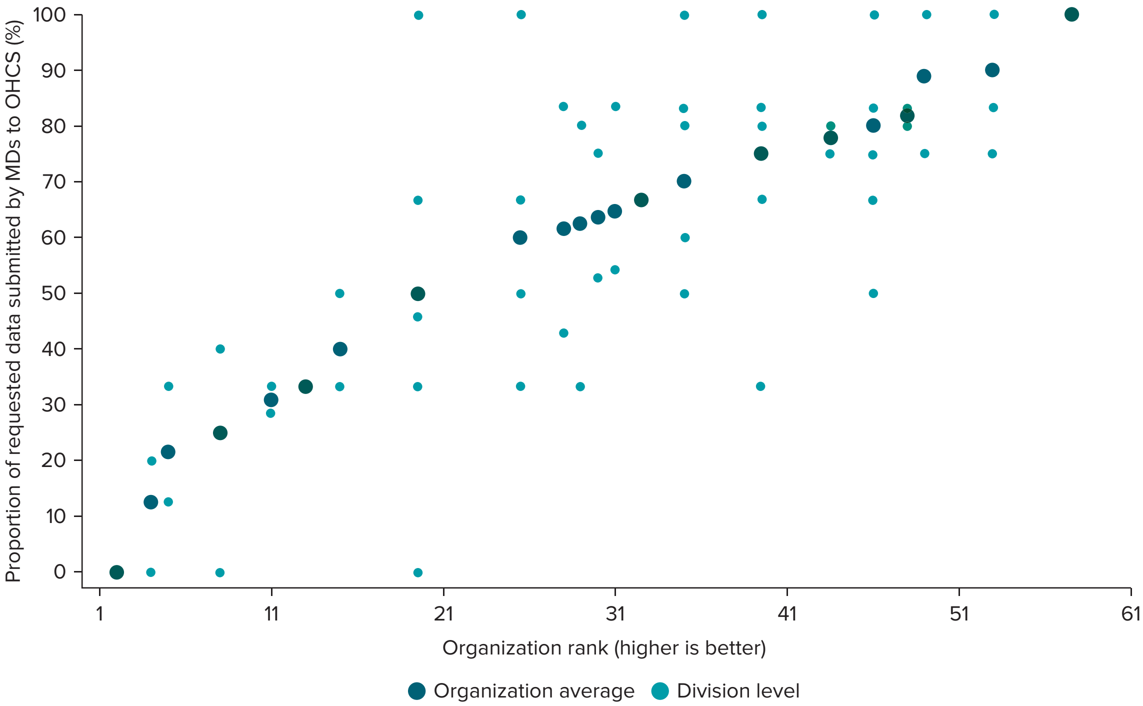
A similar picture is found in Liberia. Though not displayed here, we find similarly low responsiveness to centralized requests, with an even greater number of organizations simply not submitting any response at all.

FIGURE 13.6 Diversity in Number of Days to Receive Requested Information from Organizations and Divisions, Ghana



Source: Original figure for this publication.

FIGURE 13.7 Diversity in Proportion of Required Files Submitted by Organizations and Divisions, Ghana



Source: Original figure for this publication.
Note: MDs = ministries and departments; OHCS = Office of the Head of Civil Service.

Of the 348 units across government that we confirmed received the CSA's request, 30 units responded, 21 (70 percent) within the deadline. The quality of those submissions is even more limited, with many containing little to no usable information. Trying to undertake personnel policy making when your colleagues in the rest of the service simply refuse to answer your requests for information must be challenging.

CONCLUSION

This chapter has put forward a framework for measuring process quality in public administration: identifying evidence of transparent, logical, and equitable decision-making throughout government. Though it is a fundamental part of the activities of the public sector, the quality of public officials' work processes has rarely been measured for government analytics. This drives assessments of government functioning and productivity toward frontline services and limits analysts' capacity to assess where in the long chain of government processes dysfunction might be occurring.

We have distinguished between internal process productivity, the quality of administrative processes for activities confined within a particular administrative unit, and external process productivity, the quality of administrative processes for activities in which units interact. We have made this distinction because accountability and professional dynamics vary distinctly between the two cases but also because appropriate measurement varies as well. We have then applied our framework to two case studies, concerning general government processes in the government of Ghana and the appraisal process in the government of Liberia. We have shown that in these settings, the quality of government processes is generally poor but highly varied, with some organizations and units effectively adhering to government processes and a higher overall quality of administration.

Such measures of process quality in public administration open up three areas of government analytics: assessments of variation in process quality and associated productivity within and across organizations in the same country, comparisons of process quality across countries, and assessments of public sector process quality over time. In this way, government analysts can pinpoint where government procedure is not being adhered to, how different processes relate to public sector productivity, and what the dynamics are across individuals and organizational units.

Strengthening the quality of government processes would require increasing and updating the knowledge of public officials on appropriate ways of handling government work, strengthening senior officers' supervision, and reinforcing their capacity to hold staff to account for poorly adhering to government processes. As the world's public administrations become increasingly digital, the ability to detect substandard processes will become more automated, but the continued assessment of which processes lead to improved productivity will require the use of this information for analysis. We hope this chapter has provided a framework for such work.

NOTES

The authors gratefully acknowledge funding from the International Growth Centre; Economic Development and Institutions; the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice; and the USAID-, Sweden-, and World Bank–sponsored Liberia Public Sector Modernization Project. We thank Nyetuan Mulbah, Francesco Raffaelli, and Andre Cazor Katz for excellent research assistance and the heads of Ghana's and Liberia's civil services under which the work was implemented, Nana Agyekum-Dwamena (Ghana) and Puchu Leonard and James A. Thompson (Liberia). We thank Mrs. Rejoice Dankwa, Mr. Godwin Brocke, Patience Coleman, Stefan Dercon, Erika Deserranno, Aisha Nansamba, Dorothy Kiepeeh, Smile Kwawukume, Vincent Pons, Imran Rasul, and George B. Wah for their guidance. All errors are our own. Finally, this paper was published after the passing of our coauthor, Felix Nyarko Ampong, and we therefore dedicate the work to him.

[1] See the measures under "Instructional leadership" in table 29.3 of chapter 29.

[2] It should be noted that some of the indicators of proper procurement and customs procedures are versions of measures of process productivity.

[3] Frontier empirical evidence on what bureaucrats do showcased in chapter 17 implies that almost three-quarters of bureaucratic work is related to undertaking bureaucratic processes, such as monitoring, training, and personnel management; financial and budget management; and so forth. It would seem that process productivity is key to the productivity of the public sector.

[4] An intermediate approach is Hollyer, Rosendorff, and Vreeland (2017), who use reporting to the World Development Indicators as a measure of government transparency.

[5] The OHCS has a Public Records and Archives Administration Department (PRAAD), whose aim is to facilitate and promote good government processes and record-keeping practices across ministries and departments. Officials are trained in relevant processes upon entry to the public service, as well as at regular in-service trainings.

[6] At the end-of-year review, employees are supposed to assess their own performance against 10 servicewide standards in what we refer to as Form 2. They are further assessed by their supervisors on these 10 servicewide indicators, as well as on their individual overall performance and behavior in the workplace, in Form 3.

[7] Importantly, there have been efforts to engage on the PMS between CSA and ministries or agencies, to train hundreds of supervisors and staff on the PMS cycle, and to assign individuals in each public agency to act as focal points on issues related to the rollout of the PMS. Still, limited political will to adopt the process in a timely manner; its paper-based format; and disconnect from any recognition, rewards, or sanctions system remain persistent challenges.

[8] We employed senior and retired civil servants in Ghana to review the extent to which randomly chosen unit files followed appropriate government processes, whereas, in Liberia, this was done by enumerators from an external survey firm.

[9] The sampled files were assessed by three assistant management analysts from the Management Services Department of the OHCS. During the piloting period, the tool was adjusted and improved to reflect the records management practices within the Ghanaian civil service. Files in the sample are indicatively opened in 2015, not confidential, and not related to personal or financial subjects.

[10] The files were assessed by enumerators from Liberia-based survey firm BRAC.

[11] In this case, the tool allowed a "not applicable" option. In 54 percent of the files assessed, documents did not require a specific deadline.

12. With an estimated total workforce of 7,099 in the units assessed, based on 2017 staff lists, only 28 percent, 22 percent and 17 percent of staff had completed at least one of the PMS forms in 2017, 2018, and 2019, respectively.
13. In the units with any adoption in that year, 65 percent, 67 percent, and 71 percent of staff had filled in at least one PMS form in 2017, 2018, and 2019, respectively.
14. A new administration came into office in 2018, and numerous pay reforms that resulted in pay cuts for some in 2018 and 2019 may have impacted civil servants' motivation and prioritization of the PMS process.

REFERENCES

Alesina, Alberto, and Guido Tabellini. 2007. "Bureaucrats or Politicians? Part I: A Single Policy Task." *American Economic Review* 97 (1) (March): 169–79. https://doi.org/10.1257/aer.97.1.169.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2021. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training." *American Economic Journal: Economic Policy* 13 (1) (February): 36–66. https://doi.org/10.1257/pol.20190664.

Bedoya, Guadalupe, Amy Dolinger, Khama Rogo, Njeri Mwaura, Francis Wafula, Jorge Coarasa, Ana Goicoechea, and Jishnu Das. 2017. "Observations of Infection Prevention and Control Practices in Primary Health Care, Kenya." *Bulletin of the World Health Organization* 95 (7) (July): 503–16. https://doi.org/10.2471/BLT.16.179499.

Berliner, Daniel, Benjamin E. Bagozzi, Brian Palmer-Rubin, and Aaron Erlich. 2021. "The Political Logic of Government Disclosure: Evidence from Information Requests in Mexico." *The Journal of Politics* 83 (1): 229–45. https://doi.org/10.1086/709148.

Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. https://ideas.repec.org/p/nbr/nberwo/23350.html.

Chong, Alberto, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2014. "Letter Grading Government Efficiency." *Journal of the European Economic Association* 12 (2): 277–99. https://doi.org/10.1111/jeea.12076.

CSA and USAID-GEMS (The Civil Service Agency and the USAID Governance and Economic Management Support Project). 2016. *Performance Management Policy Manual for the Civil Service of Liberia*. Monrovia, Liberia: Civil Service Agency. https://csa.gov.lr/doc/Performance%20Management%20System%20Manual.pdf.

Daniels, Benjamin, Amy Dolinger, Guadalupe Bedoya, Khama Rogo, Ana Goicoechea, Jorge Coarasa, Francis Wafula, Njeri Mwaura, Redemptar Kimeu, and Jishnu Das. 2017. "Use of Standardised Patients to Assess Quality of Healthcare in Nairobi, Kenya: A Pilot, Cross-Sectional Study with International Comparisons." *BMJ Global Health* 2 (2): e000333. https://doi.org/10.1136/bmjgh-2017-000333.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *The Journal of Human Resources* 37 (4): 696–727. https://doi.org/10.2307/3069614.

Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. https://doi.org/10.3982/ECTA19244.

Forte, D. Othniel. 2010. "Civil Service Reform in Post Conflict Liberia." Unpublished manuscript. https://www.academia.edu/12454955/Civil Service Reform in Post Conflict Liberia.

Friedman, Jonathan. 2012. "Building Civil Service Capacity: Post-Conflict Liberia, 2006–2011." Innovations for Successful Societies, Princeton University, August 2012. https://successfulsocieties.princeton.edu/publications/building-civil-service-capacity-post-conflict-liberia-2006-2011.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2017. "Measuring Transparency." *Political Analysis* 22 (4): 413–34. https://doi.org/10.1093/pan/mpu001.

Peisakhin, Leonid, and Paul Pinto. 2010. "Is Transparency an Effective Anti-Corruption Strategy? Evidence from a Field Experiment in India." *Regulation & Governance* 4 (3): 261–80. https://doi.org/10.1111/j.1748-5991.2010.01081.x.

PRAAD (Public Records and Archives Administration Department). 2007. *Records Office Procedures Manual*. Accra: Government of Ghana.

Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal* 128 (608): 413–46. https://doi.org/10.1111/ecoj.12418.

Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2): 259–77. https://doi.org/10.1093/jopart/muaa034.

Wafula, Francis, Amy Dolinger, Benjamin Daniels, Njeri Mwaura, Guadalupe Bedoya, Khama Rogo, Ana Goicoechea, Jishnu Das, and Bernard Olayo. 2017. “Examining the Quality of Medicines at Kenyan Healthcare Facilities: A Validation of an Alternative Post-Market Surveillance Model that Uses Standardized Patients.” *Drugs—Real World Outcomes* 4 (1): 53–63. https://doi.org/10.1007/s40801-016-0100-7.

Wehner, Joachim, and John Poole. 2015. “Responsiveness of UK Local Governments to FOIA Requests.” LSE Department of Government Working Papers, London School of Economics and Political Science, London.

Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.

Wood, Abby K., and David E. Lewis. 2017. “Agency Performance Challenges and Agency Politicization.” *Journal of Public Administration Research and Theory* 27 (4) (June): 581–95. https://doi.org/10.1093/jopart/mux014.

World Bank. 2014. *Liberia—Public Sector Modernization Project.* Washington, DC: World Bank. http://documents.worldbank.org/curated/en/53084146 8263672030/Liberia-Public-Sector-Modernization-Project.

CHAPTER 14

Government Analytics Using Customs Data

Alice Duhaut

SUMMARY

Many government agencies have multidimensional missions, in which achieving one objective can reduce the attainment of another organizational objective. This presents particular challenges to government analytics. Incomplete measurement of objectives risks encouraging the attainment of the measured objectives while unknowingly impairing other objectives. This chapter showcases how government analytics can be applied in such contexts, using the example of customs agencies. Customs agencies typically have three core objectives: facilitating trade, collecting revenue, and ensuring the security and safety of the goods entering or exiting the country. Attaining one objective (for example, the greater safety of traded goods) can come at the expense of another (for example, facilitating trade). This puts a premium on the effective measurement of all dimensions of a customs mission, which requires triangulating different data sources. This chapter showcases how this can be done, deriving indicators for trade facilitation (for example, the costs of the process—in particular, time delays), revenue collection (for example, trade volume and revenue collected based on the assessed value), and safety (for example, the number of goods in infraction seized). The chapter also underscores how a wider use of the customs database itself could help measure performance, by combining it with other data collection methods, such as the World Customs Organization (WCO) Time Release Study (TRS) and exciting developments in GPS tracking data.

ANALYTICS IN PRACTICE

- Government organizations with multidimensional missions—such as customs—typically need to integrate multiple data sources to ensure they measure performance holistically and avoid measuring and focusing on some goals but not others. In customs, the efficiency of the border-crossing process, and the customs agents and other agencies involved in it, should be evaluated with both traditional tools—the World Customs Organization (WCO) Time Release Study Plus (TRS+) and monitoring and evaluation metrics—and new or underused data sources—such as GPS data—to provide a way to reduce

Alice Duhaut is an economist in the World Bank's Development Impact Evaluation (DIME) Department.

the cost and increase the frequency of the indicators used to monitor border activities. An important element of the consolidation is to ensure the validity of the data used, match the relevant time stamps to the mapped process, and program indicators and queries to automatize reports.

- Data from different sources are likely to provide a different view, and even different takes, on the process. Measurement validation and triangulation are important components in analyzing customs data. It is thus important to invest in understanding the data routinely collected and to analyze them outside of survey periods. To complement the measures derived from the traditional TRS+, we recommend using customs database data to study time delays under customs' or other border agencies' control and the revenue collected. This requires understanding the full customs process and ensuring entries are not duplicated or incomplete, as might be the case if customs declarations for a shipment can be resubmitted under a different regime (for example, when the importer wants the goods to leave the warehouse and be released).
- Data should be standardized and rendered into reports for easy and fast consumption. Standardization of the extraction process, indicators, questions, and data treatment helps reproduce reports at a high frequency. From user surveys, information on the performance of the customs agent can also be collected.
- Valuation of goods in customs is challenging. To provide a holistic assessment, there are multiple techniques available to measure the value of goods in customs. In particular, comparing the value of goods when they leave a country of origin to their destination may assist in identifying the true value of goods. While valuation is a difficult process, and the World Trade Organization (WTO) rules describe how individual items' values should be evaluated, comparing what is declared at a country's borders to what is declared for similar goods of similar origin in peer countries can provide information on international trade taxes, the duties and excises collected, and the timeliness of the process. This indicator can flag where the value collected at customs is lower than expected.
- Time is an important consideration in customs, but the relevant checkpoints along the customs process against which it is measured must be defined (for example, if the clearance of the goods is considered the endpoint of a time analysis). Time delays can be studied in association with the mapping process to determine the relevant operations: one common operation studied based on Automated System for Customs Data (ASYCUDA) data is the time between assessment and clearance excluding the payment of taxes and duties. This exclusion is important because payment issues can be the cause of a lot of the delays, and such findings would have different policy implications.

INTRODUCTION

Many government agencies have multidimensional missions, in which achieving one objective can reduce the attainment of another organizational objective. For instance, in some countries, financial regulators are mandated to develop financial services while also protecting consumers, or environmental agencies are mandated to both protect and develop natural resources. Organizations with such multidimensional missions with conflicting goals present particular challenges to government analytics. Incomplete measurement of objectives risks encouraging the attainment of the measured objectives while unknowingly impairing other objectives. Yet different objectives can often only be measured through very different types of data. This chapter showcases how government analytics can be applied in such contexts, using the example of customs agencies. By showcasing the integration of different data sources to measure multidimensional mission attainment holistically, the chapter complements other chapters in *The Government Analytics Handbook* that detail the use of one particular form of data—such as case data in chapter 15 or task data in chapter 17.

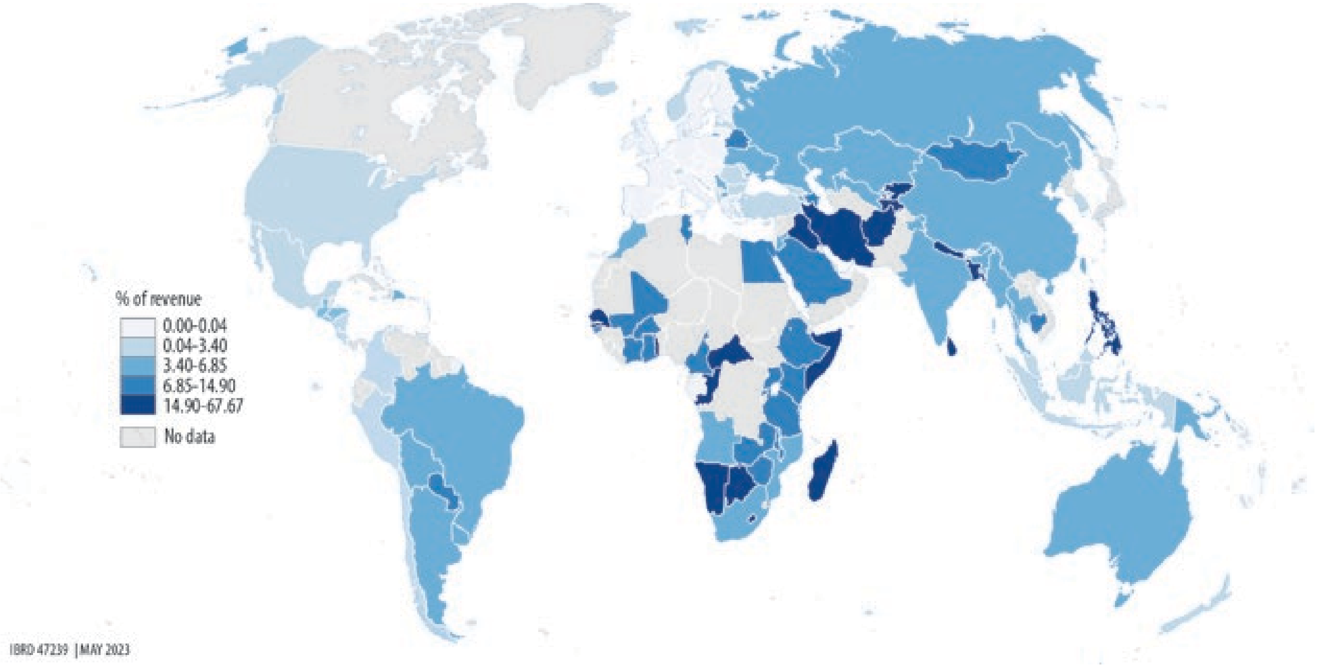
Among agencies tasked with multidimensional missions, customs is arguably a key one. Customs operations are located within the international borders of a country and are responsible for the processing of export and import goods. This includes several steps, from reviewing goods declarations to risk assessments, inspections, clearance, and risk collection. Revenue collection from customs is particularly important in low- and middle-income countries. Frequently, it represents a substantial share of state revenue and, at times, is also used for collecting fees requested by other government agencies.

In high-income countries, customs and other import duties represent, on average, 3.8 percent of state revenue, but, as illustrated in map 14.1, this value rises for countries with lower average incomes. For upper-middle-income countries, it stands at 8.9 percent, for lower-middle-income countries, at 11 percent, and for low-income countries, at 20 percent. For some countries in Sub-Saharan Africa, South Asia, and the Pacific islands, customs and import duties provide over one-third of all tax revenue. In addition to its key role in revenue collection, customs ensures borders' integrity and is the point of entry for goods coming into or going out of the country. For these reasons, the performance of customs operations has substantial implications on the fiscal sustainability and trade engagement of countries. The challenge of improving customs performance can thus be viewed from the vantage point of the following three distinct missions: the facilitation of trade across borders, the collection of revenue, and the protection of the safety of people and the security of goods coming through the borders. Working toward these three missions simultaneously involves trade-offs: making progress toward one goal can undermine the achievement of another. For example, facilitating trade means improving the customs process to reduce its total duration, including inspection and screening times. If the frontline agents were to perform fewer inspections, this would lead to a faster border crossing. However, this would likely have revenue implications, as proper product classification and tax collection would be more prone to errors.

Examples of initiatives undertaken toward those three key objectives, as well as some associated challenges, are illustrated in box 14.1 on the case of Malawi.

This puts a premium on measuring the performance of agencies with multi-dimensional missions—such as customs—holistically to ensure these trade-offs are accounted for. With this in mind, this chapter provides an empirical guide to assessing customs performance across these three objectives, as well as

MAP 14.1 Customs and Other Import Duties as a Percentage of Tax Revenue



Source: World Development Indicators (latest available values).

BOX 14.1 Interactions with Other Agencies: The Case of Malawi

The Republic of Malawi is a landlocked country in southeastern Africa (see map B14.1.1). In Malawi, 14 agencies are present at the border. The agencies perform the inspections related to their missions: for instance, the Malawi Bureau of Standards ensures that the foodstuff coming in to the country respects Malawian standards. Together, these agencies strive to improve performance as related to the three principal objectives of customs operations below:

Trade facilitation: Malawian customs agents strive to improve the flow of goods and services across the border. One example of their efforts in this sphere is upgrading to the Automated System for Customs Data (ASYCUDA) World system in 2018. The new system facilitates trading across the border by, among other things, allowing web access for businesses, enabling the round-the-clock submission of customs declarations, and providing customized data extraction features.

MAP B14.1.1 Location of Malawi

IBRD 47220 |
APRIL 2023
TANZANIA
MALAWI
ZAMBIA
ZIMBABWE
MOZAMBIQUE

Source: World Bank.

Revenue collection: Customs and other revenue duties collected in Malawi were equal to MK 88.3 billion in 2019, which represented 8.9 percent of all state tax revenue.

Protection of the safety of people and security of goods coming through the borders: The Malawi Revenue Authority restricts the import of certain classes of goods by requiring import licenses. These include military uniforms, ammunition and guns, fertilizer, pharmaceuticals, gold, and several types of foodstuff. In addition, all animals and animal products need to be certified as disease-free. Importation of most types of meat products further requires prior written permission from the Minister of Industry and Trade.

(continues on next page)

BOX 14.1 Interactions with Other Agencies: The Case of Malawi *(continued)*

In the future, the agencies will be connected to the customs database, and customs will perform joint inspections. While the reduction in the number of agencies at the border is likely to reduce the burden on drivers or transporters, customs will also have more responsibilities, and measuring their performance after the reform against the preintervention situation might be complicated. It is thus necessary to create indicators that will reflect the scope of the mission as well as be easy to implement with the existing data.

outlining the diverse data necessary to perform these assessments. There are multiple choices available for practitioners when building both performance indicators and customs databases. Practitioners should prioritize indicators that enable them to accomplish a particular policy objective, while considering the data and human resources constraints they face when developing them. Because modifications in how customs operates affect other policy areas—trade policy and fiscal revenue—any change in how data are ingested and consumed should be coordinated with other government agencies.

The chapter is structured as follows. First, it provides institutional context on how customs operates and the international policy framework governing customs data collection and trade policy. Section 3 reviews the academic literature on customs, emphasizing its role in trade facilitation and fiscal revenue. Section 4 outlines the data infrastructure requirements for analyzing customs performance and generating indicators. Section 5 presents different types of indicators used to measure customs performance, and the final section concludes.

INSTITUTIONAL CONTEXT

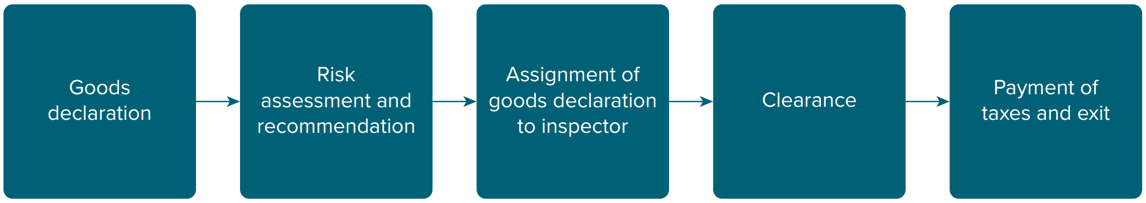
Customs Process Overview

The customs process follows a linear structure, from the formal declaration of goods to the payment of taxes and exit. The process starts with the submission of a goods declaration to the customs administration by the importer or exporter, or by a broker acting on behalf of the importer or exporter (figure 14.1). This can be done remotely or at the border, depending on the country. The goods declaration usually lists a description of the items, with the classification, weight or quantities, origin, and value of the items in the shipments. Supporting documents, such as an invoice and bill of landing, are submitted along with the declaration. In addition, the customs declaration contains the declarant's assessment of the taxes and duties to be paid.

The next step is risk assessment. This step can take place before the submission is made or when it is made, as soon as the goods arrive at customs. An initial screening is conducted through a customs database risk model, analyzing the risk level of a declaration and issuing recommendations at the product level. The risk department usually issues a color-coded clearance channel, in which the color indicates whether the documents or the goods have to be inspected, sends flags for potential fraud or discrepancies in the declaration, and, potentially, sends comments to help the inspector assess the correct valuation of the shipment.

The customs process then moves to the assessment of the declaration by an inspector. Based on the documentation submitted by the declarant and the diagnostics provided by the risk department, the inspector can overrule the clearance channel recommendations. If the green channel is recommended, nothing happens, and the shipment goes through customs uninspected. If the yellow channel is recommended, the documents submitted along with the goods' declaration are reviewed. If the red channel is recommended, the goods are inspected—either by scanning the container or opening the cargo. Based on the information

FIGURE 14.1 Diagram of Customs Process


Goods declaration
Risk assessment and recommendation
Assignment of goods declaration to inspector
Clearance
Payment of taxes and exit

Source: Original figure for this publication.

accumulated, the inspector produces a report on the declaration. The report lists any adjustments to the classification of the goods, the origin, product characteristics or quantity, and, importantly, the value assessed, as well as the taxes and duties to be paid. It can also include penalties to be paid in case of fraud.

In the last two stages, the goods are cleared and the taxes and duties are paid. The goods are released upon proof of payment. The term *clearance* means the accomplishment of all formalities necessary to allow goods to enter home use or to be exported. *Release* means that the goods are physically placed at the disposal of the transporter or importer.

International Trade Policy Framework

While work to efficiently regulate customs operations is done on the domestic front by customs authorities, international organizations play a significant role as well. Since trade is, by nature, the international flow of goods, multiple international trade policy frameworks have been designed to regulate it and provide guidance for domestic customs authorities. These frameworks have been built and advocated for by a set of international organizations, including the World Trade Organization (WTO), with its rules on customs valuation, the World Customs Organization (WCO), the voice of the international customs community, and the United Nations Conference on Trade and Development (UNCTAD). This section provides practitioners with an overview of these different international trade policy frameworks and agreements.

The WTO trade facilitation agreement (TFA) reached at the 2013 Bali Ministerial Conference includes provisions related to customs operations. Intended to expedite the movement, release, and clearance of goods, the agreement sets up procedures for effective communication between customs authorities and other entities directly involved in customs compliance issues. As a result, all WTO members can benefit from technical assistance and capacity building related to any area of everyday customs work. In particular, the TFA, which finally entered into force in February 2017 after being ratified by two-thirds of WTO members, was followed in July 2014 by the launch of another important tool, the Trade Facilitation Agreement Facility (TFAF). It was the first time in WTO history that the obligation to implement an agreement was linked to the capacity of the country to do so.

The mission of the WTO and other international organizations, such as the WCO, is broad. The WTO and WCO cooperate on a number of initiatives: customs valuation, market access, rules of origin, information technology agreement, and trade facilitation. Among numerous examples of such cooperation is the WTO's Agreement on Customs Valuation, which established the Technical Committee on Customs Valuation under the rule of the WCO. In the area of technical assistance, according to the WTO, the main focus remains on negotiations surrounding technical assistance. Another example is the Harmonized Commodity Description and Coding System, or "Harmonized System," a classification of goods under the lead of the WCO, which the WTO thoroughly follows. Established by the Tokyo Round agreement, the WCO's Technical Committee on Customs Valuation and the General Agreement on Tariffs and Trade (GATT) and WTO Committee on Customs Valuation provide advice and case studies on customs valuation. These international efforts provide a legal framework to regulate customs operations so that each member state determines the value of goods in a *neutral* and *uniform* way.

Historically, general principles for an international system of valuation were established under the GATT Article VII. The agreement sets the actual value of a good, the price at which merchandise is sold under competitive conditions. It was the first agreement for customs valuation that highlighted the importance of competitive conditions for the determination of the sale price and stated that the price under established rules should be related to either comparable quantities or quantities not less favorable. At the same time, the need to simplify and harmonize international trade procedures coexists with growing pressure from the international trading community to minimize the intervention of the government in commercial transactions (Widdowson 2007). WTO rules on customs valuation highlight the discretionary autonomy that customs authorities must retain to fulfill their duties in promoting food safety and security and fighting illegal practices.

Measurement of time as a critical component for efficient customs operations has been dictated by the WCO Time Release Study (TRS) as well. The TRS is a methodology to measure, using data-driven approaches, the time that it usually takes to release cargo. It is a part of the Performance Measurement Mechanism (PMM) thoroughly monitored by WCO. Aimed at data-driven decision-making, the TRS helps customs agencies see opportunities for further improvement of the processes involved in realizing and accepting cargo.

THE MULTIDIMENSIONAL MISSION OF CUSTOMS AGENCIES

The mission of customs agencies typically translates into three core objectives: trade facilitation, fiscal revenue, and security and food safety. In outlining these objectives, this chapter presents evidence from research exploring how these goals can be pursued, as well as a detailed discussion of their analytical approach. The first cluster of research studies examines the role of customs and nontechnical barriers in trade facilitation. The second subsection provides an extensive discussion of customs as a source of fiscal revenue, with its associated challenges in fighting fraud and illegal practices, such as corruption. The last subsection presents studies that improve our understanding of how customs can promote product safety and ensure security.

Objective One: The Role of Customs in Trade Facilitation

Scholarly interest in customs research stems from its potential to serve as a tool for trade facilitation. What follows is an overview of the evidence to date. For example, Fernandes, Hillberry, and Alćantara (2021) evaluate Albanian reforms that sharply decreased the number of physical inspections of import shipments. There are clear indications that reduced inspections increase imports substantially. And there is no compelling evidence that the reforms gave rise to evasive behaviors. Similarly, for exports, Martincus, Carballo, and Graziano (2015) focus on time as a critical barrier to trade. Using a unique data set that consists of the universe of Uruguay's export transactions over the period 2002–11, they demonstrate that delays have a substantial negative impact on firms' exports. Furthermore, this effect is more pronounced for newcomers.

A seminal research paper that looks at the measurement of time as an instrumental component for the efficient functioning of customs is by Djankov, Freund, and Pham (2010). The authors examine how time delays affect export volumes. To measure time, the total export delay is considered. This means that the time delay does not include the time spent in a home country, on procedures, or in transit. It consists of the time spent when a container is at the border, transportation from the border to the post, and getting to the ship. The logic is that trade volumes can impact home country trade times; the effect on transit times abroad is likely negligible. Nevertheless, Djankov, Freund, and Pham (2010) estimate a difference gravity equation showing that each additional day a product is delayed prior to being shipped reduces trade by more than 1 percent. Delays have a relatively more significant impact on exports of time-sensitive goods, such as perishable agricultural products. Hence, it is important to measure and study how changes in customs operations can facilitate trade.

Objective Two: Customs as a Source of Fiscal Revenue

Another key policy objective for customs offices is increasing fiscal revenue. Several studies discuss interventions and propose mechanisms to improve local tax collection practices or incentivize inspectors posted in a given tax collection location. This is not surprising since there is evidence that trade tax revenues collected at the border constitute a large part of GDP, particularly for developing, low-income countries. Baunsgaard and Keen (2010) show, using a panel of 117 countries, that the inability to find alternative sources of revenue may hinder trade liberalization. Results suggest that high-income countries recovered from the revenue they lost during the past wave of trade liberalization, but the same does not apply to emerging markets, where recovery from trade liberalization is weaker.

Another major issue is the presence of tax evasion and corruption in customs administrations. Defining corruption following Bardhan (2006), Dutt and Traca (2010) show that in most cases, corrupt bureaucrats tax trade through either *extortion* or *evasion*. The former refers to a bureaucrat's demanding bribes from exporters for doing his duties, while the latter refers to a situation in which an exporter pays off a public servant to receive preferential treatment, like a lower tariff rate or the lowering of regulatory standards. Evasion may be trade-enhancing in an environment with high tariffs because it allows an exporter to effectively reduce the tariff rate by paying a bribe. However, in order to develop in a sustainable fashion, countries need to combat corruption more efficiently. In particular, developing economies are often in dire need of increasing state fiscal revenue via the rigorous implementation of customs rules, to be able to finance their development policies.

In seeking to increase tax revenues while reducing corruption, researchers and policy makers have been conducting experiments to identify optimal policies (Cantens, Raballand, and Bilangna 2019). One method that is relatively straightforward is mirror analysis, which compares the exports for a given country with the imports for its export client, or vice versa (WCO 2015). This approach is often limited by difficulties in obtaining detailed customs data. When implemented in Madagascar by Chalendard, Raballand, and Rakotoarisoa (2019), this method helped to identify the probability of fraud in the context of customs operations reforms.

Technology can help customs improve its mission while reducing fraud. In a natural experiment in Columbia, Laajaj, Eslava, and Kinda (2019) find that the computerization of imports led to an increase of six log points in the firm's value, with consequences for employment and tax collection. However, Chalendard et al. (2021) show that, through manipulation of the IT system, some customs inspectors and IT specialists were able to manipulate the assignment of import declarations. This was identified by measuring deviations from random assignments prescribed by official rules. Deviant declarations are found to be at greater risk of tax evasion, less likely to be deemed fraudulent, and cleared faster.

Another experiment analyzing policies to curb fraud was conducted in Madagascar (Chalendard et al. 2020). The authors investigated whether providing better information to customs inspectors and monitoring their actions could affect tax revenue and fraud detection. Results from the experiment show that monitoring incentivizes agents to scan more shipments, but they do not necessarily detect more fraud. Relatedly, Khan, Khwaja, and Olken (2019) propose a mechanism to improve the performance of public servants in collecting tax revenue, given their significance in enforcing and determining tax liabilities. Evaluating a two-year field experiment with 525 property tax inspectors in Pakistan, the authors stress the potential of periodic merit-based postings in enhancing bureaucratic performance.

Objective Three: Security and Food Safety

Customs authorities play an essential role as regulators of food safety and security. Although disruptions in total trade volume due to food safety are relatively rare (Buzby 2003), international organizations such as the WCO assist customs in the event of natural disasters and food crises. In June 2010, the WCO established an ad hoc working group to find ways for customs authorities to quickly react to such emergencies. The WTO, in turn, supports food security practices through the work of its Agriculture Committee and an Agricultural

Market Information System (AMIS) by a recommendation of the UN High-Level Task Force on the Global Food Security Crisis.

The role of customs authorities and their food security practices revolves around two fundamental issues: consumers do not always judge food security properly, and there are substantial differences between countries in terms of the regulation of food safety. The notion of trade security differs considerably in developed countries and developing ones (Diaz-Bonilla et al. 2000). Additionally, there are substantial risks of contamination due to trade. Ercsey-Ravasz et al. (2012) provide evidence that given the international agro-food trade network, the speed of potential contamination is extremely high because it is not possible to track the country of origin of different food products.

Safety is another key concern for customs authorities and is often associated with operations to reduce the illegal trade of products. The academic literature has documented how illegal trade in goods operates. In the European Union, Świerczyń ska (2016) provides a list of legal solutions implemented to sustain the twofold goal of customs authorities to combat the illegal trade in goods and, at the same time, decrease control measures that increase the cost of trade. In the Islamic Republic of Iran, Farzanegan (2009) estimates that a penalty rate on smuggling contributed to reducing illegal trade, using historical data from 1970 to 2002.

DATA REQUIREMENTS

Data Sources

Before delving into the definition of customs performance indicators, it is useful to explain the data requirements for measuring them. The first and fundamental source of data is the customs database. The most common system in low- and middle-income countries is the ASYCUDA. It is used by 100 countries and territories around the globe. This is a system designed by the United Nations Conference on Trade and Development (UNCTAD). Its purpose is to compile information pertaining to customs declarations, with customs office or border post information, frontline inspectors assigned to the case, potential changes in the clearance channel, irregularities, and final value assessments. In addition, this database lists goods by their characteristics, as well as the taxes and duties due. It was also designed with the goal of generating broad-ranging data for statistical and economic analysis of trade and customs performance. Box 14.2 illustrates the basics of the ASYCUDA's structure.

However, ASYCUDA data are rarely used outside of aggregate statistics of revenue collection. Most of the time, studies of time delays are based on a TRS. A TRS measures the time required for the release and/or clearance of goods, from the time of arrival at the border until the physical release of cargo. A TRS is conducted over a predefined period of time, during which several declarations are followed by the surveyor at some border posts. The surveyor observes all steps until release and makes note of the time spent and the associated costs. As noted by the WCO, the tool is useful to produce a pre-reform benchmark and needs to be repeated often to follow the evolution at a particular border post. However, intercountry comparisons are limited given differences in capacity and infrastructure (WCO 2018).

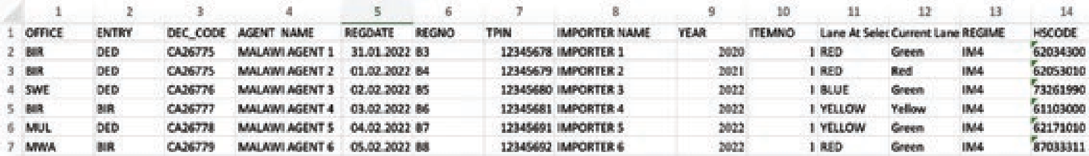
The information coming from the country databases is usually shared at an aggregated annual level with the UN Statistical Division. This information is treated and aggregated by the Harmonized System, typically using eight- or six-digit codes. The Harmonized System is a standardized classification of traded products based on numerical categories. The system is managed by the WCO and is regularly updated. Each product is described using eight digits.[1] It is used by customs authorities around the world to identify products when assessing duties and taxes and for gathering statistics. The vast range of product categories that customs agents regularly handle is illustrated by figure 14.2. It provides an overview of the total value of imports, classified according to 22 sections of the Harmonized System, across the 50 largest ports of entry in the United States.

BOX 14.2 ASYCUDA Data Structure

The Automated System for Customs Data (ASYCUDA) database is composed of a series of modules. Each module corresponds to a set of users. The customs broker module gives brokers secure access to the system to fill in a declaration. The customs office module covers declaration processing and is accessible to customs office agents. The accounting module is accessible to auditors only. The operations—registration of the declaration, assignment to an inspector, inspection results, change in value, clearance, and release—all have a time stamp associated with them, but merging this information in one report can be complicated because they are stored in different tables of the relational database.

A typical extract from ASYCUDA data thus contains information on the entry point for a specific declaration, the number of items declared, the agent and importer name, the year, and the registration date (see figure B14.2.1). ASYCUDA data also register *free on board* value—value outside insurance claims and ownership rights on the shipment—and value-added taxes (VAT), duties, and excise values for a chosen declaration, as well as exchange rate information and the currency with which payment for goods has been made (see figure B14.2.2).

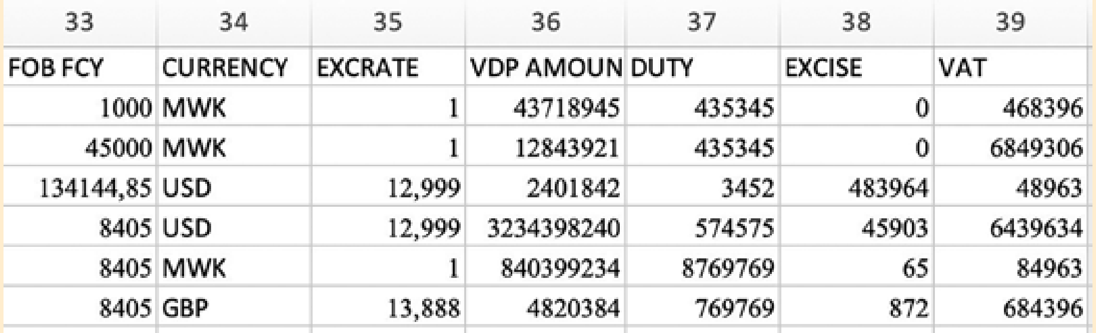
FIGURE B14.2.1 Example of an ASYCUDA Extract: Basic Variables



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	OFFICE	ENTRY	DEC_CODE	AGENT NAME	REGDATE	REGNO	TPIN	IMPORTER NAME	YEAR	ITEMNO	Lane At Selec	Current Lane	REGIME	HSCODE
2	BIR	DED	CA26775	MALAWI AGENT 1	31.01.2022	83	12345678	IMPORTER 1	2020	1	RED	Green	IM4	62034300
3	BIR	DED	CA26775	MALAWI AGENT 2	01.02.2022	84	12345679	IMPORTER 2	2021	1	RED	Red	IM4	62053010
4	SWE	DED	CA26776	MALAWI AGENT 3	02.02.2022	85	12345680	IMPORTER 3	2022	1	BLUE	Green	IM4	73261990
5	BIR	BIR	CA26777	MALAWI AGENT 4	03.02.2022	86	12345681	IMPORTER 4	2022	1	YELLOW	Yellow	IM4	61103000
6	MUL	DED	CA26778	MALAWI AGENT 5	04.02.2022	87	12345691	IMPORTER 5	2022	1	YELLOW	Green	IM4	62171010
7	MWA	BIR	CA26779	MALAWI AGENT 6	05.02.2022	88	12345692	IMPORTER 6	2022	1	RED	Green	IM4	87033311

Source: Automated System for Customs Data (ASYCUDA), United Nations Conference on Trade and Development.

FIGURE B14.2.2 Example of an ASYCUDA Extract: Duty, Excise, and Value-Added Taxes Variables



33	34	35	36	37	38	39
FOB FCY	CURRENCY	EXCRATE	VDP AMOUN	DUTY	EXCISE	VAT
1000	MWK	1	43718945	435345	0	468396
45000	MWK	1	12843921	435345	0	6849306
134144,85	USD	12,999	2401842	3452	483964	48963
8405	USD	12,999	3234398240	574575	45903	6439634
8405	MWK	1	840399234	8769769	65	84963
8405	GBP	13,888	4820384	769769	872	684396

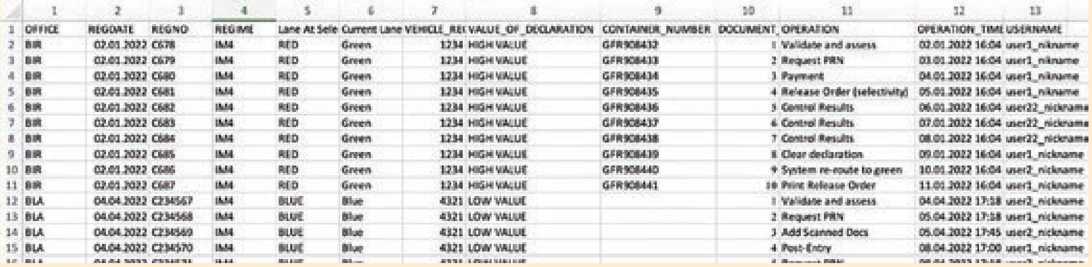
Source: Automated System for Customs Data (ASYCUDA), United Nations Conference on Trade and Development.

The typical time stamp data are associated with a particular action—such as a change in lane selectivity or in payments due, among others. Linking all the tables, one can extract tailored reports, as in figure B14.2.3, to create indicators of time delays between different actions depending on lane selectivity or type of declaration.

(continues on next page)

BOX 14.2 ASYCUDA Data Structure *(continued)*

FIGURE B14.2.3 Example of an ASYCUDA Extract: Time Stamps



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	OFFICE	REGDATE	REGNO	REGIME	Lane At Sele	Current Lane	VEHICLE_REI	VALUE_OF_DECLARATION	CONTAINER_NUMBER	DOCUMENT_	OPERATION	OPERATION_TIME	USERNAME
2	BIR	02.01.2022	C678	IM4	RED	Green	1234	HIGH VALUE	GFR908432	1	Validate and assess	02.01.2022 16:04	user1_nikname
3	BIR	02.01.2022	C679	IM4	RED	Green	1234	HIGH VALUE	GFR908433	2	Request PRN	03.01.2022 16:04	user1_nikname
4	BIR	02.01.2022	C680	IM4	RED	Green	1234	HIGH VALUE	GFR908434	3	Payment	04.01.2022 16:04	user1_nikname
5	BIR	02.01.2022	C681	IM4	RED	Green	1234	HIGH VALUE	GFR908435	4	Release Order (selectivity)	05.01.2022 16:04	user1_nikname
6	BIR	02.01.2022	C682	IM4	RED	Green	1234	HIGH VALUE	GFR908436	5	Control Results	06.01.2022 16:04	user22_nickname
7	BIR	02.01.2022	C683	IM4	RED	Green	1234	HIGH VALUE	GFR908437	6	Control Results	07.01.2022 16:04	user22_nickname
8	BIR	02.01.2022	C684	IM4	RED	Green	1234	HIGH VALUE	GFR908438	7	Control Results	08.01.2022 16:04	user22_nickname
9	BIR	02.01.2022	C685	IM4	RED	Green	1234	HIGH VALUE	GFR908439	8	Clear declaration	09.01.2022 16:04	user1_nickname
10	BIR	02.01.2022	C686	IM4	RED	Green	1234	HIGH VALUE	GFR908440	9	System re-route to green	10.01.2022 16:04	user2_nickname
11	BIR	02.01.2022	C687	IM4	RED	Green	1234	HIGH VALUE	GFR908441	10	Print Release Order	11.01.2022 16:04	user1_nickname
12	BLA	04.04.2022	C234567	IM4	BLUE	Blue	4321	LOW VALUE		1	Validate and assess	04.04.2022 17:58	user2_nickname
13	BLA	04.04.2022	C234568	IM4	BLUE	Blue	4321	LOW VALUE		2	Request PRN	05.04.2022 17:58	user1_nickname
14	BLA	04.04.2022	C234569	IM4	BLUE	Blue	4321	LOW VALUE		3	Add Scanned Docs	05.04.2022 17:45	user2_nickname
15	BLA	04.04.2022	C234570	IM4	BLUE	Blue	4321	LOW VALUE		4	Post-Entry	08.04.2022 17:00	user1_nickname

Source: Automated System for Customs Data (ASYCUDA), United Nations Conference on Trade and Development.

Another source of data is trader perception surveys. The focus of this type of survey is, as the name suggests, traders, importers, and exporters who directly engage in international trade. For example, traders might think that transport costs not related to border crossing are the most important costs faced when trading across borders, but these costs are unlikely to be shown in regular trade statistics. The burden of import or export certificates and clearance-associated costs is usually not represented either. The issue with these surveys is how to harmonize perception questions across countries to make sure they cover the same issues: what is experienced as a delay might be business as usual in another country, or traders might be reluctant to answer truthfully.

Finally, an emerging source of data is based on GPS trackers. This data source provides an objective time measure for border crossing and also captures the time spent on the road. These data can be used to observe the time spent at the border. Used in conjunction with time stamps, they show what share of time delays is attributable to customs operations as opposed to, for instance, difficulties linked to parking infrastructure. While these data are usually privately collected by firms providing transponders or insurers, some transport corridor authorities or public databases collect and provide these tracking data. One example of such a resource in Southern and Eastern Africa is administered by the World Bank's corridor team.[2]

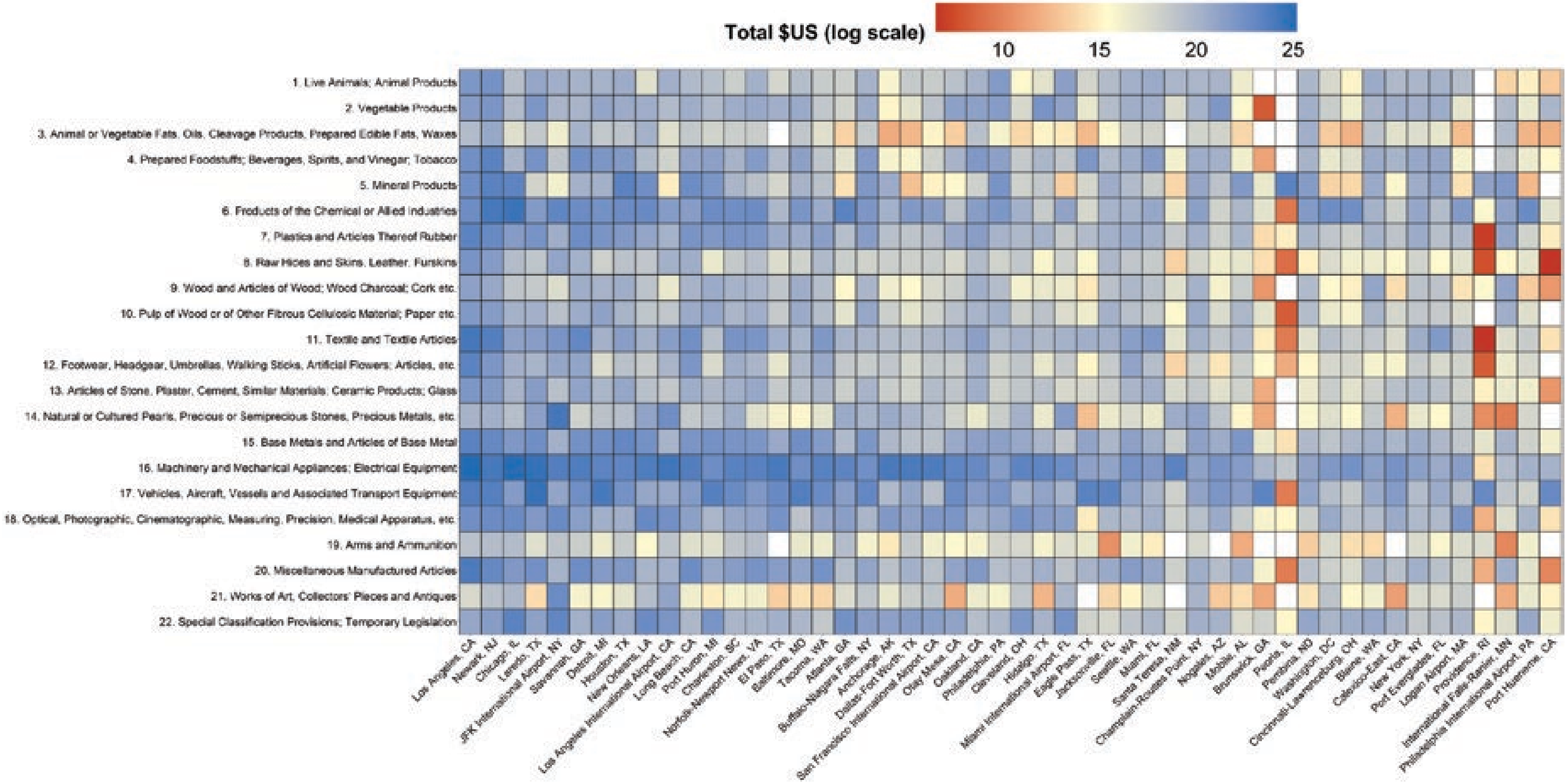
PERFORMANCE INDICATORS

While previous sections have discussed the types of data sources that can be used to measure customs performance, this section describes how customs data can be developed into indicators to measure and further the three key objectives of the multidimensional mission of customs: trade facilitation, revenue collection, and food safety and security.

Indicators for Trade Facilitation

Indicators related to trade facilitation usually focus on the time spent at the border and for clearance. This is part of the standard assessment of the WCO, the African Customs Union, and the TRS+ implemented by the World Bank. Of course, different border posts and different categories of goods will have different clearance times.

FIGURE 14.2 Value of Different Product Categories Imported to the United States for 50 Largest Ports of Entry, as Appraised by US Customs and Border Protection



Source: USA Trade Online, US Census Bureau: Economic Indicators Division.

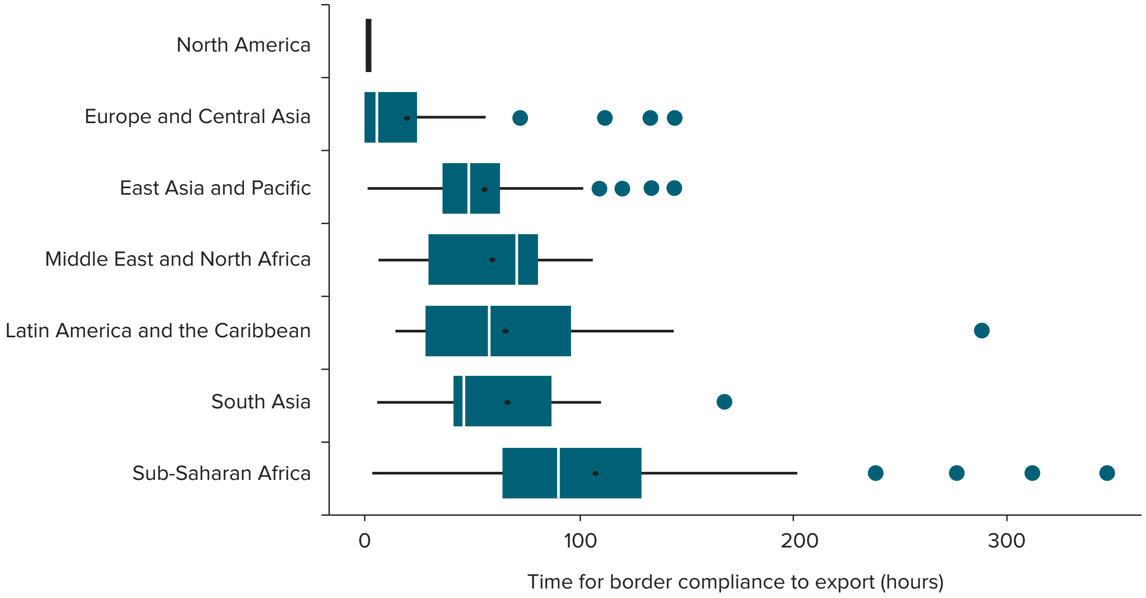
Note: Goods on the *y* axis are grouped according to the 22 sections of the Harmonized System. Some section labels are shortened due to space considerations. The *x* axis displays the 50 largest ports of entry in the United States by the total value of all goods imported, in decreasing order.

Figure 14.3 presents an example of a TRS indicator in the form of cross-country and regional disaggregation of border compliance times. Exporters across countries face vastly different times to process through customs. They are close to zero in the Northern American trade involving the United States and Canada, as well as in intra-EU trade. However, they increase more than threefold for Central Asian countries. On the other end of the spectrum, the largest delays are experienced by Sub-Saharan African exporters, where the mean border compliance time is 107 hours, and over 200 hours for several countries in Central Africa.

Not only are these processing times intrinsically heterogeneous, but the data used to measure them also paint a different picture of the customs process. The routinely collected time stamps from the customs database, the ASYCUDA or another, will show the date of the first submission and clearance. However, if the submission is made far in advance—for example, when arriving at the port, while the country itself is still far off—the time will be artificially long. In addition, as mentioned, if other agencies have to clear the goods while under customs custody, the time stamps will reflect a longer process. Indicators should take into account this heterogeneity in measurement approaches.

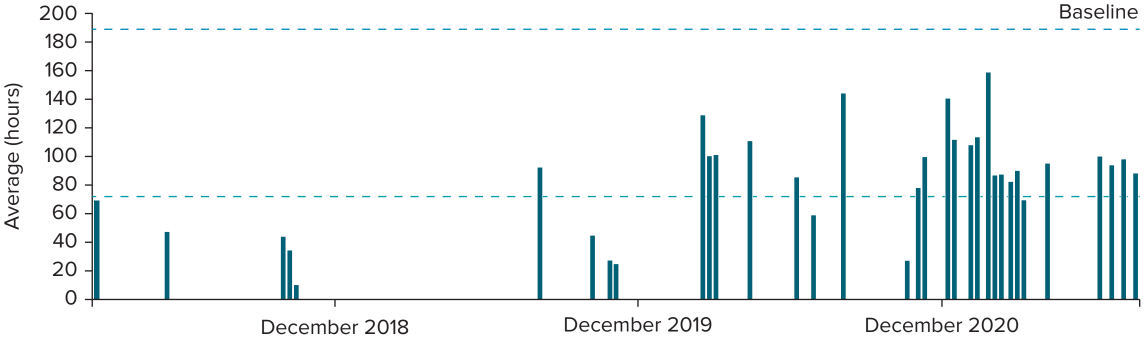
One possibility is, therefore, to look at the time necessary between the moment the frontline inspector is assigned to the declaration and the moment they clear it. While some agencies may delay the process by requesting additional inspection and clearances, this is less likely to be the case. In the ASYCUDA or other databases, this would correspond to the time difference between the time for assessment and the time at release. An example of such an indicator used for monitoring this time is depicted in figure 14.4. This figure displays the average time between the issuance of a release order and the issuance of a certificate of export at Malaba, on the Northern Corridor between Kenya and Uganda. The TRS follows a declaration at the border from when it is submitted to when the truck arrives and gives a snapshot of the border-crossing process at a moment in time, such that elements related to noncustoms delays can be isolated.

FIGURE 14.3 Border Compliance Times in Cross-Country and Regional View



Source: Doing Business database, World Bank.
Note: The component indicator is computed based on the methodology in the Doing Business 2016–20 studies. The boxes in the plot represent the interquartile range (IQR) of the variable—that is, the distance between the 25th and 75th percentiles in the distribution of respective values. The lines in the middle of the box represent the medians, whereas the dots represent means. The time is calculated in hours. The measure includes time for customs clearance and inspection procedures conducted by other agencies. If all customs clearance and other inspections take place at the port or border at the same time, the time estimate for border compliance takes this simultaneity into account.

FIGURE 14.4 Example of Indicators to Measure Performance: Transit Time on the Northern Corridor



Source: Original figure for this publication.

Finally, the same indicators can be based on surveys of traders to recover their perception of the delays, using a question to estimate how many days it takes between the moment a shipment reaches the border point and when it can be cleared from the border post. In Malawi, a survey is being conducted in this way. Early results show a reported average of two to three days once traders get notified their shipment is at the border.

Indicators for Revenue Collection

The revenue-collection objective focuses on how much revenue is collected at the border. This is intrinsically difficult to do—see box 14.3 on the problem of valuation—and, therefore, constructing the theoretical revenue that could have been collected requires considerable effort. Hence, this is something that the customs administration rarely does, unless misdeclaration or fraud is discovered. Otherwise, the declared value stays and the revenue collected is assumed to be the revenue that could have been collected by customs. However, not all misdeclaration or fraud is discovered. Hence, assuming that some of the incorrect declarations are missed, it is possible to look at the revenue that could have been collected if the items followed a similar price for other goods of the same class and origin. This is considered one of the acceptable valuation methods by the WTO. While the scholarly literature usually calls this *reference prices*, this clashes with the meaning of the reference prices used by the WTO: it is not an artificial set of prices but a comparison with similar goods' prices.

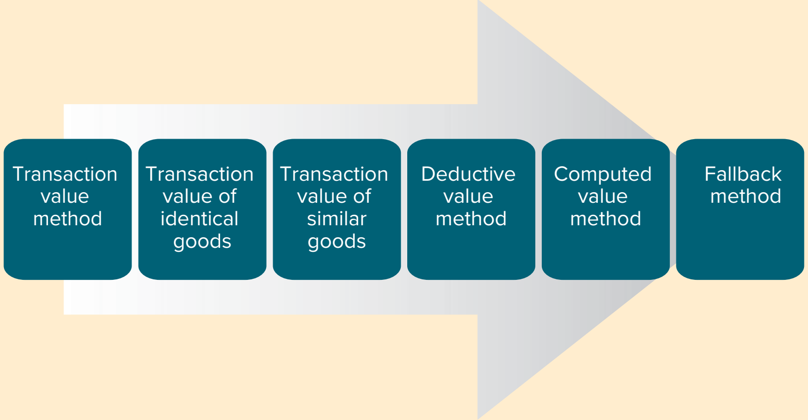
Evaluating the value of an item is intrinsically hard, as the inspector doesn't have precise information on the goods outside of what is listed on the declaration. The WTO agreement establishes rules for the valuation of imported goods that must be applied by all member countries. The WTO mandates using the transaction value supported by invoices and relevant documentation as the assessed value unless there is something missing or suspicion of fraud. In this case, the customs administration is authorized to use other valuation methods. The first method is using the transaction value of identical goods—same goods, same country of origin, same producer, whenever possible. The second method is using the transaction value of similar goods—same function or design, same country of origin, and whenever possible, same producer. Additional methods are outlined in figure B14.3.1. Customs is prohibited from using the same goods produced nationally as a comparison point, and from using arbitrary or fictitious values, such as minimal values or thresholds.

To refine this analysis, it is possible to use it in conjunction with the mirror gap: given the quantities of similar goods declared by the exporting country, how many are missing from the importing country import declarations and vice versa? The quantities declared for import and export in the origin country should be the same. This can give a rough idea of what revenues should be collected—or are missing—on either end. For an example of the use of such data for customs reform, see box 14.4. However, as mentioned earlier, these trade data sets are not updated as frequently as the customs data themselves. Hence, some of these gaps might be an artifact of the data. Another possibility is to reconcile the data at the

BOX 14.3 The Problem of Valuation

Evaluating the value of an item is intrinsically hard, as the inspector does not have precise information on the goods outside of what is listed on the declaration. The World Trade Organization (WTO) agreement establishes rules for the valuation of imported goods that must be applied by all member countries. The WTO mandates using the transaction value supported by invoices and relevant documentation as the assessed value unless there is something missing or suspicion of fraud. In this case, the customs administration is authorized to use other valuation methods. The first method is using the transaction value of identical goods—same goods, same country of origin, same producer, whenever possible. The second method is using the transaction value of similar goods—same function or design, same country of origin, and whenever possible, same producer. Additional methods are outlined in figure B14.3.1. Customs is prohibited from using the same goods produced nationally as a comparison point, and from using arbitrary or fictitious values, such as minimal values or thresholds.

FIGURE B14.3.1 World Trade Organization Valuation Methods, Arranged Sequentially



Source: Based on WTO Agreement on Customs Valuation (ACV) of 1994.

individual level, linking exporters' declarations from a country to another country's importers' declarations. This level of analysis can highlight value discrepancies and the potential mistakes or omissions of customs frontline agents.

Indicators for Food Safety and Security

There is relatively less work on safety because the data are harder to come by. The seized goods could indicate either an increase in customs activity or criminal activity. The TRS and ASYCUDA data can provide a good indication of risk management operations, both in terms of value recovery and physical inspection for safety. In Brazil, for example, the rate of physical inspections performed by customs was found to be around 2 percent during the most recent TRS (Receita Federal do Brasil 2020). However, 12 other government agencies were often involved in the process, granting licenses or permissions necessary for import. Around 60 percent of the declarations required involvement by another agency, whether or not the process required a physical inspection. The delays noted in the TRS process for Brazil thus reflect the need for other agencies' licenses and inspections. Another example is that, for goods under the jurisdiction of health authorities, around one-quarter to a third of the time is actually due to delays in paying the licensing fee.

BOX 14.4 Information and Customs Performance: The Case of Madagascar

The Republic of Madagascar is an island country lying off the southeastern coast of Africa (see map B14.4.1). In an experiment conducted in Madagascar, Chalendard et al. (2020) measure customs indicators and how they change when customs agents are given additional information. Madagascar is among the countries that rely heavily on customs and other import duties—16.9 percent of the total tax revenue going to Antananarivo proceeds from this source. At the same time, the performance of particular customs inspectors in Madagascar can be highly impactful because each inspector is responsible for a considerable value of import revenues. In the sample of Chalendard et al. (2020), every inspector handles around US$10 million in import revenues per year. Therefore, ensuring the good performance of its customs officials is a vital interest of the Malagasy authorities.

MAP B14.4.1 Location of Madagascar

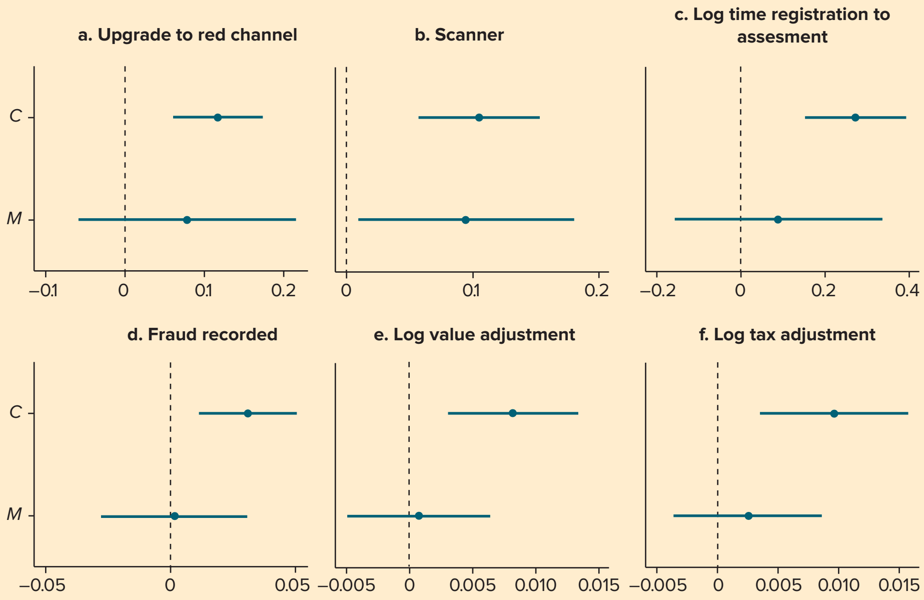
Source: World Bank.

(continues on next page)

BOX 14.4 Information and Customs Performance: The Case of Madagascar *(continued)*

Chalendard et al. (2020) investigate the role of information provision and monitoring in a randomized setting. One group of officials in their study was provided with a set of detailed risk-analysis comments on high-risk customs declarations (this group is labeled with *C* in figure B14.4.1). Officials in another group were told they would be more intensively monitored throughout a period of study (this group is labeled with *M* in the figure). Figure B14.4.1 shows that monitoring has an impact only on the increased frequency of customs officials' scanning containerized goods. In contrast, additional comments about high-risk declarations also lead the officials to more frequently upgrade inspections to the red channel and declare more cases of fraud detection and larger value adjustment. However, this also increases screening times and leads to only small improvements in tax collection, especially for declarations supposed to yield large tax revenues.

FIGURE B14.4.1 Changes in Malagasy Customs Officials' Performance



Source: Original figure for this publication.
Note: The label *C* on the *y* axis indicates the group of inspectors who were provided with comments; the label *M* indicates the group of inspectors who were told they would be monitored. *X*-axis units are regression coefficients.

CONCLUSION

What lessons for practice should be considered by the practitioner interested in exploring customs data for analytics?

First, an initial diagnosis through the TRS can provide a broad overview of the customs process. This can be done either at the beginning of a project or by using baseline data from past exercises. The TRS can provide useful indicators on what part of the clearance process suffers from a bottleneck. This is commonly done by the revenue administration before an overhaul of its process. This can be extended with a trader survey, which asks traders about the most sensitive aspects of the process, which are unlikely to be captured during the TRS. For example, the issues of speed money—or bribes to speed up the process—or other issues with any of the agencies involved might not be seen by the TRS surveyors but nevertheless influence traders' decisions to import or export.

Second, protocols to ensure data confidentiality while providing external access should be set in place. The anonymity of taxpayers is an important governmental concern, and some administrations are prevented from sharing taxpayer information with third parties. If protocols are set in place, these data can be shared while respecting these anonymity concerns, allowing practitioners and outside researchers to build customs performance indicators and opening the door to further research. These protocols include the deidentification of data whenever possible, such that researchers have access to deidentified tables only. This can be done via the hashing of the tables. Beyond security concerns, ASYCUDA tables might need to be merged and extracted, which can prove challenging in low-capacity settings. A useful solution is to support client engagement by requesting the data needed to build the basic indicators and assemble the data on a safe server. If necessary, the data can be deidentified by the client team based on a hashing code provided by the researcher, a procedure described on the World Bank's Development Impact Evaluation (DIME) Wiki.[3]

Third, stakeholders may resist additional measurement efforts. Some stakeholders may be reticent to use anything other than the TRS, as it is new and requires more effort from the ASYCUDA team. On top of that, while the TRS provides a narrative of the sources of delays, ASYCUDA data offer an often harsher view of the clearance process, as they also include steps that depend on the taxpayer—such as paying taxes. Because the ASYCUDA aggregates so much data, it can incorporate more outliers and influence the mean. This contrasts with the TRS, which is often done in a week, with the inspectors being aware of it. Researchers should thus expect discrepancies with the reported TRS, especially if the survey was done a while back. Thus, triangulating the different sources of data is important, as well as using the TRS results to comment on ASYCUDA-based indicators.

Finally, we suggest first investing in easy-to-produce indicators, such as revenue recovered and revenue recovered compared to similar products of the same type, as well as the easiest types of delays. These indicators should be triangulated with the TRS, if available, or with trader surveys. Further refinement of the indicators could include more precise measures of delays to distinguish tax compliance and the actions of customs, but these should be done once the more foundational indicators are measured and set in place. Of course, as outlined in the introduction, when measuring only select indicators of an organization with a multidimensional mission—such as customs—analysts need to remain cognizant of risks of effort substitution toward measurable indicators and to devise strategies to expand measurement to all core objectives of customs over time.

NOTES

The author would like to thank Iana Miachenkova for excellent research assistance and acknowledges the support of the Umbrella Trade Trust Fund for the Malawi Trade Facilitation Impact Evaluation.

[1] An example of an eight-digit description is 08051000, which corresponds to fresh oranges. Each product belongs, at the broadest level, to one of 22 Harmonized System sections. These are, however, not marked in the product code. Instead, each section is composed of one or more chapters, and the first two digits of the code refer to a specific chapter: in this case, chapter 08: "Edible Fruit and Nuts; Peel of Citrus Fruit or Melons." The next two digits stand for a heading within that chapter: heading 05: "Citrus fruit, fresh or dried." The following two digits stand for subheading 10: "Guavas, mangoes and mangosteens: Oranges." The last two digits can further specify more fine-grain divisions of product category if these exist. In this case, no further specification is indicated by 00.

[2] Their website is accessible at https://www.corridorperformancemonitoringsystem.com/geozone-route-catalogue.

[3] See DIME Wiki, s.v. "De-identification," last modified November 17, 2020, 20:10, https://dimewiki.worldbank.org/De-identification.

REFERENCES

Bardhan, Pranab. 2006. "The Economist's Approach to the Problem of Corruption." *World Development* 34 (2): 341–48. https://doi.org/10.1016/j.worlddev.2005.03.011.

Baunsgaard, Thomas, and Michael Keen. 2010. "Tax Revenue and (or?) Trade Liberalization." *Journal of Public Economics* 94 (9–10): 563–77. https://doi.org/10.1016/j.jpubeco.2009.11.007.

Buzby, Jean C., ed. 2003. *International Trade and Food Safety: Economic Theory and Case Studies.* Agriculture Economic Report 828. Washington, DC: Economic Research Service, US Department of Agriculture. https://www.ers.usda.gov/publications/pub-details/?pubid=41618.

Cantens, Thomas, Gaël Raballand, and Samson Bilangna. 2019. "Reforming Customs by Measuring Performance: A Cameroon Case Study." *World Customs Journal* 4 (2): 55–74.

Chalendard, Cyril, Alice Duhaut, Ana M. Fernandes, Aaditya Mattoo, Gaël Raballand, and Bob Rijkers. 2020. "Does Better Information Curb Customs Fraud?" CESifo Working Paper 8371, Munich Society for the Promotion of Economic Research, Munich. https://doi.org/10.2139/ssrn.3633656.

Chalendard, Cyril, Ana M. Fernandes, Gaël Raballand, and Bob Rijkers. 2021. "Corruption in Customs." CESifo Working Paper 9489, Munich Society for the Promotion of Economic Research, Munich. https://doi.org/10.2139/ssrn.3998027.

Chalendard, Cyril, Gaël Raballand, and Antsa Rakotoarisoa. 2019. "The Use of Detailed Statistical Data in Customs Reforms: The Case of Madagascar." *Development Policy Review* 37 (4): 546–63. https://doi.org/10.1111/dpr.12352.

Diaz-Bonilla, Eugenio, Marcelle Thomas, Sherman Robinson, and Andrea Cattaneo. 2000. "Food Security and Trade Negotiations in the World Trade Organization: A Cluster Analysis of Country Groups." TMD Discussion Paper 59, Trade and Macroeconomics Division, International Food Policy Research Institute, Washington, DC. https://www.ifpri.org/publication/food-security-and-trade-negotiations-world-trade-organization.

Djankov, Simeon, Caroline Freund, and Cong S. Pham. 2010. "Trading on Time." *The Review of Economics and Statistics* 92 (1): 166–73. https://doi.org/10.1162/rest.2009.11498.

Dutt, Pushan, and Daniel Traca. 2010. "Corruption and Bilateral Trade Flows: Extortion or Evasion?" *The Review of Economics and Statistics* 92 (4): 843–60. https://doi.org/10.1162/REST_a_00034.

Ercsey-Ravasz, Mária, Zoltán Toroczkai, Zoltán Lakner, and József Baranyi. 2012. "Complexity of the International Agro-Food Trade Network and Its Impact on Food Safety." *PLoS One* 7 (5): e37810. https://doi.org/10.1371/journal.pone.0037810.

Farzanegan, Mohammad Reza. 2009. "Illegal Trade in the Iranian Economy: Evidence from a Structural Model." *European Journal of Political Economy* 25 (4): 489–507. https://doi.org/10.1016/j.ejpoleco.2009.02.008.

Fernandes, Ana Margarida, Russell Hillberry, and Alejandra Mendoza Alcántara. 2021. "Trade Effects of Customs Reform: Evidence from Albania." *The World Bank Economic Review* 35 (1): 34–57. https://doi.org/10.1093/wber/lhz017.

Khan, Adnan Q., Asim Ijaz Khwaja, and Benjamin A. Olken. 2019. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." *American Economic Review* 109 (1): 237–70. https://doi.org/10.1257/aer.20180277.

Laajaj, Rachid, Marcela Eslava, and Tidiane Kinda. 2019. "The Costs of Bureaucracy and Corruption at Customs: Evidence from the Computerization of Imports in Colombia." Documento CEDE 2019-08, Centro de Estudios sobre Desarrollo Económico, Facultad de Economía, Universidad de los Andes, Bogotá.

Malawi Revenue Authority. 2019. *Malawi Time Release Study Report 2019.* Lilongwe: Malawi Revenue Authority. https://www.mra.mw/assets/upload/downloads/MalawiTimeReleaseStudyReport2019FN.pdf.

Martincus, Christian Volpe, Jerónimo Carballo, and Alejandro Graziano. 2015. "Customs." *Journal of International Economics* 96 (1): 119–37. https://doi.org/10.1016/j.jinteco.2015.01.011.

Receita Federal do Brasil. 2020. *Time Release Study: June 2020.* Brasília: Receita Federal do Brasil. https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/dados-abertos/resultados/estatisticascomercioexterior/estudos-e-analises/TRSReport.pdf.

Świerczyńska, Jolanta. 2016. "The Reduction of Barriers in Customs as One of the Measures Taken by the Customs Service in the Process of Ensuring Security and Safety of Trade." *Studia Ekonomiczne* 266: 212–22.

WCO (World Customs Organization). 2015. *Tools for Reducing Revenue Risks and the Revenue Gap: (I) Mirror Analysis Guide, Including Case Study (Cameroon).* Brussels: World Customs Organization. https://www.wcoesarocb.org/wp-content/uploads/2017/03/11-Mirror-analysis-guide-FINAL-EN.pdf.

WCO (World Customs Organization). 2018. *Guide to Measure the Time Required for the Release of Goods.* Version 3. Brussels: World Customs Organization. https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/instruments-and-tools/tools/time-release-study/trs-guideen.pdf?db=web.

Widdowson, David. 2007. "The Changing Role of Customs: Evolution or Revolution." *World Customs Journal* 1 (1): 31–37.

CHAPTER 15

Government Analytics Using Administrative Case Data

Michael Carlos Best, Alessandra Fenizia, and Adnan Qadir Khan

SUMMARY

Measuring the performance of government agencies is notoriously hard due to a lack of comparable data. At the same time, governments around the world generate an immense amount of data that detail their day-to-day operations. In this chapter, we focus on three functions of government that represent the bulk of its operations and that are fairly standardized: social security programs, public procurement, and tax collection. We discuss how public sector organizations can use existing administrative case data and repurpose them to construct objective measures of performance. We argue that it is paramount to compare cases that are homogeneous or to construct a metric that captures the complexity of a case. We also argue that metrics of government performance should capture both the volume of services provided as well as their quality. With these considerations in mind, case data can be the core of a diagnostic system with the potential to transform the speed and quality of public service delivery.

ANALYTICS IN PRACTICE

- Governments generate immense amounts of data that detail their day-to-day operations. These data can be repurposed to measure the performance of government agencies. Such data can provide objective comparisons of agency performance, allowing for an assessment of the quality of public administration across jurisdictions, regions, managers, and time.
- Such operational data provide objective records of bureaucratic performance. It is important to construct objective measures of organizational performance and individual performance rather than relying only on subjective evaluations such as performance appraisals.

Michael Carlos Best is an assistant professor in the Department of Economics, Columbia University. Alessandra Fenizia is an assistant professor in the Department of Economics, George Washington University. Adnan Qadir Khan is a professor at the School of Public Policy, London School of Economics.

- A prerequisite for constructing a comprehensive measure of performance for a public organization is obtaining a record of all the tasks undertaken by the organization. This may be difficult in practice because government agencies undertake a wide range of tasks, and they may not keep detailed records for all of them.
- One area of government activity where records are objective measures of performance and often relatively comprehensive is case management. Case management data are the records of responses by public officials to requests for public services or the fulfillment of public responsibilities. This chapter argues for the use of administrative data on the processing of cases by public officials as a monitoring tool for government performance and as a core input for government analytics. Relevant measures should capture both the *volume* and *quality* of cases processed.
- To construct an objective measure of performance using case data, one should ensure that cases are comparable to one another. This could entail comparing cases only within a homogeneous category or constructing a metric that captures the complexity of a case. For example, a social security claim that clearly meets the requirements of regulations and does not reference other data systems is a less complicated case to process than one in which there are ambiguities in eligibility and external validation is required. A corresponding metric of complexity might be based on the time spent on an "average" case of that type, allowing for complexity to be defined by the actual performance of public officials.

INTRODUCTION

In order to implement government policy, the apparatus of the state generates a vast trove of administrative databases tracking the deliberations, actions, and decisions of public officials in the execution of their duties. These data are collected in order to coordinate throughout a large, complex organization delivering a host of services to citizens and to preserve records of how decisions are reached to provide accountability for decisions made in the name of the public.

These data are not, typically, collected for the express purpose of measuring the performance of government officials. But as governments become more and more digitalized, these records contain ever-richer details on the work that is carried out throughout government. This presents an opportunity to repurpose existing data, and possibly extend its reach, to achieve the goal of measuring performance. In turn, such data can then be used to motivate government officials and hold them accountable. Ultimately, a greater ability to *measure* performance can help governments to *monitor* performance. This can improve efficiency in the public sector to deliver more and better services to citizens with the human and material resources the government has available.

Using administrative data has the distinct advantage that the data are already being collected for other purposes. As such, the additional costs of using them to measure performance are largely technical issues surrounding granting access to the data, protecting their confidentiality appropriately, and setting up the information technology (IT) infrastructure to perform statistical analysis on them. These obstacles are typically much simpler to overcome than the obstacles to launching new surveys of public officials or citizens to measure performance.

Set against this advantage, the primary disadvantage of using administrative data to measure performance is that they were not designed to be used for that purpose. As a result, a great deal of careful thought and work must go into how to repurpose the data for performance measurement. This involves thinking carefully about what outputs are being produced, how to measure their quantity and quality, and how to operationalize them within the constraints of the available data. Sometimes, this requires collecting additional data (either through a survey or from external sources) and linking them to the administrative data.

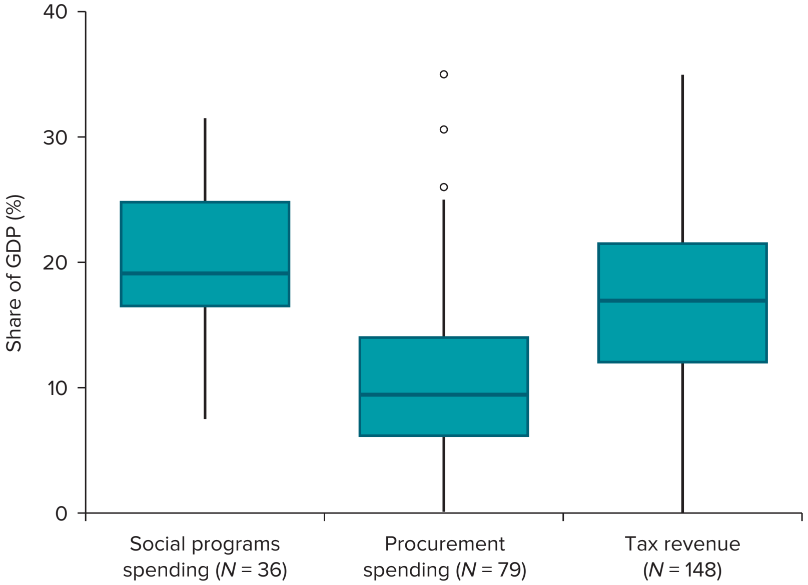
A large share of government operations involve the processing of case files or cases. Case data are the records of responses by public officials to requests for public services or the fulfillment of public responsibilities. A case file is typically a collection of records regarding an application. The nature of the applications varies widely. For one, thousands of claimants file applications every day to receive government services, such as welfare transfers, to gain access to government-sponsored childcare, or to obtain licenses and permits. Public sector organizations around the world initiate auctions to purchase goods and services from private sector suppliers. And millions of citizens and firms all over the globe file taxes every year.

In this chapter, we highlight examples from recent academic work trying to develop new methods to measure performance using administrative data on the processing of government casework. The academic papers provide a window into how similar data from public administrations around the world can be repurposed for analytical purposes.

Our examples cover three important realms of government operations—the delivery of social programs, the collection of taxes, and the procurement of material inputs—that together span a large part of what modern governments do. Figure 15.1 shows that spending on social programs and procurement and tax revenues jointly amount to more than 30 percent of a country's gross domestic product (GDP) on average. While there is some variation in the size of social programming, procurement spending, and tax revenues, these three functions of government represent a large share of government operations in all countries.

Since all governments engage in these activities, exploring potential alternative uses of the data generated in the process is of broad interest. In addition, operations in these areas are usually fairly standardized, tending to boost the quality of related data, which in turn can be used to generate more accurate insights. In all three cases, we highlight the importance of carefully specifying the outputs that are to be measured before undertaking an analysis, as well as how to conceptualize data quality.

FIGURE 15.1 Cross-Country Scale of the Three Sectors Discussed in the Chapter Relative to National Gross Domestic Product



Source: Original figure based on data from the Organisation for Economic Co-operation and Development (OECD) (social programs spending), the World Bank Development Indicators (tax revenue), and the World Bank Global Public Procurement Database (procurement spending).
Note: The box represents the interquartile range (IQR)—the distance between the 25th and 75th percentiles in the distribution of each variable. The line in the middle of the box represents the median. Whiskers—that is, the lines extending from the box—represent values lying within 1.5 of the IQR from the median. Outliers lying beyond that range are represented by dots, where one dot represents a country. The value of *N* shows the number of country-level observations in each column. GDP = gross domestic product.

We also provide some details on the technical methods used to operationalize these concepts and turn them into concrete performance measures and on how these performance measures are then used in the academic arena. In the conclusion, we discuss how policy makers can use these types of measures in other ways, as well as some important limitations to these approaches. The intention of our exposition of these cases is not to argue that the approach taken in the specific papers we review is optimal for every setting but rather to showcase a way to approach the analysis of government administrative case data.

CASE DATA IN ADMINISTRATION

A General Structure for the Analysis of Case Data

Government casework involves a series of standardized elements, each of which can be associated with a measure of the performance of public administration. Casework typically revolves around a set of protocols—perhaps standardized forms that applicants must fill in to apply for social security payments—that make common measures feasible. Cases are processed by government officials, again, frequently in a relatively standardized way.[1] For this reason, measures of performance can be used to judge how efficiently and effectively public officials worked through the relevant protocols. Case data are therefore made up of the records of cases and their processing, including details of the application or case and characteristics that can be analyzed. For example, in electronic case management systems, time and date stamps record exactly when cases were submitted, acted upon by officials, and then resolved. The speed of multiple stages of case processing can thus be easily calculated. Similarly, a decision is often made on a case and a response is sent to the applicant, such as a confirmation to a taxpayer that they have paid their taxes.

To use data on the processing of such cases to monitor and analyze government capabilities, we have to overcome two main challenges. Claims are diverse in how challenging or "complex" the associated case is. A case that involves a claim where a claimant clearly meets the required criteria is less complex than one in which eligibility is ambiguous on one or more margins. In some cases, evaluating the claimant's eligibility may be fairly straightforward, involving verification of the veracity of a few supporting documents provided by the applicant. In other cases, it may require the officer to request access to a separate archive to pull the claimant's records.

Thus, we first have to construct a common measure of task complexity that allows us to compare claims of different types. Second, we must ensure that any such measure is not easy to manipulate by government staff and is as objective as possible. For example, to minimize the risk of manipulation of these types of metrics, the tracking of claims should be done by a centralized computer system. Allowing employees to self-report their output and log it onto a computer may leave room for opportunistic behavior aimed at artificially inflating the measure of output. Employees may report processing a higher volume of claims or more complex claims than they actually did. One way around this is to complement electronic records with field observations of a representative sample of tasks at hand that is regularly updated. This approach minimizes the risk that the performance measures become outdated or disentangled from the constantly evolving work environment of public officials.

With these pieces in place, case data can be a source of government analytics. These data can provide objective comparisons of agency performance, allowing for an assessment of the quality of public administration across jurisdictions, regions, managers, and time. Rather than comparing simple output across offices, it is often useful to compare a measure of output per worker (or per unit of time). These measures capture the productivity of the average worker (or the average hour) in each office and are not affected by differences in office size. For instance, larger offices typically process a larger quantity of various cases by virtue of having more workers devoted to back-office operations. However, the fact that larger offices process more cases does not necessarily imply that they are more productive.

A major limitation of evaluating the performance of public sector offices based solely on output or productivity is that these measures reflect production volume and do not capture the quality of the service provided. For example, imagine an official who rubber-stamped applications for a claim. Looking only at production volumes, the official would seem very productive. However, the officer has de facto awarded welfare transfers to all claimants regardless of their eligibility status. Conditioning on, or including in analysis, a measure of complexity would not adjust for the official's quality of service. Rather, a separate metric related to the quality of decision-making must be constructed to address this concern.

Extending Analytics Insights

Government agencies can significantly increase the impact of existing administrative data by going beyond a basic analysis of the administrative data they hold. First, they can build assessments of the accuracy of their case data. For example, governments can collect additional data on the accuracy of tax assessment, say, from randomly selected tax units, which will enable them to construct more comprehensive performance measures of tax staff and establish more credible audit and citizen grievance redress mechanisms.

Second, the digitization of case data allows for the use of machine-learning and artificial intelligence algorithms to create better valuation measures, such as to detect clerical and other types of error, flag suspected fraud cases, or classify taxpayer groups in a (more) automated fashion. Further discussion of this topic is provided in chapter 16 of *The Government Analytics Handbook*, and a case study of a similar system is provided in case study 9.2 in chapter 9.

Authorities can also make anonymized case data publicly available, and this increased transparency can enable whistleblowing and peer pressure mechanisms. As one of the following case studies shows, there is precedent for doing this in Pakistan, where the entire tax directory for federal taxes has been published annually for the past decade.

Finally, case data can be integrated with political data to create better measures of politicians' performance at the local government level and thus enhance political accountability. For example, updates to cadastre records, which are crucial for accurate property valuations for tax purposes, were found to be crucially linked to electoral pressures on local officials in Brazil (Christensen and Garfias 2021).

The rest of this paper presents case studies that highlight the analysis and use of case data, focusing on measuring case volume, complexity, and quality, as well as describing ways to strengthen this analysis by linking to other data sources.

SOCIAL SECURITY CLAIMS DATA

Social security claims data include records relating to old-age programs and social welfare programs, such as unemployment benefits, maternity leave, and subsidies to the poor. Most governments around the world already regularly collect claims data in an electronic format. For this reason, these data can be repurposed to perform quantitative analysis to better understand the performance of the social security system overall, the challenges facing individual public sector offices, and design solutions to address them.

In this section, we discuss a recent academic paper that uses detailed claims data from the Italian Social Security Agency (ISSA) to construct a measure of the performance of public offices and evaluate the effectiveness of ISSA managers. Fenizia (2022) exploits the rotation of managers across sites to estimate the productivity of public sector managers. This study finds significant heterogeneity in the effectiveness of these managers: some managers are very productive and improve the performance of the offices where they work, while others do not. The increase in office productivity brought about by talented managers is mainly driven by changes in personnel practices.

A case in this setting is the process of assessment by a social security officer of the validity of a claim for social security payments to an individual. A key advantage to studying the ISSA is that the tasks employees perform are fairly standardized, and the agency keeps detailed records of all applications and welfare transfers. This allows Fenizia (2022) to construct a comprehensive measure of performance that encompasses all the activities employees perform.

The obvious volume-based measure of productivity in this context is the number of social security claims of a particular type processed by an office in a particular time period divided by the full-time equivalent of workers of that office during that time. Map 15.1 describes how this measure varies across Italian regions, showcasing how such data can be used in government analytics. The figure indicates which regions are more productive than others and thus where investments might be needed in the quality of management or staff.

The first concern with analyzing this sort of data is that some cases may be more complex to process than others. In many settings, it is possible to measure only the output stemming from a subset of activities rather than the associated complexity. In these settings, the measure of performance only reflects the activities being measured and may be harder to interpret. For example, imagine that an agency performs two types of tasks: task A is observable, but task B is not. The measure of performance will reflect only the output from task A. If this measure were to decline over time, this could be driven by a worsening of performance in the agency overall or by the fact that resources had been reallocated from task A to task B. The following section discusses how to construct a measure of complexity using the time spent on an "average" case of a particular type.

The second concern is that production volumes do not reflect the quality of the service provided. After the discussion of complexity, the following section evaluates the strengths and weaknesses of two proxies of quality of service that can be derived from claims data.

Complexity

Virtually all government agencies that administer old-age and welfare programs process a variety of different claims. While it is relatively straightforward to keep track of the number of incoming and processed claims, it is more challenging to construct a measure of performance for public offices that can be meaningfully compared across sites.

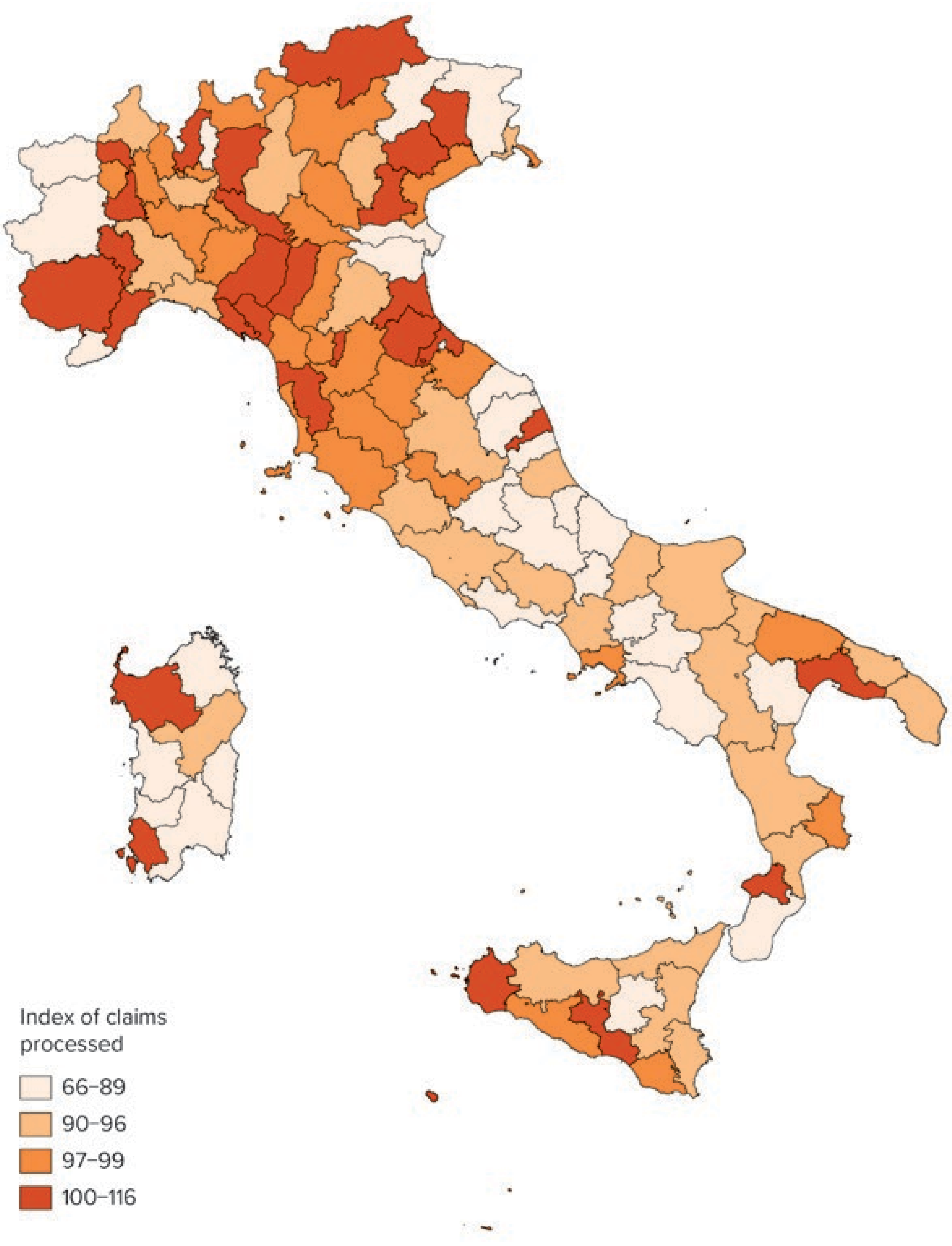
A naive solution might involve counting the number of claims processed by each office. Despite being simple and transparent, this measure suffers from a major draw-back: it does not take into account task complexity. Some claims might be very quick to process, while others might require a lot of time and resources. As mentioned above, in some cases, officers have simply to verify that the documentation provided by the applicant is complete and up-to-date. In other cases, officers may have to acquire further documentation from their internal archives or from other entities. If different offices process a different mix of paperwork, simply counting the number of claims processed would not correctly reflect differences in task complexity across sites. The naive metric would overstate the performance of offices that process simpler claims relative to those that process more sophisticated paperwork.

A solution is to use a complexity-adjusted measure of claims processed. For example, the ISSA constructs a measure of output for public offices that combines the number of claims processed by each site with a measure of their complexity. Specifically, the ISSA has grouped all claim types into more than 1,000 fine categories. Each category is constructed to group highly comparable claims that are equally complex. Each category is assigned a weight representing how much time it should take to process that specific claim type.

Figure 15.2 illustrates the distribution of expected processing time (that is, weights) for the most common types of pensions and welfare transfers. The expected processing time for most pensions ranges between 31 and 38 minutes, with a median of 30 minutes. The expected processing time is more variable for welfare transfers, reflecting the fact that these products are much more heterogeneous. Most of these claims take between 17 and 41 minutes to process, with a median processing time of 28 minutes.

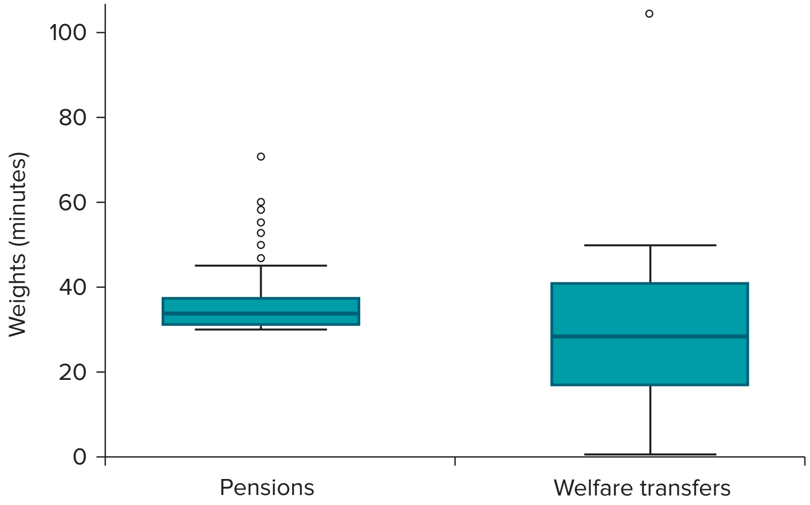
Importantly, the ISSA complexity-adjustment formula uses objective weights as opposed to subjective scores. As part of the ISSA quality control department, there is a team devoted to measuring weights and keeping them up-to-date. To construct the weight for product *v*, this team selects an excellent, an average, and a mediocre office and picks a representative sample of product *v* claims from each office. Then the team visits each site and records the amount of time each employee took to process each claim. The weight is constructed by averaging all measurements across employees and offices, and it represents the time spent processing an "average" case of that type. The same weights apply to all offices at a given time to ensure that all offices are evaluated using the same standards. Weights can change in response to a technological

MAP 15.1 Variations in Productivity of Processing Social Security Cases, Subregions of Italy



Source: Fenizia 2022, using Italian Social Security Agency data.
Note: The key refers to the number of social security claims of a particular type that are processed by an office in a particular time period divided by the full-time equivalent of workers of that office during that time.

FIGURE 15.2 Expected Processing Time for Most Common Types of Pensions and Welfare Transfers, Italy



Source: Fenizia 2022, based on Italian Social Security Agency data.
Note: This figure illustrates the distribution of the expected processing time (that is, weights) for the most common types of pensions and welfare transfers. The box represents the interquartile range (IQR)—the distance between the 25th and 75th percentiles in the distribution of the weights. The line in the middle of the box represents the median. Whiskers represent values lying within 1.5 of the IQR from the median. Outliers are represented by circles.

improvement, if the time required to process a specific claim shortens, or when the paperwork associated with a claim changes.

The ISSA also ensures that the weights are measured accurately and that there are no opportunities for arbitrage. For example, if processing product *b* takes, on average, 10 minutes, and the weight associated with it is equal to 20 minutes, officers have an incentive to process as many *b* claims as possible. By doing so, they artificially increase the output of the office. Similarly, if product *b* is assigned a weight of 5 minutes when it takes 10 minutes on average to process, officers may be inclined to give priority to other claim types. To minimize arbitrage, the ISSA tracks backlog by product. If the backlog for a given product increases (decreases) across several offices, this may be an indication that the weight associated with it is too low (high). Therefore, the ISSA reevaluates the weights associated with the products that experienced large changes in backlog.

The weights are used to aggregate the number of claims of different types processed by each office *i* into a single output measure. The aggregation consists in multiplying the number of product *v* claims processed (c_{vi}) with their corresponding weight (w_v) and then summing across categories as follows:

$$\text{output}_i = \sum_{\upsilon=1}^{V} c_{\upsilon i} \times w_{\upsilon} \tag{15.1}$$

This output metric reflects the *theoretical* amount of time that it *should* have taken to process the claims that were effectively processed.

Although the procedure described above is largely specific to the ISSA and its mandate related to social security, similar measures are used in manufacturing firms across the world. These measures are especially popular in the garment sector, where the standard minute value (SMV) has become the standard.

Quality

In the case of social security claims, a straightforward measure of the quality of service provided is the error rate (that is, the fraction of claims that were processed incorrectly). There are two types of mistakes: a government agency may erroneously give a beneficiary money, or it may erroneously deny a transfer. Keeping track of the errors found when a denied beneficiary files an appeal only catches the latter type of mistake.

This is why, to construct a comprehensive measure of the office error rate and discourage fraudulent behavior, it is paramount to regularly audit a random subset of claims processed by each office.

Agencies may combine the error rate with a second proxy for quality: timeliness in claim processing. While timeliness is an important dimension of the service provided, a drawback of this measure is that it is mechanically correlated with office productivity. In other words, holding constant other office characteristics, offices that process claims quickly are also those that deliver a high level of output.

Extending Administrative Data

Alternative approaches to measuring the quality of service provided include using subjective customer satisfaction ratings. The main challenge when using customer ratings is that the subset of customers who choose to provide feedback is not representative because customers with more extreme (either positive or negative) opinions are more likely to provide a review (Schoenmueller, Netzer, and Stahl 2020).[2]

This limitation can potentially be overcome by conducting regular surveys of a representative sample of all customers. The US Social Security Administration (SSA) implements a range of such surveys both by phone and in person across different groups of customers (online users of SSA services, callers to the SSA phone number, and visitors to SSA field offices). Although it does not eliminate the possibility that the most (un)happy customers will be more likely to respond to a survey invitation, it does mitigate this concern by targeting a sample of all customers. An indication of average customer satisfaction can also be obtained from surveys conducted by third parties. For example, the different dimensions of services provided by US government agencies are regularly evaluated as one of the topics covered in the American Customer Satisfaction Index (ACSI), which is used to measure the general satisfaction of American customers with various goods and services.

PROCUREMENT RECORDS

Public procurement—the purchase of goods and services by governments from private sector suppliers—is one of the core functions of the state. Public procurement represents a large portion of governments' budgets and a sizeable fraction of the economy, representing 12 percent of world GDP (Bosio et al. 2022). Public procurement also tends to be a highly technocratic, legalistic process generating large volumes of documents recording every step of the procurement purchase in great detail. These data are generated and recorded as part of the government's procedures in order to uphold the transparency and accountability of the procurement process—core goals of a well-functioning procurement system. However, these same data, either by themselves or in conjunction with additional data, can also be used to measure the performance of the officials and public entities in charge of carrying out procurement.

This section builds on chapter 12 of the *Handbook* to showcase how the indicators outlined in detail there can be considered as individual case data and to showcase the benefits of complementing administrative data with experimental variation. Here, we discuss two recent academic papers that develop methods to use administrative databases on public procurement to construct measures of procurement performance. Best, Hjort, and Szakonyi (2017) use detailed procurement data from Russia spanning all procurement transactions between 2011 and 2016 to construct measures of procurement performance. They show that there are big differences across purchases in how effectively the purchase is carried out, which can be attributed in roughly equal proportions to the effectiveness of the individual civil servants tasked with procurement and the effectiveness of the public entities they represent. They also show how procurement policy can be tailored to the capacity of the implementing bureaucracy in order to offset weaknesses in implementation capacity.

Bandiera et al. (2021) use existing procurement data from Punjab, Pakistan, and supplement it with additional data collected from purchasing offices to construct performance measures. This paper is an example of how a randomized controlled trial (RCT) can be used to complement government administrative data to better understand the impact of personnel policies and other aspects of public administration. By introducing experiments into government, such initiatives amplify the potential benefits of the analysis of public administration data. Bandiera et al. (2021) show that granting procurement officers additional autonomy to spend public money improves procurement performance, especially when the officers' supervisors caused significant delays in approvals.

Complexity

A procurement case may be characterized by a differing number of features of the good or service being procured and by a wide range of requirements on those features. For example, the procurement of pencils has far fewer features for the procurement officer to assess than the procurement of a vehicle. For this reason, when comparing the productivity of procurement agents and agencies, it is important to have a measure of the nature of the procurement cases they have to process.

Best, Hjort, and Szakonyi (2017) use publicly available administrative data from Russia to construct measures of performance based on public procurement. Since 2011, a centralized procurement website has provided information to the public and suppliers about all purchases.[3] They use data from this website on the universe of electronic auction requests, review protocols, auction protocols, and contracts from January 1, 2011, through December 31, 2016. The data cover 6.5 million auction announcements for the purchase of 21 million items. However, purchases of services and works contracts are highly idiosyncratic, making comparisons across purchases impossible, so they are dropped from the sample, resulting in a sample of 15 million purchases of relatively homogeneous goods.

To use these data to measure performance, there are two key challenges to overcome. First, the main measure of performance uses prices paid for identical items, requiring precise measures of the items being procured. Second, prices are not the only outcome that matters in public procurement, and so they use administrative data to construct measures of spending quality as well.

The main measure of performance used in Best, Hjort, and Szakonyi (2017) is the price paid for each purchase, holding constant the precise nature of the item being procured. Holding constant the item being procured is crucial to avoid conflating differences in prices paid with differences in the precise variety of item being procured. As described in more detail in appendix F.1, they use the text of the final contracts, in which the precise nature of the good purchased is laid out, to classify purchases, using text analysis methods, into narrow product categories within which quality differences are likely to be negligible.

The method proceeds in three steps. First, the goods descriptions in contracts are converted into vectors of word tokens. Second, they use the universe of Russian Federation customs declarations to train a classification algorithm to assign goods descriptions a 10-digit Harmonized System product code and apply it to the goods descriptions in the procurement data. Third, for goods that are not reliably classified in the second step, either because the goods are nontraded or because their description is insufficiently specific, they develop a clustering algorithm that combines goods descriptions that use similar language into clusters similar to the categories from the second step. Just as in the case of claims data discussed in the preceding section, here it can be seen that the key issue in analyzing case complexity is comparing "apples to apples." Although many procedures in public administration come with a set of standardized procedures, the actual complexity of each task is highly variable, and, therefore, its accurate evaluation is the key to understanding the performance of public officials. To achieve this, highly detailed metrics might be required. In the case of ISSA claims data, this metric was a continuous weight—time judged as necessary to complete a specific task based on primary data obtained during field observations in various social security offices. In the case of procured goods, the metric used is categorical but narrow enough to avoid classifying goods of a different nature as comparable. It is also not based on field measurements but rather relies on secondary data from descriptions in Russian Federation customs declarations and advanced classification algorithms.

Quality

Sourcing inputs at low prices is the primary goal of public procurement, but it is not the only outcome that matters.[4] Successful procurement purchases should also be smoothly executed. Contracts should not need to be unduly renegotiated or terminated, and goods should be delivered as specified, without delays. These outcomes reflect the quality of public spending and may conflict with the goal of achieving low prices. If this problem is severe, then it would be misleading to deem purchases effective if they achieve low prices but this is offset by poor performance on spending quality.

To address this, Best, Hjort, and Szakonyi (2017) build direct measures of spending quality by combining a number of proxies for the quality of the nonprice outcomes of a procurement purchase. Specifically, they use six proxies: the number of contract renegotiations, the size of any cost overrun, the length of any delays, whether the end user complained about the execution of the contract, whether the contract was contested and canceled, and whether the product delivered was deemed to be of low quality or banned for use in Russia because it didn't meet official standards.

To summarize spending quality in a single number, they take the six quality proxies and create an index of spending quality y_i as the average of the six proxies after standardizing each one to have mean zero and standard deviation one, as follows (Kling, Liebman, and Katz 2007):

$$y_i = \tfrac{1}{6}\sum_{k=1}^{6}(y_i^k - \overline{y}^k)/\sigma^k. \qquad (15.2)$$

This is done because the proxies are in different units of measurement and because some proxies will be more variable than others. For a deviation in a proxy to be judged as "large," this approach conditions it on what other deviations we observe for that proxy. For example, there may be many complaints but very few contract cancellations. In that case, one would want to weight a cancellation more heavily than a complaint, in accordance with how rare, and thus significant, a cancellation is. With these measures in hand, Best, Hjort, and Szakonyi (2017) show that there are big differences across purchases in how effectively the purchase is carried out. They also decompose these differences into the part that can be attributed to the effectiveness of the individual public servants working on the purchase and the part that can be attributed to the agency that is receiving the item being purchased. They show that both contribute roughly equally to the differences in effectiveness and that together they explain around 40 percent of the variation in government performance. They also show how these differences in effectiveness contribute to differences in how policy changes manifest in performance outcomes.

They argue that policy that is tailored to the capacity of the implementing bureaucracy can offset overall weaknesses in implementation capacity. The analysis provides an example of how the analytics of public administration can lead to direct implications for the policies that govern it.

Extending Administrative Data

Existing administrative data can sometimes prove insufficient to measure productivity in public administration, but the required information can nevertheless be obtained by the targeted data collection efforts of governments and researchers. Bandiera et al. (2021) use administrative data from Punjab, Pakistan, to measure procurement performance. In their case, the existing administrative data were not sufficiently detailed to implement their preferred method of performance measurement, and so they worked with the government to design and implement an additional administrative database capturing detailed information about the products being purchased by procurement officers.

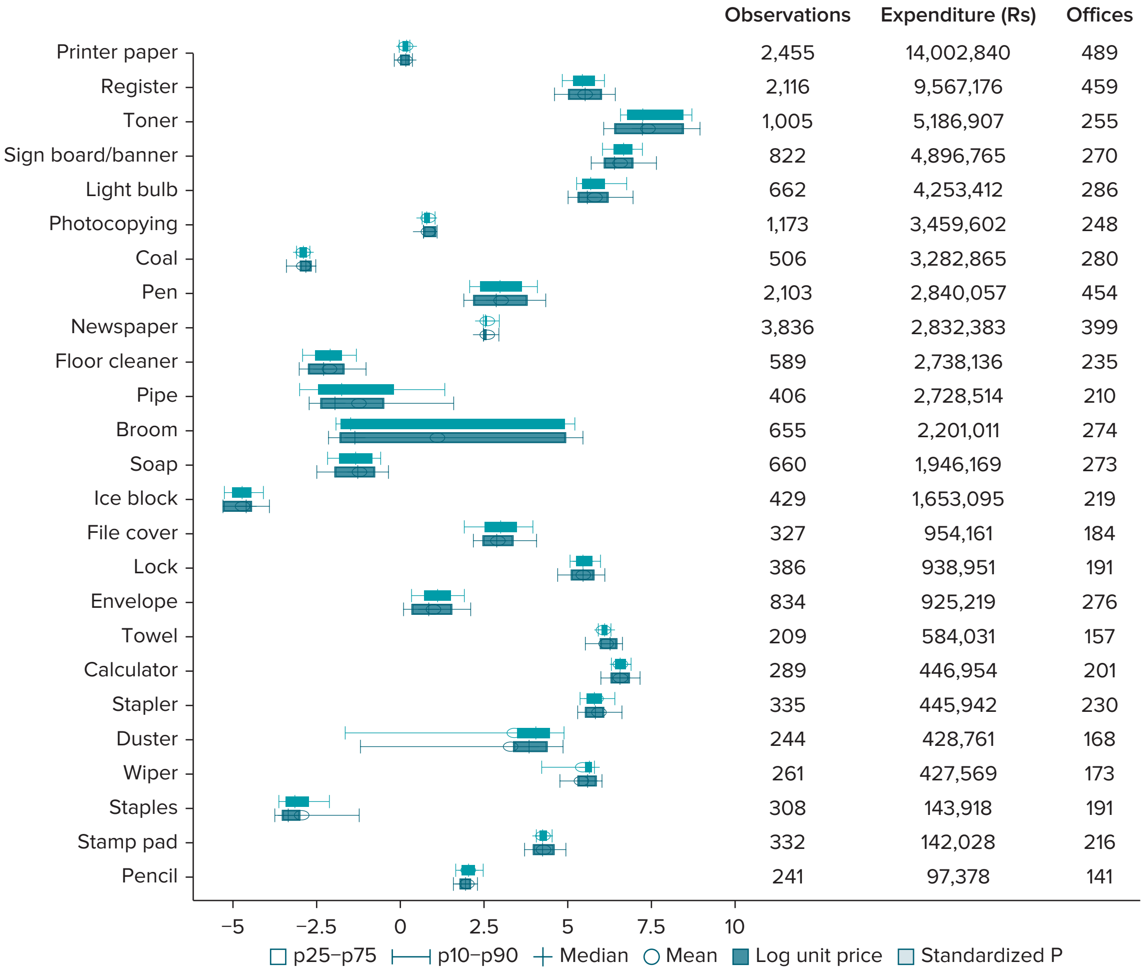
The government of Punjab considers the primary purpose of public procurement to be ensuring that "the object of procurement brings value for money to the procuring agency" (PPRA 2014). In line with this, they developed a measure of bureaucratic performance that seeks to measure value for money in the form of the unit prices paid for the items being purchased, adjusted for the precise variety of the item being purchased.

They proceed in two steps. First, they restrict attention to homogeneous goods for which it is possible to gather detailed enough data to adequately measure the variety of the item being purchased. Second, they partner with the Punjab IT Board to build an e-governance platform—the Punjab Online Procurement System (POPS). This web-based platform allows offices to enter detailed data on the attributes of the items they are purchasing. Over one thousand civil servants were trained in the use of POPS, and the departments they worked with required the offices in the study to enter details of their purchases of generic goods into POPS. To ensure the accuracy of the data, offices were randomly visited to physically verify the attributes entered into POPS and collect any missing attributes required.

After the POPS platform was run for the two-year project and the data the officers entered were cleaned, the analysis data set consists of the 25 most frequently purchased goods—a total of 21,503 purchases. Dropping the top and bottom 1 percent of unit prices results in a data set of 21,183 observations.[5] Figure 15.3 shows summary statistics of the purchases in the POPS data set. The 25 items are remarkably homogeneous goods, such as printing paper and other stationery items, cleaning products, and other office products. While each individual purchase is small, these homogeneous items form a significant part of the procurement: generic goods are 53 percent of the typical office's budget in the sample.

To use these data on prices to measure procurement performance, they again need to be able to compare purchases of exactly the same item. The goods in the analysis are chosen precisely because they are extremely

FIGURE 15.3 Summary Statistics on the 25 Most Commonly Purchased Goods in the Punjab Online Procurement System, 2014–16



	Observations	Expenditure (Rs)	Offices
Printer paper	2,455	14,002,840	489
Register	2,116	9,567,176	459
Toner	1,005	5,186,907	255
Sign board/banner	822	4,896,765	270
Light bulb	662	4,253,412	286
Photocopying	1,173	3,459,602	248
Coal	506	3,282,865	280
Pen	2,103	2,840,057	454
Newspaper	3,836	2,832,383	399
Floor cleaner	589	2,738,136	235
Pipe	406	2,728,514	210
Broom	655	2,201,011	274
Soap	660	1,946,169	273
Ice block	429	1,653,095	219
File cover	327	954,161	184
Lock	386	938,951	191
Envelope	834	925,219	276
Towel	209	584,031	157
Calculator	289	446,954	201
Stapler	335	445,942	230
Duster	244	428,761	168
Wiper	261	427,569	173
Staples	308	143,918	191
Stamp pad	332	142,028	216
Pencil	241	97,378	141

Source: Bandiera et al. 2021.
Note: The figure displays summary statistics for purchases of the goods in the purchase sample. The figure summarizes the log unit prices paid for the goods, the number of purchases of each good, and the total expenditure on the good (in rupees) in the sample.

homogeneous. Nevertheless, there may still be some differentiation across items, and so Bandiera et al. (2021) use four measures of the variety of the goods being purchased. First, they use the full set of attributes collected in POPS for each good. This measure has the advantage of being very detailed but comes at the cost of being high dimensional. The three other measures reduce the dimensionality of the variety controls. To construct the second and third measures, they run hedonic regressions to attach prices to each of the goods' attributes. They run regressions of the form

$$p_{igto} = \mathbf{X}_{igto}\lambda_g + \rho_g q_{igto} + \gamma_g + \varepsilon_{igto}, \tag{15.3}$$

where p_{igto} is the log unit price paid in purchase i of good g at time t by office o, q_{igto} is the quantity purchased, γ_g are goods fixed effects, and $\mathbf{X}_{igto}$ are the attributes of good g.

The second, *scalar*, measure of goods variety uses the estimated prices for the attributes $\hat{\lambda}_g$ to construct a scalar measure $\upsilon_{igto} = \sum_{j \in A(g)} \hat{\lambda}_j X_j$, where $A(g)$ is the set of attributes of item g. The third, *coarse*, measure studies the estimated $\hat{\lambda}_g$s for each item and partitions purchases into high- and low-price varieties based on the $\hat{\lambda}_g$s that are strong predictors of prices in the control group. Finally, the *machine-learning* measure develops a variant of a random forest algorithm to allow for nonlinearities and interactions between attributes that regression (15.1) rules out. Appendix F.2 provides further details. This effort provides a way to homogenize the type and quality of goods on which government analytics can be performed.

Extending Administrative Data

Extending administrative data does not only imply the collection of further data. Rather, it can imply an extension in the methods used for analysis. A particularly powerful extension is to embed an RCT into data collection. In this way, the data collected reflect groups that have received a policy intervention purely by chance. Comparing measures of case processing between these groups thus allows one to look for differences that are due purely to the policy intervention and not some other mediating factor.

With the above performance measure in hand, Bandiera et al. (2021) perform just such a field experiment in which one group of procurement officers is granted greater autonomy over the procurement process (essentially reducing the amount of paperwork required and streamlining the preapproval of purchases by government monitors), another group is offered a financial bonus based on their performance, and a third group is offered both. By embedding an experiment into their analysis, they find that granting autonomy causes a reduction in prices by around 9 percent, illustrating that in settings where monitoring induces inefficiency, granting frontline public servants more autonomy can improve performance.

PROPERTY TAX DATA

Taxation is critical for development; however, tax systems throughout the developing world collect substantially less revenue as a share of GDP than their counterparts in the developed world.[6] Weak enforcement, informational constraints, and tax morale provide some explanation. This is also true for property taxes, despite their greater visibility and contribution to local public goods. Khan, Khwaja, and Olken (2016, 2019) describe a long collaboration with the Excise and Taxation Department in Punjab, Pakistan, on different mechanisms for incentivizing property tax collectors—through performance-pay and performance-based postings. Once again, these papers provide insight into how case data, and in this subsection, case data related to the taxation of individual properties, can be combined with experimental variation to improve the measurement of and insights related to the performance of public administration.

The urban property tax in Punjab is levied on the gross annual rental value (GARV) of the property, which is computed by formula. Specifically, the GARV is determined by measuring the square footage of the land and buildings on the property, and then multiplying by standardized values from a valuation table depending only on property location, use, and occupancy type. These valuation tables divide the province into seven categories (A–G) according to the extent of facilities and infrastructure in the area, with a different rate for each category. Rates further vary by residential, commercial, or industrial status, whether the property is owner occupied or rented, and location. Taxes are paid into designated bank branches.

The Excise and Taxation Department collects regular administrative data. Each quarter, as part of their normal reporting requirements, tax inspectors report their revenue collected during the fiscal year cumulatively through the end of the quarter, which they compile from tax-paid receipts retrieved from the national bank. In addition, they report their total assessed tax base both before exemptions are granted and after exemptions have been granted. These records are compiled separately for current-year taxes and arrears.

In theory, the performance of property tax collectors should be easy to monitor because the key measure of performance, tax revenue, is less subject to measurement issues than other areas of government work. However, in practice, measurement related to the performance of tax inspectors faces many challenges. It is not ex ante obvious how much credibility to give to reported tax revenues at the unit level in Punjab, given that the tax department's internal cross-checks are usually run at a higher level of aggregation. Given multiple reporting templates with slightly varying assumptions in use in the province, all officers can overstate the revenues they have generated without their misreporting being effectively detected. Similarly, the continuously evolving environment in which tax collectors operate introduces further complications to understanding relative performance. For example, the boundaries of tax administrative units (called "tax circles" in Punjab) are continuously being changed, and tax circle boundaries do not overlap with the boundaries of political units.

For these reasons, gaining a coherent measure of the taxes collected and the performance of tax officials and agencies can be a challenging task. Since reported tax revenues are a function of the tax base, exemption rate, and collection rate, comparing collection alone is not reflective of performance. Finally, given concerns over multitasking, performance on revenue collection has to be matched with performance on nonrevenue outcomes, especially on the accuracy of tax assessments and citizen/taxpayer satisfaction.

Complexity

Rather than generating novel measures of complexity or clever systems for categorization, as in the social security and procurement cases, complexity was made more homogeneous in this context by standardizing the reporting templates and matching boundaries. The approach to ensuring a common level of complexity in case data can thus be relatively simple in some settings.

Quality

In the work in Punjab, to ensure the accuracy of the administrative data unit level, an additional reverification program was instituted, involving cross-checking the department's administrative records against bank records. This entailed selecting a subset of circles, obtaining the individual records of payment received from the bank for each property, and manually tallying the sum from the thousands of properties in each circle to ensure that it matched the department total.

The project found virtually no systematic discrepancies between the administrative data received from the department and the findings of this independent verification; the average difference between the independent verification and what the circle had reported revealed underreporting of −0.28 percent, or about zero. In general, if rightly conducted, data diagnostics and audits can ensure the accuracy of administrative data, help flag issues before policy decisions are based on such data, and align incentives for truthful reporting.

Extending Administrative Data

Once again, Khan, Khwaja, and Olken (2016) showcase the power of introducing experimentation into government analytics. They ran a large-scale field experiment in which all property tax units in the province were experimentally allocated into one of three performance-pay schemes or a control. After two years, incentivized units had 9.4 log points higher revenue than controls, which translates to a 46 percent higher growth rate. The revenue gains accrued due to a small number of properties that became taxed at their true value, which was substantially more than they had been taxed at previously. The majority of properties in incentivized areas, in fact, paid no more taxes but instead reported higher bribes. The results are consistent with a collusive setting in which performance pay increases collectors' bargaining power over taxpayers, who either have to pay higher bribes to avoid being reassessed or pay substantially higher taxes if collusion breaks down. The paper shows that performance pay for tax collectors has the potential to raise revenues but might come at a cost if it increases the bargaining power of tax collectors relative to taxpayers.

The paper also highlights the limitations of relying on existing administrative data for areas where multitasking can be a concern and where existing systems capture only some aspects of performance—for instance, administrative data usually capture revenue collection but not nonrevenue outcomes, like the accuracy of tax assessments and taxpayer satisfaction. To capture these nonrevenue outcomes, as well as owner and property characteristics to examine any heterogeneous effects, Khan, Khwaja, and Olken (2016) conduct a random property survey.

The survey is based on two distinct samples. The first, the "general population sample," consists of roughly 12,000 properties selected by randomly sampling five GPS coordinates in each circle and then surveying a total of five (randomly chosen) properties around that coordinate. These properties therefore represent the picture for the typical property in a tax circle. The second sample, referred to as the "reassessed sample," consists of slightly more than 4,000 properties (roughly 10 per circle) sampled from an administrative list of properties that are newly assessed or reassessed. These properties were then located in the field and surveyed. The purpose of this survey was to oversample the (few) properties that experience such changes each year in order to examine the impacts on such properties separately.

These survey data are used to determine the GARV of the property, which is the main measure of a property's tax value before exemptions and reductions are applied and, unlike tax assessed, is a continuous function of the underlying property characteristics and, hence, much more robust to measurement error. To measure under- or overtaxation, the "tax gap" is determined as

$$\text{Tax Gap} = \frac{(GARV_{Inspector} - GARV_{Survey})}{(GARV_{Inspector} + GARV_{Survey})}. \tag{15.4}$$

Taxpayer satisfaction is measured based on two survey questions about the quality and results of interactions with the tax department. Accuracy is measured as one minus the absolute value of the difference between the GARV as measured by the survey and the official GARV, as measured from the tax department's administrative records, divided by the average of these two values.

Khan, Khwaja, and Olken (2019), in a subsequent project, examine the impact of performance-based postings in the same setting and rely primarily on administrative data. They propose a performance-ranked serial dictatorship mechanism, whereby public servants sequentially choose desired locations in order of performance. They evaluate this using a two-year field experiment with 525 property tax inspectors. The mechanism increases annual tax revenue growth by 30–41 percent. Inspectors who the model predicts face high equilibrium incentives under the scheme indeed increase performance more. These results highlight the potential of periodic merit-based postings in enhancing bureaucratic performance.[7]

CONCLUSION

In this chapter, we have discussed how public sector organizations can use administrative data to construct measures of performance across three important realms of government operations: the delivery of social security programs, the procurement of material inputs, and tax collection. Agencies whose primary work consists of processing claims can use their existing records to construct a measure of the volume of services provided (that is, a complexity-adjusted index of claims processed) and proxies for the quality of service (that is, the error rate and timeliness in claim processing). Similarly, government organizations purchasing goods and services can leverage their existing procurement records to construct two measures of performance: the price paid for homogeneous goods and an index of spending quality that combines information on the number of contract renegotiations, cost overrun, the length of delays, complaints, contract cancellations, and whether the product delivered did not meet minimum quality standards. When the administrative data are not sufficiently detailed, governments can develop a platform that standardizes the procurement process and collects the underlying data. Finally, taxation authorities can construct reliable measures of tax revenue by standardizing the process through which tax collectors report the taxes they have collected and instituting a set of automatic checks to ensure data accuracy.

Better measures of performance may help governments improve the effectiveness of public service provision. For example, policy makers can use these performance measures to identify the best-performing offices, learn about "best practices," and export them to the underperforming sites. Government agencies can also use these metrics to identify understaffed sites and reallocate resources toward them. Moreover, governments can *monitor* the performance of public offices and intervene promptly when a challenge arises. Finally, they can use these measures to design incentive schemes aimed at improving public service provision.

Administrative records typically include large amounts of data, and performing statistical analyses on them involves some practical challenges. First, not all public sector organizations employ workers who have the technical skills to repurpose data for performance measurement and carry out the statistical analyses. This challenge can be addressed by partnering with external researchers experienced in this area. Second, governments should take all necessary steps to protect data confidentiality when granting access to their internal records. This may involve anonymizing data to protect the identity of the subjects being studied, transferring data through secure protocols, and ensuring that data are stored on a secure server. In some cases, government organizations may also invest in their own IT infrastructure, such as a large server to store data and a set of workstations through which researchers can access anonymized administrative records.

The approaches described in this chapter have the potential to promote evidence-based policy making within government organizations, resulting in more effective public service provision. An example of such impacts comes from the tax analytics work described in this chapter. Over the course of the research collaborations discussed, the Punjabi tax authorities began to digitize and geocode unit data at the property level. This database is now being regularly updated. Tax notices are now issued through an automated process, supporting tax staff still responsible for field work and for updating property status—for example, covered area, usage (residential, commercial, or industrial), and status (owner-occupied or rented)—and for providing the information relevant for deciding on exemptions. This reduces the human interface between tax collectors and taxpayers. It allows for more sophisticated analysis and data visualization conducted at more granular levels—for example, at neighborhood levels—in real time. The data are now being used by the Urban Unit in Pakistan, different government agencies, and by analysts to address a range of policy questions.

NOTES

1. Many governments put effort into standardizing case data to increase their capacity to undertake analytics. For example, a number of countries have introduced the Standard Audit File for Tax (SAF-T) for all taxpayers, a protocol for the data collected on each case (OECD 2017).
2. To evaluate the performance of government agencies, it is also important to account for the fact that many government agencies also have front-office operations. Measuring productivity in any customer-facing setting is challenging. While some agencies use customer ratings, the ISSA measures front-office output using the inputs—the amount of time employees spend on front-office duties. Thus, the measure bluntly captures the value of staffing the office without adjusting for the number of customers served or the complexity of their demands. An agency may also consider constructing a measure of front-office operations analogous to the one used for claim processing. The additional challenge is that allowing front-office employees to self-report their output may incentivize employees to misreport the activities they undertake.
3. The website can be accessed at http://zakupki.gov.ru/.
4. Article 1 of Federal Law 94 (FZ-94), which transformed Russia's public procurement system in 2005, declares the aim of procurement to be the "effective, efficient use of budget funds." The law also introduced minimum price as the key criterion for selecting winners for most types of selection mechanisms (Yakovlev, Yakobson, and Yudkevich 2010).
5. The majority of these outliers are the result of officers adding or omitting zeros in the number of units purchased.
6. According to 2018 World Bank data, tax revenue as a share of gross domestic product stood at 11.4 percent in lower-middle-income countries, compared to 15.3 percent in high-income countries.
7. In ongoing work with the tax authorities and the local government, Khwaja et al. (2020) examine strengthening the social compact between citizens/taxpayers and the government by linking the (property) taxes citizens pay with the services they receive at the neighborhood level. Combining administrative data from tax and municipal agencies at the neighborhood level provides local-level measures of variation in public service provision, tax and fiscal gap, administrative performance, and sociopolitical dynamics.

REFERENCES

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats." *Quarterly Journal of Economics* 136 (4): 2195–242. https://doi.org10.1093/qje/qjab029.

Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w23350.

Bosio, Erica, Simeon Djankov, Edward Glaeser, and Andrei Shleifer. 2022. "Public Procurement in Law and Practice." *American Economic Review* 112 (4): 1091–117. https://doi.org/10.1257/aer.20200738.

Christensen, Darin, and Francisco Garfias. 2021. "The Politics of Property Taxation: Fiscal Infrastructure and Electoral Incentives in Brazil." *The Journal of Politics* 83 (4): 1399–416. https://doi.org/10.1086/711902.

Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. https://doi.org/10.3982/ECTA19244.

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2016. "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors." *The Quarterly Journal of Economics* 131 (1): 219–71. https://doi.org/10.1093/qje/qjv042.

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2019. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." *American Economic Review* 109 (1): 237–70. https://doi.org/10.1257/aer.20180277.

Khwaja, Asim Ijaz, Osman Haq, Adnan Qadir Khan, Benjamin Olken, and Mahvish Shaukat. 2020. *Rebuilding the Social Compact: Urban Service Delivery and Property Taxes in Pakistan.* 3ie Impact Evaluation Report 117. New Delhi: International Institute for Impact Evaluation (3ie). https://doi.org/10.23846/DPW1IE117.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119. https://doi.org/10.1111/j.1468-0262.2007.00733.x.

OECD (Organisation for Economic Co-operation and Development). 2017. *The Changing Tax Compliance Environment and the Role of Audit*. Paris: OECD Publishing. https://doi.org/10.1787/9789264282186-en.

PPRA (Punjab Procurement Regulatory Authority). 2014. *Punjab Procurement Rules 2014*. No. ADMN (PPRA) 10–2/2013. Lahore: Government of the Punjab Services General Administration Department. https://ppra.punjab.gov.pk/system/files/Final%20Notified%20PPR-2014%20%28ammended%20upto%2006.01.2016%29.pdf.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl. 2020. “The Polarity of Online Reviews: Prevalence, Drivers and Implications.” *Journal of Marketing Research* 57 (5). https://doi.org/10.1177/0022243720941832.

Yakovlev, Andrei, Lev Yakobson, and Maria Yudkevich. 2010. “The Public Procurement System in Russia: Road toward a New Quality.” 4th International Public Procurement Conference, Seoul, Republic of Korea, August 26–28. http://ippa.org/images/PROCEEDINGS/IPPC4/01ComparativeProcurement/Paper1-9.pdf.

CHAPTER 16

Government Analytics Using Machine Learning

Sandeep Bhupatiraju, Daniel Chen, Slava Jankin, Galileu Kim, Maximilian Kupi, and Manuel Ramos Maqueda

SUMMARY

The use of machine learning offers new opportunities for improving the productivity of the public sector. The increasing availability of public sector data and algorithmic approaches provides a conducive environment for machine learning for government analytics. However, the successful deployment of machine-learning solutions requires first developing data infrastructure of the required quality to feed these algorithms, as well as building the human capital necessary to develop them. Ethical principles regarding the use of machine-learning technologies must be defined and respected, particularly for the justice system. This chapter provides an overview of potential applications of machine learning in the public sector and in the justice system specifically, as well as the necessary steps to develop them sustainably and ethically. It then analyzes the case of machine-learning deployment in India to illustrate this process in practice.

ANALYTICS IN PRACTICE

- Machine learning is fundamentally a methodological approach: it defines a performance indicator and uses collected data to train an algorithm to improve this indicator. Because of this relatively broad definition, machine learning includes different algorithms and may be applied in a variety of domains, from payroll fraud detection to court rulings. This flexibility requires practitioners to make key design decisions: what kind of performance indicator will be used? What training data and algorithm will be deployed? These decisions may substantially alter the machine-learning algorithm's results. Making these decisions thus requires close collaboration between machine-learning engineers, domain experts, and the agencies that will use the technology.

The authors' names are listed alphabetically. Sandeep Bhupatiraju is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Daniel Chen is a senior economist at DIME. Slava Jankin is a professor of data science and public policy at the Hertie School. Galileu Kim is a research analyst at DIME. Maximilian Kupi is a PhD candidate at the Hertie School. Manuel Ramos Maqueda is a research analyst at DIME.

- Machine learning can leverage large amounts of administrative data to improve the functioning of public administration, particularly in policy domains where the volume of tasks is large and data are abundant but human resources are constrained. Governments generate large amounts of administrative data on an almost-daily basis, but these data are seldom used to improve the production function of public administration. At the same time, civil servants are constrained in the amount of time they can dedicate to complete tasks—as well as in the amount of information they have readily available. Machine learning can process large amounts of administrative data, structuring them around performance indicators that algorithms are well suited to optimize. For example, machine-learning algorithms can be trained, using procurement data, to predict whether new, incoming contracts are irregular or not, at a scale and speed which far exceed human capacity.
- While machine learning can offer efficiency gains in public administration, governments need to be aware of their role in generating and using measurements in public administration, as well as their ethical responsibilities. Machine-learning algorithms require extensive data and measurements on both citizens and civil servants, but these data are often collected without their consent. As a result, governments should be transparent about how these data are being used to enhance public administration and how these technologies are being used to affect public administration. Care should be taken not to reproduce biases, such as racial or gender discrimination, in the machine-learning algorithms.
- To fully reap the benefits of machine learning, governments must undertake long-term investments in data infrastructure and human capital. Before machine learning is implemented, investments in data infrastructure must take place. Data quality needs to be improved and data pipelines must be developed to train the algorithm. Additionally, specialized machine-learning engineers must be hired and trained to implement the technology in the public sector. These investments are costly and require long-term planning: governments should not expect machine-learning technologies to be developed overnight.
- Machine-learning experts should collaborate with subject matter experts to guide how machine-learning technology will benefit the civil servants who will use it. Besides the technical knowledge to develop and operate machine-learning technologies, substantial levels of domain and political expertise, as well as awareness about potential ethical and legal pitfalls, are necessary to ensure the effective use of a machine-learning solution. For example, if the machine-learning technology is meant to assist judges in reducing racial bias, judicial experts and judges themselves should be consulted to ensure that relevant performance indicators and data are used. Including civil servants in the development of the solution also facilitates the adoption of the new technology.
- Machine learning is not a panacea, and practitioners should be aware of the limitations of the approach to fully leverage its benefits. Algorithms are limited to what is measurable by data, and performance indicators may reflect the bias of machine-learning engineers and the data themselves. Improving a particular performance indicator may not necessarily be the best way of achieving a policy goal. As a result, machine-learning applications should not be considered as a substitute for policy making but as a tool to complement and enhance decisions made by government agencies and their civil servants.

INTRODUCTION

Machine learning is a discipline that focuses on the development of computer systems (machines) that, through the analysis of training data, can improve their performance (learn) (Jordan and Mitchell 2015). Recent advances in data collection and processing power have expanded opportunities for machine-learning applications in a variety of fields. Advances in machine learning have brought tangible benefits in the worlds of business and medicine and in large-scale systems more generally (Brynjolfsson and McAfee 2017). However, this growth in opportunity has often led to excessive optimism about what machine learning can accomplish, as well as a tendency to downplay the potential steep costs of deploying machine-learning technologies (Chen and Asch 2017).

We start our discussion by offering a general definition of machine learning. What distinguishes machine learning from other methodological approaches is the definition of a learning problem under a statistical framework. Following Jordan and Mitchell (2015, 225), we define a learning problem as a "problem of improving some measure of performance when executing some task, through … training experience." The following example illustrates a machine-learning approach. Suppose a government is interested in reducing irregularities in its payroll system. One measure of performance would be the proportion of irregular paychecks correctly identified by the machine. The training experience—or data—would be a collection of paychecks manually classified by payroll analysts (see case study 2 in chapter 9). The learning problem would thus define a statistical model that learned how best to predict irregular paychecks by being trained on historical payroll data.[1]

In this chapter, we discuss applications of machine learning to public administration. We outline the data infrastructure and human capital requirements for developing machine-learning applications, as well as potential complementarities with public servant surveys. As noted in chapter 9 of *The Government Analytics Handbook*, the foundational step for any form of data analytics—including machine learning—is the development of a robust data infrastructure. We also highlight ethical concerns regarding the development and deployment of machine-learning applications, which relate directly to the discussion in chapter 6. Despite our focus on machine learning, we consider the broader shift to a data-driven and statistically informed culture—regardless of the implementation of algorithms—to be often already sufficient for bringing substantial benefits to public service delivery. These benefits include organizational changes, data literacy, and performance monitoring. With the transition to data-driven policy making, machine-learning applications are a natural next step in government analytics, automatically leveraging data to improve the performance of public administration through well-defined performance metrics.

Following this broader discussion of machine learning in public administration, we focus on machine-learning applications in justice systems. Within public administration, the justice system generates a large number of case rulings linking legal cases and actors (training data) to ideally fair rulings (performance). Thus, a potential machine-learning learning problem is: can we identify and reduce racial bias in court rulings by training an algorithm on a collection of case rulings? Instead of defining what a fair ruling is, we might define what an unfair ruling is. For instance, the decision of the judge should not be influenced by extraneous factors, such as the time of day or the race of the defendant. By identifying cases in which the decision has been influenced by such biases, a machine-learning model can potentially identify and ultimately prevent unfair rulings. This approach thus allows machine learning to improve the quality and fairness of judicial decisions (Ramos-Maqueda and Chen 2021). Despite this promise, limited data literacy and statistical training inhibit applications of data analytics in general—and machine learning specifically—in the judiciary.

Machine learning is not a panacea: it requires significant investments before any of its benefits come to fruition. As we detail throughout this chapter, machine-learning applications require investments in data infrastructure and the development of the human capital necessary to develop and deploy machine-learning algorithms. Undertaking these investments—and often long cycles of development—requires resources and long-term horizons before the benefits of the approach become apparent. Additionally, ethical concerns regarding embedded racial or gender biases in training data highlight how technologies can inadvertently reproduce the same human biases they were designed to eliminate. Thus, initial optimism regarding the revolutionary potential of machine-learning approaches should be balanced by a recognition of their limitations (Cross 2020).

Practitioners have much to gain from deploying machine learning in public administration. Defining performance metrics and automating the training of algorithms through large-scale data can improve the functioning of public administration—particularly when oriented toward well-defined tasks with an abundance of quantitative data. Performance improvement can come simply from embedding a data-driven approach to government functioning. There is not always a need for sophisticated approaches in machine learning to make progress. In fact, machine learning generally comes only after more basic steps have been taken on data management and analytics in public administration, as highlighted in other chapters of the *Handbook*.

This chapter is structured as follows. Section 1 describes applications of machine learning in public administration. Section 2 then outlines a road map to applying machine learning in the public sector, focusing on data infrastructure requirements and human capital needs. Section 3 shifts our focus from public administration in general to the justice system. In doing so, it highlights applications of machine learning in the justice system, as well as the data infrastructure and human capital required to implement them. Section 4 presents a case study from India illustrating machine-learning approaches to justice in practice. Section 5 moves beyond descriptive analysis to outline how machine learning can be used to assist causal inference. Finally, we conclude.

MACHINE LEARNING FOR PUBLIC ADMINISTRATION

The use of machine learning is spreading across many functional areas of public administration. While European Union (EU) governments focus on service delivery and public engagement, other areas, such as internal management and law enforcement, are progressively being targeted for the deployment of machine-learning solutions to increase their efficiency and effectiveness (Misuraca and van Noordt 2020). The applications are diverse, from detecting COVID-19 outbreaks to simulating the impact of changes in macroeconomic policy. Machine-learning applications thus provide novel ways for governments to use their data to improve public administration. Chapter 15 of the *Handbook* highlights how machine learning can be used to detect similarities between goods in public procurement.[2]

The use of machine learning provides a few advantages compared with more standard analytical approaches. Standard data analytics provides the analyst with tools bounded by the analyst's capacity to investigate connections between variables in the data—often the coefficients in a regression specification. However, in many public administration settings, the analyst is confronted with factors—individual or organizational—that may influence a policy outcome without the analyst's knowledge. Machine learning enables the exploration of relationships between variables in a principled, and often unsupervised, way. However, causality in machine learning is a relatively recent development, and it presents considerable challenges.[3] Potential applications, therefore, focus less on causal interventions or experiments and more on solutions that, based on given data, best perform an accurate prediction.

The primary focus of this section is on applications of machine learning for administrative data.[4] However, governments may leverage public servant surveys to complement this analysis, particularly for personnel data. For example, a government may be interested in better understanding job satisfaction and how it relates to staff turnover. While a machine-learning analysis could be useful in identifying potential patterns in civil service exit from the full population of interest as a function of demographics (sex, age, education, or race), it might not provide much information about the attitudes of the staff who are at risk of exiting. A public servant survey provides a complement for answering this kind of "why" question, but it may not be large enough to find general patterns in the first place—particularly if it is not linked to administrative data on exits, which is often difficult to do. Thus, machine-learning applications on human resources data outperform surveys at identifying certain kinds of patterns, but they need to be complemented by surveys explaining these patterns and highlighting potential interventions that might address problems.

Machine-Learning Applications

Applications of machine learning can be subsumed under the following three categories.

Detection and Prediction

Machine learning can help policy makers detect and predict destructive events, improving the design and implementation of adequate policy measures. This is the largest application of machine-learning approaches, addressing issues such as COVID-19 outbreaks, fake news, hate speech, tax fraud, military aggression, terrorist activity, cyberattacks, natural disasters, street crime, and traffic congestion—to mention only a few. While the detection and prediction of destructive events is only the first step toward effective government intervention, it is an important instrument for effective policy making.

For example, in Delhi, over 7,500 CCTV cameras, automatic traffic lights, and 1,000 LED signs are equipped with sensors and cameras that collect traffic data, which a machine-learning system processes into real-time insights. Local authorities can then decide how to balance traffic flow in real time and identify traffic patterns and congestion trends in order to plan for the long-term mitigation of traffic problems (Devanesan 2020). Besides these benefits, which are geared toward improving general traffic flow, these systems are also used by the Delhi Police to track and enforce traffic violations, such as speeding or illegal parking (Lal 2021).

More generally within public administration, machine-learning approaches have been used to improve the machinery of government itself. For example, chapter 15 of the *Handbook* highlights how machine-learning techniques have been applied to detect corruption in public procurement. One prominent example of this application is the use of decision tree models (random forest and gradient boosting machine) to detect the presence of Mafia activity in procurement contracts in Italy (Fazekas, Sberna, and Vannucci 2021). In Brazil, federal agencies have deployed machine learning to detect evidence of corruption in federal transfers to municipal governments, as well as in irregularities in paychecks issued to civil servants, as described in case study 2 in chapter 9.[5]

Simulation and Evaluation

Simulating and evaluating the impact of future policy measures is another widespread application area for machine learning. Simulating the potential costs of a policy measure against its expected benefits has become an increasingly relevant tool for governments. For example, in the United States, a simulation known as the National Planning Scenario 1 allows policy makers to simulate what might happen if Washington, DC, were subject to a nuclear attack (Waldrop 2018). Whether policies are designed to stimulate the economy or to contain the spread of a virus, simulation and evaluation provide valuable insight to policy makers before implementation, allowing them to choose which policies maximize intended effects.

Personalization and Automation

Machine learning can also be applied to the personalization and automation of government processes and services. For example, policy makers may customize digital government services for parents to every life stage of their newborn child or tailor the provision of health care services to each patient's particular needs. Additionally, the automation of repetitive tasks leaves more time for public servants to do other tasks. All in all, these novel technologies may help governments be more efficient in their use of time and increase their responsiveness to citizens' needs.

A medical example illustrates this approach. There has been growing interest within the US federal government in using machine learning to improve public health outcomes. A series of pilots to develop such machine-learning solutions have been rolled out. These include the prediction of potential adverse drug reactions using medical reports, the classification of whether a child is likely to have autism based on medical records, and the prediction of unplanned hospital admissions and adverse events (Engstrom et al. 2020, appendix). Another study has found, through the application of machine-learning techniques, that physicians overtest low-risk patients but simultaneously undertest high-risk patients (Mullainathan and Obermeyer 2022).

Practical Steps for Machine Learning in Public Administration

The implementation of machine learning in public administration comprises two key steps. The first is building a high-quality data infrastructure to feed the necessary training data to the machine-learning algorithm. Because public administration data infrastructures are often developed without machine-learning applications in mind, adaptation is often necessary. New data pipelines need to integrate public sector information systems that previously operated in isolation, such as public procurement and budget data. Data standardization practices, such as ensuring that variables in different data tables are named consistently, and other quality checks need to be in place to ensure that the data fed to the machine-learning system are accurate and comprehensive.

Another key step is developing the human capital necessary to deploy machine learning. Before fulfilling the promise of automated, self-learning algorithms, a team of human developers is necessary to set the system in place. In fact, the entire pipeline, from data infrastructure to the training of the algorithm to disseminating actionable insights for policy makers, has to be designed by humans. Having an in-house team capable of developing and maintaining machine-learning applications is crucial. Continuous collaboration between the machine-learning implementation team and policy colleagues who will use its insights ensures that applications are adapted for and stay relevant to public administration's needs.

In the following sections, we dig deeper into these steps. In so doing, we highlight examples of strategies to ensure that both the data infrastructure and the human capital requirements are in place to deploy machine learning in public administration.

Public Sector Data Infrastructure for Machine Learning

Machine learning requires large volumes of data. These data should be of high quality: they should be comprehensive, covering all measurements necessary for the algorithm, and complete, reducing to the extent possible any gaps in measurement that may arise. A robust data infrastructure ensures that these two principles are respected and is a prerequisite for any machine-learning application. The implementation process may require upgrading legacy information systems or integrating new systems into old ones to process the resulting, often large, data sets. Practitioners may benefit from referring to the *Handbook*'s wider discussion of how to reform data infrastructure for analytical insights; this discussion provides lessons that apply to machine-learning settings as well (chapter 9).

Some types of data structure may be more amenable to machine learning than others. Machine-learning applications often require structured data—with well-defined formats and measurements—so policy areas that traditionally deal with structured data, such as finance or budget data, lend themselves particularly well to it. (For an overview of using budgetary data for analytics, see chapter 11.) At the same time, governments produce unstructured data—which lack a predefined data format—such as written documents, meeting recordings, and satellite imagery. To take full advantage of this range of data, practitioners should develop a flexible storage solution that accommodates different types of data. This flexibility should be complemented by thorough documentation of data collection and standardization practices, as well as by measures to ensure compliance with data security regulations, such as the EU's General Data Protection Regulation (GDPR).

The deployment of this data infrastructure requires long-term, costly investment. Data engineers and information technology technicians should partner with the machine-learning implementation team to define data requirements, identify relevant variables, and connect the machine-learning applications to the data infrastructure. The development of a robust data infrastructure is foundational for the effective deployment of machine learning in public administration and should always precede it. Since the data infrastructure is embedded within public administration, its development requires careful coordination between machine-learning engineers, data engineers, and their institutional counterparts who own permissions to the data. Data should be integrated across government agencies to ensure that the largest pool of data is made available for training the application. Open communication between teams and agencies is therefore key.

Human Capital Requirements for Machine Learning in Public Administration

A sustainable machine-learning application is often best achieved by building on in-house human capital. This ensures that the developed solutions are in line with existing government regulations and that policy choices are encoded faithfully. Furthermore, in-house machine-learning experts will be more likely to possess the necessary subject and political expertise required to implement machine-learning solutions in a policy area. Finally, even if an agency decides to rely on external service providers, a certain level of embedded expertise is required to know what is technically possible and feasible, as well as to make informed judgments about the quality of contractor-provided solutions.

To build the necessary skills infrastructure in government organizations, it is first necessary to understand what competencies are needed for machine-learning developers. Naturally, knowledge about machine-learning and deep-learning algorithms is necessary. Beyond this basic knowledge, methods for dealing with large-scale data and databases in general and knowledge about distributed computing systems are also key. To successfully develop, deploy, and operate machine learning in government, familiarity with human-centered design and acquaintance with the legal and ethical frameworks in public administration are important. Finally, policy-area expertise and knowledge about governance and policy making in general enable machine-learning applications to be anchored in the operational needs of government.

Integrating the necessary skills infrastructure within government organizations often proves to be difficult. Hence, governments should follow one or more of the following best practices. Machine-learning talent does not usually follow the classical tenure path of public sector officials. Lateral entries or dedicated programs that allow entries for a limited amount of time can be effective methods for attracting these specialists into government offices. Furthermore, adapting job-classification schemes to include machine-learning-related job categories and increasing salaries and career prospects to better compete with comparable private sector job placements are advisable strategies. Ultimately, it is important to raise awareness among the target talent group about the motivating challenges (for example, social impact) and rewarding benefits (for example, job stability and work-life balance) of public sector work.[6]

Once in government, machine-learning experts' work can benefit greatly from exchange and knowledge sharing with colleagues. Establishing so-called communities of practice to cross knowledge boundaries within and across agencies can help gain legitimacy in relation to relevant stakeholders and foster collaboration among different agencies. Including nontech colleagues in these communities can also ensure that machine-learning applications are developed in a user-friendly manner and integrate well into the daily activities of public servants.[7] Another often-applied practice is the establishment of excellence centers that offer research, support, and training services and help agencies stay on the cutting edge of machine-learning technology. Finally, open communication, such as through blogs or dedicated events, can help other departments take notice and learn from each other's experiences.

Collaborations with external experts and research institutions can be another effective approach to bringing external expertise into a specific project while maintaining control and monitoring quality. Besides concrete project collaborations, establishing academic partnerships, like mobility or internship programs for the temporary assignment of personnel between government agencies and universities or research centers, can help institutionalize such collaborative efforts. Also, tailoring the machine-learning-related educational offerings of partner academic institutions to the particular needs of government organizations can be a viable way to ensure an inflow of machine-learning talent. Finally, building and sustaining intersectoral and interdisciplinary networking initiatives focused on the use of machine learning in government can help establish collaborations and foster learning and exchange.

Ethical Considerations for the Deployment of Machine Learning

Ethical considerations should be at the forefront of machine-learning deployment in public administration, and of analytical applications more broadly.[8] The social contract between governments and citizens differs substantially from the one that private sector companies have with their customers. Citizens or civil servants rarely have a choice about whether to share their data with the government. This makes data security and privacy particularly sensitive because most machine-learning applications require substantial amounts of data for appropriate training. On top of more general regulations, like the GDPR, ensuring the responsible usage and sharing of data, potentially by applying adequate anonymization techniques, should be a priority for governments to ensure citizens' trust (for a more extensive discussion, see chapter 28).

Another factor that can inhibit citizens' trust stems from the rare position of governments in relation to machine-learning technologies. Governments unify the roles of user and regulator in a single entity. This makes the public sector's use of machine learning a particularly delicate target of public scrutiny. Cases in which government machine-learning systems violate citizens' rights, like the recent case of the Dutch automated surveillance system for detecting welfare fraud, pose serious threats to citizens' trust (see box 16.1). It is therefore necessary to faithfully encode legal and political choices and ensure compliance with international regulatory frameworks to ensure ethical machine-learning applications in the public sector.

Applications of machine learning in government must consider that citizens often rely only on the government for public services like social security. This is a particular challenge to using machine learning in settings where the algorithm must make a choice. For instance, regarding social security systems, an algorithm might decide whether a citizen is eligible for a particular government benefit. In this situation, the algorithm would have to compare what would happen if the citizen were granted the benefit versus if the citizen were not. The algorithms that underlie this decision-making have to make assumptions about what would happen in each scenario, and the usefulness of the final decision depends on how appropriate these underlying assumptions were. If a citizen were denied a benefit due to an algorithm's decision, who would hold the algorithm accountable?

BOX 16.1 The Precedent Case of the Netherlands' Automated Surveillance System

On February 5, 2020, the District Court of The Hague ruled that SyRI (Systeem Risico Indicatie), a machine-learning application used by the government of the Netherlands to detect welfare fraud, violated Article 8 of the European Convention on Human Rights (ECHR)—that is, the right to respect for private and family life. This case is one of the first times a court has stopped a government's use of machine-learning technologies on human rights grounds and is thus considered to offer an important legal precedent for other courts to follow.

SyRI was designed to prevent and combat fraud in areas such as social benefits, allowances, and taxes by linking and analyzing data from various government and public agencies and generating personal risk profiles. It was deployed by the Minister of Social Affairs and Employment upon the request of various public agencies, including, among others, municipalities, the Social Insurance Bank, and the Employee Insurance Agency. The system mainly used a neighborhood-oriented approach, meaning it targeted specific neighborhoods where the linked data indicated an increased risk of welfare fraud.

Although the Court agreed with the government of the Netherlands that the fight against fraud is crucial and thus that employing novel technologies offering more possibilities to prevent and combat fraud generally serves a legitimate purpose, it ruled that the way SyRI was operated did not strike a "fair balance" between social interests and violation of the private lives of citizens, as required by the ECHR. In particular, the Court stated that due to the lack of insight into the risk indicators and the operation of the risk model, the system had violated the transparency principle and that discrimination or stigmatization of citizens in problem areas could not be ruled out.

The ruling, which led to the immediate halt of SyRI and caused public uproar far beyond the Netherlands, is a telling example of the potential negative consequences of applying machine-learning systems for government purposes without adequately addressing their potential ethically adverse side effects.

Often, modeling assumptions are not directly testable and hence require a substantial level of expertise over both what assumptions the algorithm is making and the suitability of those assumptions for a given public sector setting (Athey 2017). Public policy making through machine learning therefore raises important ethical questions. Choices may be made on behalf of government officials about citizen outcomes by machines they do not fully understand. For this reason, there is tension between the use of machine-learning technology to improve public administration and the oversight required to ensure that its use is in accordance with ethical principles. This tension becomes particularly salient when the use of previous administrative data for algorithm training introduces human biases into the system. Not uncovered, these biases can lead to "discrimination at scale" in sensitive areas such as racial profiling.

Finally, most applications of machine learning for government purposes are not static and should be adapted to evolving understandings of ethical principles. For example, algorithms for detecting fraud need to be constantly updated or retrained to address new forms of misconduct uncovered by agency employees and avoid an excessive focus on past forms of misconduct. Without such updating, algorithms may be biased toward past versions of criminal conduct. Constant updating by consulting domain experts and ethical advisors is necessary to ensure the effectiveness and ethical compliance of machine-learning technologies in government.

MACHINE LEARNING FOR JUSTICE

We now turn our focus to applications of machine learning in the justice system. The justice system is an institutional setting with high-frequency data, well-documented cases, extensive textual evidence, and a host of legal actors. As such, it is a useful setting within which to explore the use of machine learning for administration in the public service. An example of a core analytical question in justice is how the characteristics of judges impact judicial outcomes, such as rulings. This is a formulation of a wider question about how the individual characteristics of public officials impact the quality of public services provided by the government. It is a question that the analytics of public administration can investigate with the right measurement, data infrastructure, and skills for analysis.

Significant progress has been made in answering this question using machine learning (Chen 2019a, 2019b; Rigano 2018). In the United States, machine learning is already used in processing bail applications, DNA analysis of crimes, gunshot detection, and crime forecasting (Epps and Warren 2020; Rigano 2018). The large volume of data from surveillance systems, digital payment platforms, newly computerized bureaucratic systems, and even social media platforms can be analyzed to detect anomalous activity, investigate potential criminal activity, and improve systems of justice. For example, after the January 6, 2021, riots at the US Capitol, investigators used machine-learning-powered facial-detection technologies to identify participants and initiate prosecutions (Harwell and Timberg 2021). Machine-learning systems can also reduce the barriers to accessing courts by providing users with timely information directly, rather than through lawyers. Sadka, Seira, and Woodruff (2017, 50) find that providing information to litigants in mediation reduces the overconfidence of litigants and nearly doubles the overall settlement rate, but this only occurs when litigants are informed directly rather than through their lawyers.

The application of machine-learning systems to justice systems is useful because slight tendencies in human behavior can have significant impacts on judicial outcomes. A growing body of work demonstrates how small external factors, most of which participants are unaware of, can leave their mark on the outcomes of legal cases. Analysis of courts in the US, France, Israel, the United Kingdom, and Chile, for example, has found that in various settings, the tone of the words used in the first 3 minutes of a hearing, the incidence of birthdays, the outcomes of sporting events, and even the time of day of a hearing or a defendant's name can affect the outcome of a case (Chen 2019a). An analysis of 18,686 judicial rulings by the 12 US circuit courts (also known as courts of appeals or federal appellate courts), collected over 77 years, illustrates that judges demonstrate considerable bias before national elections (Berdejó and Chen 2017). Similarly, there is new evidence on sequencing matters in high-stakes decisions: decisions made on previous cases affect the outcomes of current cases even though the cases have nothing to do with each other. For instance, refugee asylum judges are two percentage points more likely to deny asylum to refugees if their previous decision granted asylum (Chen, Moskowitz, and Shue 2016).

Given the abundant evidence of how bias shapes decisions made by officials in the justice system, machine-learning methods can identify these sources of bias and signal when they shape judicial outcomes. The subtlety of different forms of bias requires an approach that searches through a very large number of relationships to detect their wider effects, an approach for which machine learning may be well suited. This can result in a more streamlined system and a reduction in backlog. Such tools can identify discrimination and bias even when these are not evident to the participants in the courts themselves, thereby strengthening the credibility of the judiciary (Bhushan 2009; Galanter 1984; Kannabiran 2009). Moreover, as large backlogs of cases are a significant problem for the efficiency of the judiciary in developing countries, interest is growing in performance metrics that will improve the functioning of the judiciary.

The adoption of machine-learning systems, however, is not an easy-to-implement solution, in particular for the justice system. Despite the growing availability of judicial data, it is first necessary to process these data in a way that is amenable to machine learning. This requires the integration of data from different sources, the processing of textual data into quantifiable metrics, and the definition of indicators for learning tasks that reflect either performance objectives or the operationalization of the concepts of fairness and impartiality. This is not an easy task: it requires substantial investments in data infrastructure and human capital, as well as building a conceptual framework. Therefore, to implement machine-learning algorithms successfully, justice systems need to acquire and train teams of machine-learning engineers, subject matter experts, and legal actors to develop machine-learning algorithms that are operationally relevant. These considerations are similar to the ones highlighted in the broader consideration of public administration.

Finally, there are ethical concerns regarding machine-learning applications for judicial outcomes. Practitioners and citizens may raise questions regarding the interpretability of algorithms because technological sophistication creates a "black box" problem: increases in technology's sophistication make its operation less interpretable (Pasquale 2015). The challenge of interpretability also raises concerns about accountability and oversight for these systems. Furthermore, the gap between those who can and those who cannot access and understand these technologies exacerbates existing social divisions and intensifies polarization. For all these reasons, machine-learning tools should be seen as complements to rather than substitutes for human decision-making, in particular for institutions that make life-altering decisions, such as the judiciary.

Judicial Data Infrastructure for Machine Learning

It is increasingly recognized that "the world's most valuable resource is no longer oil, but data" (*Economist* 2017). Like oil, raw data are not valuable in and of themselves; their value arises when they are cleaned, refined, processed, and connected to other databases that allow for the generation of insights that inform decision-making. This is particularly true in the field of machine learning, which requires large amounts of data to build accurate predictive models that provide information on the process, behaviors, and results of any indicators of interest.

Judiciaries collect vast amounts of data daily. Despite the availability of data, judicial data have rarely been analyzed quantitatively. In recent years, the transition from paper to e-filing and case management systems has facilitated the systematic analysis of massive amounts of data, generating performance metrics that can be used to evaluate courts and justice actors. Furthermore, with advances in machine learning, natural language processing (NLP), and processing power, these data create valuable opportunities to apply machine-learning models to evaluate and improve justice systems (Ramos-Maqueda and Chen 2021). Nonetheless, the extent to which countries can utilize novel approaches in machine learning and data analytics will depend on the available data infrastructure. The question, then, is which data do (and should) judiciaries collect?

In the justice domain, an integrated justice system brings together data on each case and connects these data with information on the actions and decisions at each case milestone. For instance, this includes information on the case filing, initial decisions, hearings, rulings, and sentences for each case. These data should also relate to potential appeals in order to help understand the evolution of the case. By implementing NLP on the text of case filings or judicial decisions, judiciaries can automate the revision of case filings or identify relevant jurisprudence for judges and court actors, for instance. Beyond information on the justice process itself, judiciaries will gain valuable insights from connecting these data with other information, such as human resources data, information on recruitment, or data from judicial training, to understand how best to select, train, and motivate judges depending on their background and experience.

To evaluate the impact of machine-learning interventions in justice, judiciaries should ideally collect information from other agencies involved in the justice process, as well as the economic outcomes of the parties involved. In criminal justice, an interoperable data ecosystem will connect judicial data with data from the prosecutor's office, the police, and the prisons, which will enable judiciaries to understand where the case has come from and the implications of judicial sentences. In civil cases, this may include economic data on citizens and firms who participate in the justice system, such as tax data, social insurance data, or procurement data. This way, judiciaries will be able to evaluate the impact of machine-learning applications not only on the judicial process but also on the lives of citizens and the financial status of firms that use the justice system.

In addition to the external and internal databases, it is also recommended that judiciaries carry out surveys that complement administrative information with the experience of the parties and employees involved in the justice system. Administrative data will not capture important elements of user or employee satisfaction, for instance, so survey data are a necessary complement for understanding the impacts of any new machine-learning models. We also recommend surveying those who are not necessarily part of the justice system—but who could be potential users in the future—through legal needs assessments.

Finally, there are additional complexities to developing data ecosystems for machine learning. Data must be of high quality, and large volumes need to be collected and stored to make them amenable to artificial intelligence (AI) algorithms, which, in general, presuppose big data. This requires data extraction, transformation, and loading (ETL) processes designed to support AI pipelines and, in most use cases, dedicated data engineers to maintain them.

Human Capital Requirements for Machine Learning in Justice

Justice systems often have limited in-house access to the necessary human capital to develop machine-learning applications. Judicial officers are rarely experts on data analysis—which is seldom part of their training—and machine-learning engineers generally lack the domain expertise necessary to understand the functioning of the law. In courts without sufficient human capital to take advantage of available data, training public servants in even simple data analysis skills may be a valuable long-term investment for improving the functioning of courts. Nevertheless, the development of machine-learning approaches may remain out of reach for public officials whose training does not include statistical modeling or data engineering.

An alternative approach is relying on nongovernmental organizations, international organizations, or even private companies to develop machine-learning applications. An example of this approach is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool in the United States, which is an algorithm that generates recidivism risk scores to aid judges in their rulings. However, judiciaries should consider the long-term sustainability of an outsourced solution as well as ethical concerns. COMPAS has been the target of controversy due to its proprietary algorithm and the inability of public officials and citizens to understand how it operates under the hood.[9] Additionally, a reliance on external contractors often substitutes for in-house development of the necessary human capital to develop machine-learning technologies, reproducing external reliance on nonjudicial actors for both the maintenance and expansion of machine-learning solutions.

Whether in-house or externally sourced, the forms of human capital required for machine-learning applications are diverse and costly. The implementation team should include machine-learning engineers, software developers to code the user interface, data engineers to develop the data infrastructure, legal experts, and project managers to communicate the judiciary's needs to the implementation team. Because each component of the project relies on the others—there can be no user interface without a data infrastructure to feed it—the team should ensure that their timelines are aligned. A sufficient budget should be allocated to the project to cover the team's time for both the implementation and the monitoring of the technology after the first version of the application is developed.

Ethical Concerns

Developers of machine-learning applications should carefully consider the ethical implications of their use by judicial actors. Only technologies that aid human decision-making, rather than replacing it, should be adopted in the courts. There are multiple reasons for this recommendation. As noted earlier, algorithms have the "black-box" problem of interpretability—that is, it is not easy to trace the output of complex algorithms to the data inputs themselves. Additionally, biases in the decisions of judicial actors are reflected in the algorithm's training data and may be encoded into the algorithm itself. Thus, using machine-learning algorithms to inform judicial decisions without critical oversight raises the risk of replicating these biases elsewhere in the system. The inherent choice of performance metrics can also reinforce existing biases by decision-makers within the system. Addressing these issues requires a participatory and deliberative approach to the design, implementation, and evaluation of machine-learning technologies.

A reasonable demand to guarantee trust and fairness is that algorithms be interpretable. A judge may request the reason why a particular decision has been recommended by the algorithm. This transparency enhances judges' trust in the technology and allows for disagreement with its recommendations. Given the complexities of working with machine-learning algorithms, any rollout must be preceded by a phase of comprehensive study and rigorous testing of the systems themselves. Randomized controlled trials that carefully estimate the causal impacts of the adoption of these algorithms to properly evaluate their costs and benefits are essential. A carefully constructed trial can provide important benchmarks on cost, efficiency, user satisfaction, and impact on key performance metrics, all essential for a justice system to credibly serve citizens.

CASE STUDY: MACHINE LEARNING FOR JUSTICE SYSTEMS IN INDIA

This case study illustrates how machine learning was implemented in the national justice system of India by the Data and Evidence for Justice Reform (DE JURE) team. Due to India's large population and volume of cases, justice officials are often unable to effectively manage cases in a timely fashion. India has just 19 judges per million people and 27 million (2.7 crore) pending cases (Amirapu 2020; Chemin 2012; Kumar 2012). To address this, the Indian justice system has made considerable advances in the adoption of information technology and has released large volumes of data to court users and encouraged them to use electronic systems. Yet legislative, institutional, and resource constraints have limited the full impact of these advances (Amirapu 2020; Damle and Anand 2020).

In this section, we describe how the DE JURE team implemented machine-learning applications in India. We first highlight the data infrastructure requirements for implementing machine-learning applications, as well as how these applications could enhance the functioning of the justice system.

Judicial Data Infrastructure in India

In the past 15 years, considerable efforts have been made to adopt and deploy information technology systems in the courts of India. One of the most significant projects, the e-Courts project, was first launched in 2005 by the Supreme Court of India through the National Policy and Action Plan for Implementation of Information and Communication Technology (ICT) in the Indian Judiciary. The e-Courts initiative introduced technology into Indian courts in a variety of ways.

The most obvious feature of the system was the deployment of technology within courtrooms. Judges were provided with LCD touch screens, screens and projectors were connected via a local network to disseminate information to lawyers, and electronic boards at the courts displayed the queue of case numbers scheduled for hearing on a particular day. Outside the courtroom, e-filing procedures were established, and a data management architecture was created that ranged from the scanning of old cases into the electronic system to the creation of digital archives. The ICT plan also established direct electronic communication with litigants and an online case management system.

These investments eventually paved the way for the creation of the National Judicial Data Grid (NJDG), a database of 27 million cases that allows court users to view the status of pending cases and access information on past hearings. For the DE JURE team's goal to implement machine-learning tools, the most significant resources were the digital archives of cases. The team was able to scrape these publicly available digital archives to construct an e-Courts district court data set of 83 million cases from 3,289 court establishments.[10] They were able to curate details, like the act under which the case was filed, the case type (criminal or civil), the district where it originated, the parties to the case, and the history of case hearings, in a manner that made the data amenable to large-scale analysis.

A wider data ecosystem has been created by joining additional sources to the case data, including the following:

- **Data on judges:** To better understand the impact of specific judges—their identity, training, and experience—the team constructed a database of judges for the courts of India. They began this task by extracting data from editions of *The Handbook on Judges of the Supreme Court of India and the High Courts*, released by the Supreme Court of India, and appending to it information from various High Court websites. So far, the team has assembled detailed information for 2,239 judges from the handbooks for the years 2014–20. Most notably, 93.5 percent of these judges are men and 6.5 percent are women, and their range of experience covers a period spanning approximately 70 years.
- **Database of Central Acts:** This auxiliary data set is intended to give a definitive list of standardized act names. This could then be used to standardize the act names appearing in various cases. This would allow the team to analyze all cases filed under a given act. The team has, for example, examined all cases related to the Water Act of 1974 and found a total of 978 such cases at the Supreme Court and High Courts of India. The list of central (federal) acts can be viewed on the Legislative Department website of the Ministry of Law and Justice. There is currently no centralized source for all state legislation: this needs to be obtained from state websites separately.
- **Other administrative data:** Data on other institutions can be linked to the judicial data at the district as well as the state level. For example, data on Indian banks and their branches are available through the Reserve Bank of India. This database contains information on names, locations, license numbers, license dates, addresses, and other unique identifiers. The team has scraped, cleaned, and organized these data for further analysis. The database contains about 160,000 entries. The unique identifiers and location information allow the team to merge these data with banks appearing in litigation in courts that are present in the e-Courts databases. Merging these data with the legal data allows the team to examine a variety of interesting questions about the development of financial markets in an area, participation in the justice system, and the impacts of legal rulings.

The quality of these data varies significantly: there is no nationally standardized system for defining variables or reporting on them. For instance, in some states, the legal act name and section numbers are well delineated, but in other states, this is not the case. This makes it difficult to compare individual case types across courts and states (Damle and Anand 2020). There are no standardized identifiers within the data to follow a case through its potential sequence of appeals in higher courts. In a similar vein, there is no easy way to track a criminal case from its entry into the system as a first information report (FIR) to its exit as a judgment. There are inconsistencies in identifying information about participants, their attributes, and the types of laws or acts that cases relate to. There are also issues of incorrect reporting and spelling mistakes.

Machine-Learning Applications in the Courts of India

The quality of data in India's justice system is often compromised: case data are incomplete or litigants' identities are not registered. The DE JURE team has constructed a robust data pipeline to collect often-incomplete judicial data, as well as machine-learning tools to clean and prepare them for analysis. In this section, we contextualize the problem: how data quality issues in judicial data manifest themselves in India. In the following section, we describe the solution: how machine-learning tools have been designed to enhance the quality of judicial data for analysis.

Legal data released by the Indian judiciary are voluminous, messy, and complex (Damle and Anand 2020). The typical case has clear tags for some key dates (filing date), the key actors (petitioner, respondent, and judges) and the court name, but information about the type of case, the outcome of deliberations, and pertinent acts cited is often not clearly identifiable in the textual body of the judgment. Cleaning and preprocessing the data is critical for any form of analysis, especially for supervised algorithms trained on these data. Traditional empirical legal studies have typically addressed this issue by relying on small-scale data sets in which legal variables are manually coded and the scope of inference is related to a small body of legal cases pertinent to a single issue (Baxi 1986; Gauri 2009; Krishnaswamy, Sivakumar, and Bail 2014).

These traditional approaches are unable to keep up with the incoming volume of cases. In this context, machine-learning tools provide an alternate approach to detecting errors or gaps in the data and correcting them in an automated fashion. Using machine learning, it is possible to infer the identities of participants even when these data are not registered. Additionally, laws used as precedents for a ruling can be identified through text analysis. Beyond the data quality itself, machine-learning approaches can help identify biases and discrimination in judges' rulings.

Inference about the Identities of Participants

Some databases of judgments provide no identifying information about participants in the cases themselves. To better understand who participates in the courts, the team first extracts litigant names from the raw text of the judgments and then uses matching algorithms to identify the type of litigant (individual, company, or state institution). Classifying participants can be challenging. If the identification exercise involves government agencies, it is first necessary to compile all the different agencies of the state and national governments. Manually tagging entities is prohibitively time-consuming, but the existence of latent patterns in the names makes this fertile ground for machine-learning applications.

The machine-learning application relies on similarity across names for participants that belong to the same "group" to classify a particular name as belonging to that group. Using prelabeled data—individual name A belongs to group B—the machine-learning algorithm can extrapolate to unlabeled data, where an individual's name is available but not their group. Some obvious groups of interest are gender, caste, and religion, which are not recorded in judicial data but are available in other data sources. Another group of interest may be whether a participant is a government agency or not. We focus here on people's first and last names, for illustration.

The team first formats individuals' names to ensure that each individual can be identified by an honorific title, a first name, and a last name. Honorifics, such as Shri, Sri, Smt., Mr., Mrs., and Ms., enable the algorithm to directly identify an individual's gender. To extend this classification to names without an honorific, the team trains an algorithm on a publicly available corpus of labeled common Indian first names. Training this algorithm, often referred to as training a classifier, is the process by which the algorithm learns patterns within the data related to a group. Here, these patterns are the statistics of co-occurrence of letters in names, the lengths of names, and other features that allow the algorithm to determine whether a name indeed belongs to a particular group: in this case, a gender.[11]

These algorithms formalize intuitive notions of why a name belongs to a given group by identifying frequently occurring patterns within names associated with that group. Caste assignment is more complicated because the same last name could be associated with multiple caste groups. The name "Kumar," for example, could belong to a person belonging to the Scheduled Castes, the Scheduled Tribes, or the category "Other."

In the case of such names, the team generates distributions of the last name across the different caste categories. They then use this distribution to generate a prediction and combine this with the predictions of other models to ensure a robust prediction. They assign a caste to each household based on a simple majority vote among these models.

Identification of Laws and Acts

Legal texts in India's justice system do not currently employ a standardized citation style for referring to acts or laws. For example, the Hindu Marriage Act may be referred to in a variety of ways, such as "u/s 13 clause 2 of Hindu Marriage Act," "u/s 13(b) Hindu Marriage Act," or "u/s 13 of Hindu Marriage Act 1995." Again, machine-learning tools can be used to address this issue.

In this project, the DE JURE team uses a set of tools that create mathematical representations of the text in the form of vectors. Term frequency–inverse document frequency (TF–IDF) is one popular method for representing a string of words as a numerical score that reflects how frequently a word is used within a given text and how infrequently it appears in the corpus. Applying this to act names, the team uses different clustering algorithms to group particular act citations based on how similar they are numerically. This approach groups the underlying act-name data in a manner that best preserves the coherence within groups (a particular act name) and the distance across groups to make the classification.

The identification of specific acts and how often they are cited opens new opportunities for legal analysis. The team can, for example, compare the different types of acts that are cited in the different courts within India's justice system. It can allow researchers and practitioners to identify the real-time evolution of legal citation—and legal thought—as judges refer to these acts.

Using Descriptive Analysis and Machine Learning to Identify Discrimination and Bias

Bias and discrimination can occur in different areas of policy making, particularly when civil servants exercise discretion, such as in judicial rulings. Judges may favor or deny due process to plaintiffs depending on their ethnic or gender identity, undermining the rule of law and the right to impartial judgment. This challenge is, of course, not unique to judiciaries. A broad academic literature has demonstrated that bias in a human decision-maker can have conscious as well as unconscious drivers and may manifest in complex ways and in a variety of contexts that can be difficult to prove (Banerjee et al. 2009; Bertrand and Mullainathan 2004; Ewens, Tomlin, and Wang 2014; Kleinberg et al. 2018). In other settings, such as labor markets and educational institutions, algorithms—rules that take "inputs" (like the characteristics of a job applicant) and predict some outcome (like a person's salary)—have been shown to create new forms of transparency and to serve as methods for detecting discrimination (Kleinberg et al. 2020; LeCun, Bengio, and Hinton 2015).

Machine-learning algorithms can identify these biases and forms of discrimination, enabling governments to measure their prevalence and design policy changes to reduce them. In the courts of India, algorithms could help judges make critical decisions about cases (for example, dismissals or bail applications) and reduce bias in their rulings. Building such algorithms requires a rich data set that includes litigant characteristics (caste and gender), lawyer characteristics, court characteristics, case details (filing details and evidence provided), and case outcomes (such as the granting of bail or the dismissal of a case). A machine-learning engineer would develop a "learning procedure" that would aim to provide a predicted outcome from a broad range of inputs and modeling approaches, such as a neural network (Dayhoff 1990). These models differ from traditional statistical methods, such as linear regression, which are more deductive (presuming a linear fit between a few sets of variables) than inductive (allowing the data to report the best fit between a large set of variables).

These insights could be invaluable not only within the courtroom itself but also in judicial education. Experiments are currently underway, in the Judicial Academy of Peru, for example, to assess methods to improve case-based teaching by using the history of a judge's past decisions, which can reveal potential bias or error. The data are also suitable for creating personalized dashboards and interfaces that provide judges, mediators, and other decision-makers with real-time information about their own performance relative to their own previous decisions and those of others who are comparable (Kling 2006). This information can be used to augment the capabilities of judges and lawyers, increase their productivity, and reduce potential bias in their decisions.

BEYOND DESCRIPTIVE ANALYSIS: IMPACT EVALUATION IN THE JUSTICE SYSTEM

Moving beyond descriptive analysis and more correlational analysis of data, an underexplored field in the justice system is policy experiments for impact evaluation. Legal scholars and judges have long debated the merits of implementing various laws and regulations and have justified their arguments with theories about the effects of these legal rules. This situation resembles the field of medicine a century ago: before clinical trials, medical research focused on theoretical debates rather than rigorous causal evidence.

A growing body of empirical research now demonstrates that causal inference is possible in judicial studies. For example, in situations where cases are randomly assigned to judges, the random assignment itself can be used as an exogenous source of variation to evaluate the impact of judicial decisions. Since judges do not choose their cases, observed rulings reflect the judge's personal characteristics (ideological preferences or biases) and features of the case rather than the judicial process as a whole. Additionally, informational treatments can have an impact on the behavior of judges, improving their performance (box 16.2).

BOX 16.2 Leveraging Data and Technology to Improve Judicial Efficiency in Kenya

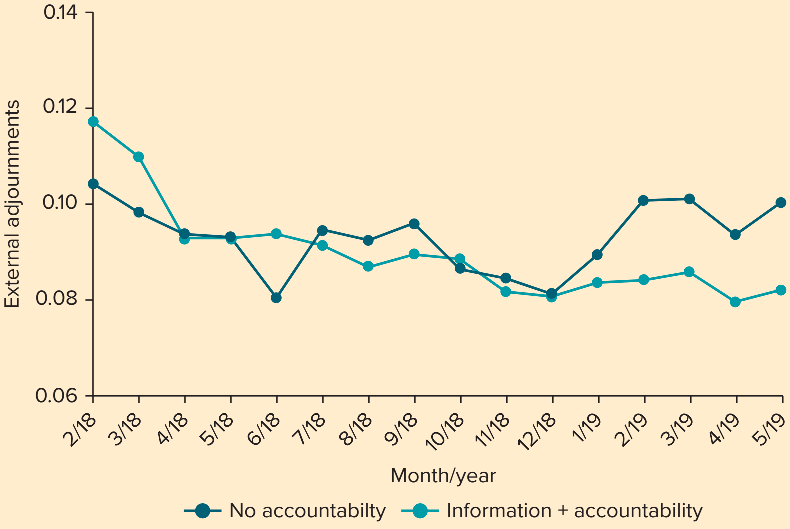
In partnership with the Judiciary of Kenya and McGill University, the World Bank's Development Impact Evaluation (DIME) Data and Evidence for Justice Reform (DE JURE) team has been leveraging underutilized administrative data to improve judicial efficiency. Through its case management system, the Judiciary of Kenya collects large amounts of administrative data on the characteristics of cases, the dates of hearings and reasons for adjournments, and other important metrics of court performance. These data are readily available for understanding and designing interventions to address challenges to the efficient delivery of justice, such as adjournments of hearings, which cause large delays in court proceedings. Despite the richness of these data, they were not being used for decision-making. DIME and the Judiciary of Kenya decided to leverage these data systems to design an algorithm identifying the greatest sources of delay for each court and presenting recommended actions. The team included performance information in a one-page feedback report. Then it studied whether this simplified, action-oriented information could reduce adjournments and improve judicial performance.

In a randomized controlled trial across all 124 court stations in Kenya, the team compared the impact of sharing the one-page feedback reports only with judges and supervisors to the impact of sharing them with Court User Committees as well, the latter acting as an additional accountability mechanism (figure B16.2.1). The team found that the one-page feedback report with the accountability mechanism reduced the number of adjournments by 20 percent over a four-month period and increased the number of cases resolved (Chemin et al. 2022). The conclusion was that the report was more effective when both tribunals and Court User Committees received it. Thus, sharing performance information with courts may be effective to improve efficiency, but it is particularly effective when this information is also shared with civil society and court stakeholders. This study served as proof of concept that utilizing data to provide information to judicial actors can reduce adjournments and increase the speed of justice, which have a downstream impact on the economic outcomes of citizens and firms.

(continues on next page)

BOX 16.2 Leveraging Data and Technology to Improve Judicial Efficiency in Kenya *(continued)*

FIGURE B16.2.1 Impact of One-Page Feedback Reports on Case Delays in Kenya



Source: DE JURE, World Bank.

Randomly assigning cases to judges predicted to be harsh or lenient allows researchers to identify the long-run causal impacts of the length of sentences (Dobbie and Song 2015). To identify the causal effect of a sentence length of eight months or eight years, a randomized controlled trial would need to randomize the sentence, which is impossible. However, assigning a defendant to a judge predicted to assign an eight-month sentence or to another judge predicted to assign an eight-year sentence allows researchers to identify the causal impact of sentence length on subsequent life outcomes.

The same conceptual framework can examine the causal effects of debt relief on individuals' earnings, employment, and mortality (Sampat and Williams 2019). This causal approach sheds light on the impact these judicial decisions can have on individuals' welfare outcomes. By applying machine learning to infer the bias, lenience, and ideological preference of a judge, researchers can identify the causal effect of these variables on the judicial system and the life outcomes of those affected by judges' decisions.

CONCLUSION

In this chapter, we argue that machine learning is a powerful tool for improving public administration in general and the justice system in particular. Machine learning, at its core, emphasizes a methodological approach centered around a learning problem: defining indicators and using evidence to improve them. Under the umbrella of this methodological approach, multiple applications are available to tackle key issues in public administration. Algorithms can be written to draw inferences about the identities of participants and study the deliberative processes they employ within courtrooms. Machine-learning tools can also convert a high volume of textual data to numerical estimates that can be used for understanding the processes and outcomes of different types of case data, including public procurement, taxes, and the systems of justice themselves.

These tools, however, have several limitations and requirements that need to be addressed before they can be effectively deployed in the courts. At the very outset, there are significant issues related to the privacy of personally identifiable information, security, and the control of legal data. Next, the algorithms require data preprocessing, training on large, high-frequency data sets, and iterative refinement with respect to the actual use cases in which they are deployed. This requires strong pilot programs that are studied as part of randomized controlled trials. Insights on data privacy and costs as well as outcomes require that these pilots be constructed on a reasonable scale.

Public administration officials often execute a range of tasks, from the ordinary to the complex, such as the adjudication of a trial. The smart application of machine learning can enhance levels of automation, productivity, and the level of information extracted from data generated in the public sector. If done right, it can help reduce noise—and this can be one step toward aiding the impersonal execution of tasks, reducing bias, enhancing predictability, and making decision rules more transparent. But none of these outcomes can be presupposed from the machine-learning approach: they depend on the ethical framework and operational relevance underlying its implementation. Machine-learning practitioners are therefore advised to take necessary precautions and develop solutions that are accountable to the public and useful for government officials.

NOTES

1. Machine learning is therefore a methodological approach anchored in a learning framework. It is not a radical departure from classical approaches to statistics and, in fact, often builds on canonical models (such as linear or logistic regressions), nor is it exempt from well-known challenges, such as bias and model misspecification.
2. The machine-learning application is part of a broader study on bureaucratic allocation available in Bandiera et al. (2021).
3. For a discussion of causality in machine learning, see Schölkopf (2022).
4. Public servant surveys, due to their smaller sample frames and limited applications in prediction, are rarely used directly for machine-learning applications.
5. Federal transfers have been analyzed using machine learning since 2018 (CGU 2018).
6. The Inter-American Development Bank has written extensively on the topic. See chapter 3 of Porrúa et al. (2021).
7. For a concrete example of this approach, see case study 2 in chapter 9 of the *Handbook* on the Brazilian federal government's experience with the development of machine learning for payroll analysis).
8. For an overview, see chapter 6 of the *Handbook*.
9. For a discussion, see Yong (2018).
10. The e-Courts data are public and can be accessed via the district court websites, the e-Courts Android/iOS app, or the district court services webpage at https://services.ecourts.gov.in/ecourtindia_v6/.
11. To reduce model overfitting, we use the majority vote from multiple trained classifiers, including a logistic regression model and a random forest classifier to make predictions on gender. A logistic regression models the probability of a binary outcome or event. A random forest classifier will use decision trees (nested if-then statements) on features of the data to make the prediction. We have also made predictions of religion and caste using similar approaches. Muslims can be recognized in the data through the distinctiveness of Muslim names: common names such as Khan and Ahmed can easily be assigned and coded, but for others, we utilize the occurrence of specific letters (such as *q* and *z*) through appropriate classifiers to identify additional names.

REFERENCES

Amirapu, Amrit. 2020. "Justice Delayed Is Growth Denied: The Effect of Slow Courts on Relationship-Specific Industries in India." *Economic Development and Cultural Change* 70 (1): 415–51. https://doi.org/10.1086/711171.

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355 (6324): 483–85. https://doi.org/10.1126/science.aal4321.

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats." *The Quarterly Journal of Economics* 136 (4): 2195–242. https://doi.org/10.1093/qje/qjab029.

Banerjee, Abhijit, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan. 2009. "Labor Market Discrimination in Delhi: Evidence from a Field Experiment." *Journal of Comparative Economics* 37 (1): 14–27. https://doi.org/10.1016/j.jce.2008.09.002.

Baxi, Upendra. 1986. *Towards a Sociology of Indian Law*. New Delhi: Satvahan.

Berdejó, Carlos, and Daniel L. Chen. 2017. "Electoral Cycles among US Courts of Appeals Judges." *The Journal of Law and Economics* 60 (3): 479–96. https://doi.org/10.1086/696237.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013. https://doi.org/10.1257/0002828042002561.

Bhushan, Prashant. 2009. "Misplaced Priorities and Class Bias of the Judiciary." *Economic and Political Weekly* 44 (14): 32–37. https://www.jstor.org/stable/40278698.

Brynjolfsson, Erik, and Andrew McAfee. 2017. "The Business of Artificial Intelligence." *Harvard Business Review*, July 18, 2017. https://hbr.org/2017/07/the-business-of-artificial-intelligence.

CGU (Controladoria-Geral da União). 2018. "Inteligência Artificial Analisará Prestação de contas em Transferências da União." Comptroller General of Brazil, October 23, 2018. https://www.gov.br/cgu/pt-br/assuntos/noticias/2018/10/inteligencia-artificial-analisara-prestacao-de-contas-em-transferencias-da-uniao.

Chemin, Matthieu. 2012. "Does Court Speed Shape Economic Activity? Evidence from a Court Reform in India." *The Journal of Law, Economics, & Organization* 28 (3): 460–85. https://doi.org/10.1093/jleo/ewq014.

Chemin, Matthieu, Daniel L. Chen, Vincenzo Di Maro, Paul Kimalu, Momanyi Mokaya, and Manuel Ramos-Maqueda. 2022. "Data Science for Justice: The Short-Term Effects of a Randomized Judicial Reform in Kenya." TSE Working Paper 22-1391, Toulouse School of Economics, Toulouse.

Chen, Daniel L. 2019a. "Judicial Analytics and the Great Transformation of American Law." *Artificial Intelligence and Law* 27: 15–42. https://doi.org/10.1007/s10506-018-9237-x.

Chen, Daniel L. 2019b. "Machine Learning and the Rule of Law." In *Law as Data: Computation, Text, and the Future of Legal Analysis*, edited by Michael A. Livermore and Daniel N. Rockmore, 433–41. Santa Fe, NM: Santa Fe Institute Press. https://doi.org/10.37911/9781947864085.16.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. 2016. "Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *The Quarterly Journal of Economics* 131 (3): 1181–242. https://doi.org/10.1093/qje/qjw017.

Chen, Jonathan H., and Steven M. Asch. 2017. "Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations." *New England Journal of Medicine* 376: 2507–09. https://doi.org/10.1056/NEJMp1702071.

Cross, Tim. 2020. "Artificial Intelligence and Its Limits: An Understanding of AI's Limitations Is Starting to Sink In." *Economist*, Technology Quarterly, June 13, 2020. https://www.economist.com/technology-quarterly/2020/06/11/an-understanding-of-ais-limitations-is-starting-to-sink-in.

Damle, Devendra, and Tushar Anand. 2020. "Problems with the e-Courts Data." National Institute of Public Finance and Policy Working Paper 314, National Institute of Public Finance and Policy, New Delhi, India. https://www.nipfp.org.in/media/medialibrary/2020/07/WP_314__2020.pdf.

Dayhoff, Judith E. 1990. *Neural Network Architectures: An Introduction*. New York: Van Nostrand Reinhold.

Devanesan, Joe. 2020. "AI-Powered Traffic Management Is Slashing Asia's Congestion Problem." *Techwire Asia*, August 28, 2020. https://techwireasia.com/2020/08/ai-powered-traffic-management-is-slashing-asias-congestion-problem.

Dobbie, Will, and Jae Song. 2015. "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection." *American Economic Review* 105 (3): 1272–311. https://doi.org/10.1257/aer.20130612.

Economist. 2017. "The World's Most Valuable Resource Is No Longer Oil, but Data." May 6, 2017. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Washington, DC: Administrative Conference of the United States. https://www.acus.gov/research-projects/agency-use-artificial-intelligence.

Epps, Willie J. Jr., and Jonathan M. Warren. 2020. "Artificial Intelligence: Now Being Deployed in the Field of Law." *The Judges' Journal* 59 (1): 16–39. https://www.americanbar.org/groups/judicial/publications/judges_journal/2020/winter/artificial-intelligence-now-being-deployed-the-field-law/.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *The Review of Economics and Statistics* 96 (1): 119–34. http://www.jstor.org/stable/43554917.

Fazekas, Mihály, Salvatore Sberna, and Alberto Vannucci. 2021. "The Extra-Legal Governance of Corruption: Tracing the Organization of Corruption in Public Procurement." *Governance: An International Journal of Policy, Administration, and Institutions* 35 (4): 1139–61. https://doi.org/10.1111/gove.12648.

Galanter, Marc. 1984. *Competing Equalities: Law and the Backward Classes in India*. Oxford: Oxford University Press.

Gauri, Varun. 2009. "Public Interest Litigation in India: Overreaching or Underachieving?" Policy Research Working Paper 5109, World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-5109.

Harwell, Drew, and Craig Timberg. 2021 "How America's Surveillance Networks Helped the FBI Catch the Capitol Mob." *Washington Post*, April 2, 2021. https://www.washingtonpost.com/technology/2021/04/02/capitol-siege-arrests-technology-fbi-privacy/.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Kannabiran, Kalpana. 2009. "Judicial Meanderings in Patriarchal Thickets: Litigating Sex Discrimination in India." *Economic and Political Weekly* 44 (44): 88–98. https://www.jstor.org/stable/25663738.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10: 113–74. https://doi.org/10.1093/jla/laz001.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2020. "Algorithms as Discrimination Detectors." *Proceedings of the National Academy of Sciences* 117 (48): 30096–100. https://doi.org/10.1073/pnas.191279011.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76. https://www.jstor.org/stable/30034076.

Krishnaswamy, Sudhir, Sindhu K. Sivakumar, and Shishir Bail. 2014. "Legal and Judicial Reform in India: A Call for Systemic and Empirical Approaches." *Journal of National Law University Delhi* 2 (1): 1–25. https://doi.org/10.1177/22774017 20140101.

Kumar, Vandana Ajay. 2012. "Judicial Delays in India: Causes & Remedies." *Journal of Law, Policy & Globalization* 4: 16–21. https://www.iiste.org/Journals/index.php/JLPG/article/view/2069.

Lal, Niharika. 2021. "How Traffic Cameras Issue E-Challans." *Times of India*, April 17, 2021. https://timesofindia.indiatimes.com/city/delhi/how-traffic-cameras-issue-e-challans/articleshow/82103731.cms.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521: 436–44. https://doi.org/10.1038/nature14539.

Misuraca, Gianluca, and Colin van Noordt. 2020. *AI Watch: Artificial Intelligence in Public Services*. EUR 30255 EN. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/039619.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics* 137.2 (May): 679–727. https://doi.org/10.1093/qje/qjab046.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Porrúa, Miguel, Mariano Lafuente, Edgardo Mosqueira, Benjamin Roseth, and Angela María Reyes, eds. 2021. *Digital Transformation and Public Employment: The Future of Government Work*. Washington, DC: Inter-American Development Bank. https://publications.iadb.org/publications/english/document/Digital-Transformation-and-Public-Employment-The-Future-of-Government-Work.pdf.

Ramos-Maqueda, Manuel, and Daniel L. Chen. 2021. "The Role of Justice in Development: The Data Revolution." Policy Research Working Paper 9720, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/35891.

Rigano, Christopher. 2018. "Using Artificial Intelligence to Address Criminal Justice Needs." National Institute of Justice, October 8, 2018. https://nij.ojp.gov/topics/articles/using-artificial-intelligence-address-criminal-justice-needs#citation--0.

Sadka, Joyce, Enrique Seira, and Christopher Woodruff. 2017. "Overconfidence and Settlement: Evidence from Mexican Labor Courts." Unpublished manuscript. http://www.enriqueseira.com/uploads/3/1/5/9/31599787/overconfidence_and_settlement_preliminary.pdf.

Sampat, Bhaven, and Heidi L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109 (1): 203–36. https://doi.org/10.1257/aer.20151398.

Schölkopf, Bernhard. 2022. "Causality for Machine Learning." In *Probabilistic and Causal Inference: The Works of Judea Pearl*, edited by Hector Geffner, Rina Dechter, and Joseph Y. Halpern, 765–804. New York: Association for Computing Machinery. https://doi.org/10.1145/3501714.3501755.

Waldrop, M. Mitchell. 2018. "Free Agents." *Science* 360 (6385): 144–47. https://doi.org/10.1126/science.360.6385.144.

Yong, ed. 2018. "A Popular Algorithm Is No Better at Predicting Crimes Than Random People." *Atlantic*, January 17, 2018. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.

CHAPTER 17

Government Analytics Using Data on Task and Project Completion

Imran Rasul, Daniel Rogger, Martin Williams, and Eleanor Florence Woodhouse

SUMMARY

Much government work consists of the completion of tasks, from creating major reports to undertaking training programs and procuring and building infrastructure. This chapter surveys a range of methods for measuring and analyzing task completion as a measure of the performance of government organizations, giving examples of where these methods have been implemented in practice. We discuss the strengths and limitations of each approach from the perspectives both of practice and research. While no single measure of task completion provides a holistic performance metric, when used appropriately, such measures can provide a powerful set of insights for analysts and managers alike.

ANALYTICS IN PRACTICE

- Much government activity can be conceived as discrete tasks: bounded pieces of work with definite outputs. Public sector planning is often organized around the achievement of specific thresholds; the completion of planning, strategy, or budgetary documents; or the delivery of infrastructure projects. *Task completion* is a useful conception of government activity because it allows analysts to assess public performance in a standardized way across organizations and types of activity.
- Assessing government performance based solely on the passing of legislation or the delivery of frontline services misses a substantial component of government work. Using a task completion approach pushes analysts to better encapsulate the breadth of work undertaken by public administration across government. It thus pushes analysts to engage with the full set of government tasks.

Imran Rasul is a professor in the Department of Economics, University College London. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Martin Williams is an associate professor in the Blavatnik School of Government, University of Oxford. Eleanor Florence Woodhouse is an assistant professor in the Department of Political Science and School of Public Policy, University College London.

- A task completion approach also allows for the investigation of which units and organizations are most likely to initiate, make progress on, and complete tasks. Though not a full picture of government work—it is complementary to the analysis of process quality or sector-specific measures of quality, for example—it allows for a rigorous approach to comparisons frequently made implicitly in budgetary and management decisions.
- Collecting data across projects on determinants of progress, such as overruns, and matching them to input data, such as budget disbursements, allows for a coherent investigation of the mechanisms driving task progress across government or within specific settings.
- Attempting to assess task completion in a consistent way across government is complicated by the fact that tasks vary in nature, size, and complexity. By collecting data on these features of a task, analysts can go some way toward alleviating concerns over the variability of the tasks being considered. For example, analysis can be undertaken within particular types of task or size, and complexity can be conditioned on in any analysis. An important distinction in the existing literature is how to integrate the analysis of tasks related to the creation of physical infrastructure and tasks related to administration.

INTRODUCTION

A fundamental question for government scholars and practitioners alike is whether governments are performing their functions well. What these functions are and what performing "well" means in practice are complex issues in the public sector, given the diverse tasks undertaken and their often indeterminate nature. Despite the importance of these questions, there is little consensus as to how to define government effectiveness in a coherent way across the public service or how to measure it within a unified approach across governments' diverse task environments (Rainey 2009; Talbot 2010). Such considerations have practical importance because government entities, such as political oversight or central budget authorities, frequently have to make implicit comparisons between the relative functioning of public agencies. For example, when drawing up a budget, public sector managers must make some comparison of the likely use of funds across units and whether these funds will eventually result in the intended outputs of those units, however varied the tasks are in scope. From an analytical perspective, the more comprehensive a measure of government functioning, the greater the capacity of analytical methods to draw insights from the best-performing parts of government.

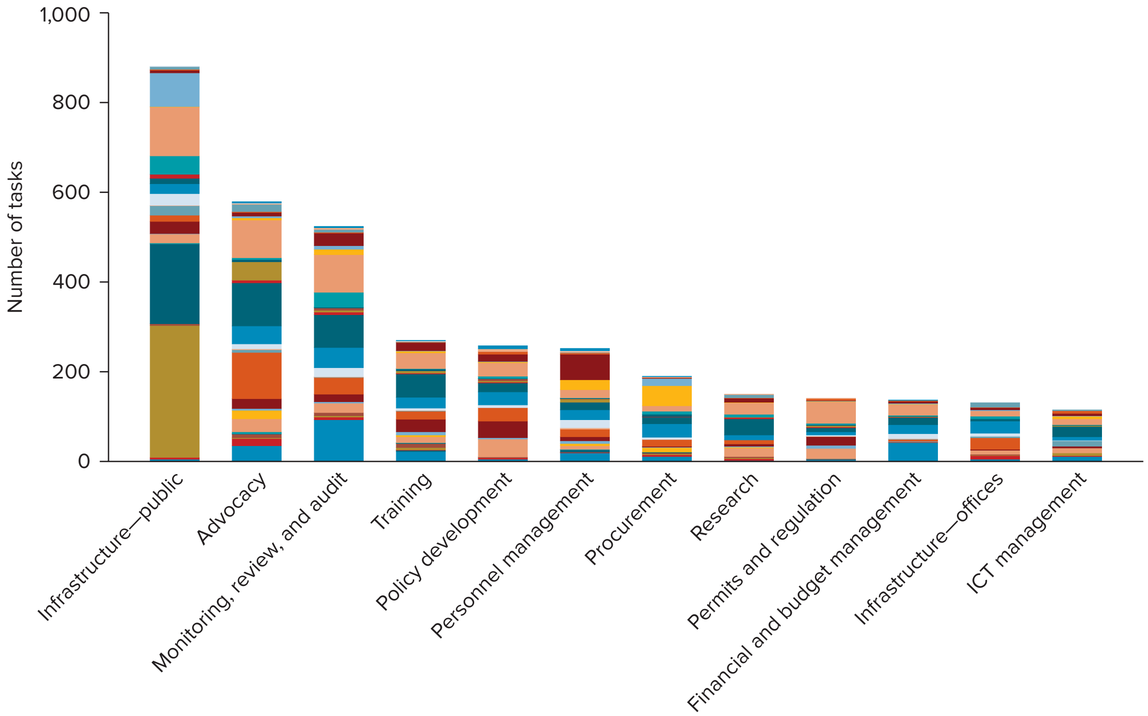
Much government activity can be conceived as discrete *tasks*: bounded pieces of work with definite outputs. Public sector planning is often organized around the achievement of specific thresholds; the completion of planning, strategy, or budgetary documents; or the delivery of projects. Government *projects* are also often conceived as bounded activities with definite outputs but frequently encompass multiple tasks within a wider conception of completion. *Task completion* (or project completion) is thus a useful conception of government activity—however that activity is conceived—because it allows analysts to assess public performance in a standardized way across organizations and types of activity. This kind of assessment contrasts with continuous regulatory monitoring, the assessment of the stability of citizens' access to frontline services, and the equity of activities related to redistribution, which are better understood as the evaluation of ongoing processes. In this chapter, we propose a way to leverage data on task completion to assess the government's effectiveness across its diverse task environments and learn from related analysis. We argue that by utilizing a unified framework for task completion, analysts can assess whether a government or government agency "does well what it is supposed to do, whether people. . . work hard and well, whether the actions and procedures of the agency and its members help achieve its mission, and, in the end, whether it actually achieves its mission" (Lee and Whitford 2009, 251; paraphrasing Rainey and Steinbauer 1999).

Task comparison is useful for three core reasons. First, task completion is a concept that can be applied across much government work, thus allowing for a broad consideration of government functioning.

We believe that by working with a task completion framework, analysts can gain a fuller and more accurate picture of the functions of government that reflects the full range of government activities—from human resource management to policy definition, infrastructure planning and implementation, service delivery, and audit and evaluation. We know little about the full distribution of tasks that public administrators undertake. As shown in figure 17.1, the few studies that do apply a task completion framework find that administrators undertake a vast range of activities—from advocacy to auditing and monitoring to planning—that go well beyond infrastructure and service delivery, the activities that are usually considered in the academic literature.

Figure 17.1 displays the frequency of the most prevalent tasks undertaken by Ghanaian public officials in their daily duties. Infrastructure provision for the public (rather than upgrading government facilities) is the most common activity, partly motivating our particular attention to it in this chapter. However, the figure indicates the broad diversity of tasks undertaken by the public service. The distinct colors in each bar of the histogram indicate different organizations undertaking that type of task. Thus, it is clear that each type of task is undertaken by many different organizations. Second, a task completion framework pushes analysts to think carefully about the characteristics of each of the tasks assessed. We define projects above as collections of tasks; an obvious question is how to apply boundaries to tasks or projects uniformly across government. There is very limited research on the characteristics of the tasks undertaken by public administrators and how to assess whether they are being undertaken adequately. The task completion framework pushes analysts to think in detail about the activities that administrators engage in and how successfully they do so. That is to say, they must think not only about whether a bridge is completed but what the full conception of the bridge project is, whether the bridge was of a complex design that was hard to implement, whether the quality of the implementation of the bridge is adequate, whether it was completed within a reasonable time frame given the complexity of the project, and so on.

FIGURE 17.1 Task Types across Organizations



Source: Rasul, Rogger, and Williams 2021.
Note: The task type classification refers to the primary classification for each output. Each color in a column represents an organization implementing tasks of that type, but the same color across columns may represent multiple organizations. Figures represent all 30 organizations with coded task data. ICT = information and communication technology.

Third, by creating comparators from across government, an integrated measurement approach yields analytical benefits that more than make up for the losses from abstraction for many types of analysis. Analysts can investigate the determinants of successful task completion from a large sample, with varying management environments, buffeted by differential shocks, and on which a greater range of statistical methods can be effectively applied. The task completion framework has the advantage of capturing a wide range of activities and being comparable across departments. Thus, the framework can pool task types to allow analysts to draw conclusions about government effectiveness more broadly, rather than, for example, allowing only for inferences about a specific type of task (for example, the delivery of a specific service, such as passport processing times, versus a range of government tasks that are indicative of administrative effectiveness more broadly). By leveraging these data—on the nature, size, and complexity of tasks—analyses can be undertaken within particular types of tasks—for example, distinguishing between the completion of physical and nonphysical outputs. Taking the analysis a step further, having measures of task completion that are comparable across teams or organizations can enable researchers to identify the determinants of task completion—although this entails its own methodological challenges and is beyond the scope of this measurement-focused chapter.

Much of the existing literature that seeks to describe how well governments perform their functions focuses on upstream steps in the public sector production process (such as budgetary inputs or processes), which are useful for management (but not so relevant to the public) (Andrews et al. 2005; Lewis 2007; Nistotskaya and Cingolani 2016; Rauch and Evans 2000). Or it focuses on final outcomes (such as goods provided or services delivered), which are relevant to the public (but not always so useful for management) (Ammons 2014; Boyne 2003; Carter, Klein, and Day 1992; Hefetz and Warner 2012). The completion of tasks and projects falls between these two approaches: it is useful for management and relevant to the public. The task completion framework helps analysts address the gap between inputs and final outcomes in terms of how they measure government performance. It gives analysts a way to engage with the full distribution of government tasks and to assess the characteristics of the tasks themselves. The task approach outlined in this chapter is closely aligned with the discussion in chapter 15 of *The Government Analytics Handbook*. There, the relevant task is the processing of an administrative case. Clearly, there are other types of tasks in government, and this chapter aims to present a framework that can encapsulate them all. Given the scale of case processing in government, however, chapter 15 presents a discussion specific to that type of task. Related arguments can be made for chapter 12 on procurement, chapter 14 on customs, and chapter 29 on indicators of service delivery. Across these chapters, the *Handbook* provides discussions of the specific analytical opportunities afforded by different types of government activity. These chapters contain some common elements, such as discussions of some form of complexity in relation to the task under focus. This chapter showcases the considerations required for an integrated approach across task types.

Through the task completion framework, we aim to encourage practitioners and scholars alike to conceive of government activity more broadly and to leverage widely available data sources, such as government progress reports or independent public expenditure reviews, to do so. As well as being widely available, these kinds of objective data are highly valuable because they usually cover a wide range of different task types performed by numerous different agencies and departments.

This chapter continues as follows. First, we conceptualize government work as task completion. Second, we use tasks related to the creation of physical infrastructure to illustrate the task completion framework. Third, we show how the framework applies to other types of tasks. Fourth, we explore how to measure task characteristics (considering the complexity of tasks and their ex ante and ex post clarity). Fifth, we discuss key challenges in integrating these measures with each other and into management practice. Finally, we conclude.

CONCEPTUALIZING GOVERNMENT WORK AS THE COMPLETION OF TASKS

Much of the literature in public administration has focused on how to measure government effectiveness by relying on the tasks of single agencies (Brown and Coulter 1983; Ho and Cho 2017; Lu 2016), on a set of agencies undertaking the same task (Fenizia 2022), or on a broad conception of the central government as a single entity (Lee and Whitford 2009). These approaches limit analysis to a single conception of government effectiveness, which in the case of a single agency or sector, can be precisely defined. However, almost by definition, this limits analysis to a subset of government work and thus raises concerns over what such analysis tells us about government performance as a whole or how performance in one area of government affects other areas.

In addition to measuring government effectiveness on the basis of a partial vision of government, many studies that have sought to investigate government effectiveness have relied on perception-based measures of effectiveness, based on the evaluations of either government employees or external stakeholders and experts (Poister and Streib 1999; Thomas, Poister, and Ertas 2009; Walker et al. 2018). Such measures are frequently available only at an aggregate or even country level because of how distant these individuals are from actual government tasks, and they frequently assess not the outputs of those tasks directly but some perception of "general effectiveness."[1]

Objective measures of government functioning have frequently been eschewed because of obstacles related to data availability, their purported inability to capture the complexity of government work, or conflicting understandings of what effectiveness means. However, many government agencies produce their own reports on the progress they have made across the full distribution of their work. Similarly, agencies often have administrative data on the totality of their activities that provide quantities related to the complexity of task completion that can be repurposed for analytics. These data are collected for management and reporting purposes as part of the daily duties of agency staff. These reports frequently contain characteristics of the tasks undertaken and progress indicators outlining how far tasks have progressed. These reports can be the basis of an integrated analysis of government functioning.

For tasks related to physical infrastructure and administration, analysts can use quantities from these reports, or similar primary data collection, to conceptualize government work in a unified task completion framework. The following discussion of the strengths and limitations or challenges of such an approach focuses on a small set of research papers that have applied a task completion framework to the assessment of government functioning. It thus aims to illustrate the utility of the task completion framework rather than being in any way comprehensive. Where relevant, we provide a number of examples of how public officials have taken a similar approach.

We rely on two simple definitions throughout the chapter. First, a *task* is the bounded activity for which a given organization, team, or individual in the government is responsible. Second, an *output* is the final product a government organization, team, or individual delivers to society. An output is the result of a successful task. In government performance assessment, outputs are defined as "the goods or services produced by government agencies (e.g., teaching hours delivered, welfare benefits assessed and paid)."[2] An example of a government task might be developing a draft competition policy or organizing a stakeholder meeting to validate the draft competition policy (Rasul, Rogger, and Williams 2021, appendix). The corresponding outputs would be the draft competition policy itself and the holding of the stakeholder meeting. These tasks are usually repeated and are completed within varying time frames, depending on the complexity and urgency of the activity at hand.

More granular guidance on how to define a task is challenged by the fact that the appropriate conception of a task will vary by the focus of the analysis. However, to illustrate common conceptions, some examples from the analyses that will be discussed in this chapter include the design, drilling, and development of a water well (including all taps linked to a single source of water); the design, construction, and finishing of a school; the renovation of a neighborhood sewage system; a full maintenance review and associated activities, such as resurfacing, to bring a road up to a functioning state as determined by local standards; the development of a new public health curriculum for primary school students; and the updating of a human resources management information system with current personnel characteristics for all health-related agencies.

By conceiving all government activity as consisting of tasks with intended outputs, analysts can construct a standardized measure of government performance and can gather multiple tasks together to assess government performance across teams within an organization, across organizations, and over time.

Government performance can be defined as the frequency with which particular government actors are able to produce outputs from corresponding tasks. We now turn to considerations in the definition of a task or project and an output in the case of physical and nonphysical outputs.

Physical Outputs

We first consider a task completion framework as it pertains to the accomplishment of physical infrastructure, or, more precisely, tasks relating to the production of physical outputs. In lower-middle-income countries in particular, the noncompletion of infrastructure projects is a widespread and costly phenomenon, with recent estimates suggesting that over one-third of the infrastructure projects started in these countries are not completed (Rasul and Rogger 2018; Williams 2017).

We focus on task completion measures developed from coding administrative data that are at least somewhat comparable across organizations and can be implemented at scale, rather than on performance audits of specific programs (for example, by national audit offices or international financial institutions' internal performance reports) or on the evaluation of performance against key performance indicators (for example, in leadership performance contracts or through central target-setting mechanisms). Many governments or government agencies have infrastructure-project-tracking databases (either electronic or in paper-based files). These records may be for implementation management, for budgeting and fiduciary reasons, or for audit and evaluation. These databases keep records of how far physical projects have been implemented relative to their planned scope.

For example, in Nigeria, Rasul and Rogger (2018) use independent engineering assessments of thousands of projects from across the government implemented by the Nigerian public service to assess the functioning of government agencies. They complement this with a management survey in the agencies responsible for the projects and examine how management practices matter for the completion rates of projects. The analysis exploits a specific period in the Nigerian public service when "the activities of public bureaucracies were subject to detailed and independent scrutiny" (2) and a special office was set up to track the quality of the project implementation of a broad subset of government activities. This was due to an effort by the presidency to independently verify the status of many of the public infrastructure projects funded by the proceeds of debt relief and implemented by agencies across the federal government. The records of this tracking initiative allowed the authors to quantify both the extent of project implementation and the assessment of the quality of the public goods provided.

A second application of the task completion framework to an empirical setting examining physical outputs is provided by Williams (2017), who collects, digitizes, and codes district annual progress reports in Ghana. These reports, which are written annually by each district's bureaucracy and submitted to the central government, include a table listing basic information about projects that were ongoing or active during the calendar year. Such reports are widely produced but not frequently available in a digital format or used for government analytics. The potential of these data for useful insights into government performance is great. Williams uses the reports on physical projects to examine the determinants of noncompletion, presenting evidence that corruption and clientelism are not to blame but rather a dynamic collective action process among political actors facing commitment problems in contexts of limited resources.

Similarly, Bancalari (2022) uses district administrative data on sewerage projects in Peru to explore the social costs of unfinished projects. She uses a combination of mortality statistics, viability studies, annual budget reports on sewerage projects (which allow her to identify unfinished and completed projects), spatial topography data, and population data in order to provide evidence that infant mortality and under-five mortality increase with increases in unfinished sewerage projects. She also finds that mayors who are better connected to the national parliament are able to complete more projects.

Beyond using administrative data, analysts have also undertaken primary fieldwork to explore the completion of physical projects. For example, Olken (2007) uses various surveys on villages, households, individuals, and the assessments of engineering experts to investigate the level of corruption involved in building roads in Indonesia. Olken is able to produce a measure of corruption in terms of missing expenditures by calculating discrepancies between official project costs and an independent engineer's estimate of costs defined by the survey responses. Primary field activity also allows analysts to undertake randomized controlled trials of potential policies to improve government functioning. In the case of Olken (2007), randomized audits of villages are used to estimate the effect of top-down monitoring on the quality of government outputs: in this case, the building of roads. Such a research design and measure are highly valuable and capture a very important feature of government activity, although they come at a high cost in terms of the resources needed to capture these government tasks.

Other papers have studied the maintenance rather than the construction of physical outputs. In these cases, task completion is the effective continuation of physical outputs. Once again using primary fieldwork to collect required data, Khwaja (2009) uses survey team site visits and household surveys to measure the maintenance of infrastructure projects in rural communities in northern Pakistan (Baltistan) as a form of task completion. Maintenance here is measured through surveys of expert engineers who assess the maintenance of infrastructure projects in terms of their physical state (that is, how they compare to their initial condition), their functional state (that is, the percentage of the initial project purpose satisfied), and their maintenance-work state (that is, the percentage of required maintenance that needs to be carried out). Khwaja (2009) uses these data to examine whether project design can improve collective success in maintaining local infrastructure. The paper presents within-community evidence that project design makes a difference to maintenance levels: "designing projects that face fewer appropriation risks through better leadership and lower complexity, eliciting greater local information through the involvement of community members in project decisions, investing in simpler and existing projects, ensuring a more equitable distribution of project returns, and emulating NGOs can substantially improve project performance even in communities with low social capital" (Khwaja 2009, 913).

We have seen several examples of "government analytics" that seek to measure the completion rate of tasks related to the provision (or maintenance) of physical outputs. From Nigerian federally approved social sector projects, such as providing dams, boreholes, and roads, to Indonesian road building, analysts have defined measures of task completion based on physical outputs. The analysis has used administrative data, existing household surveys, and primary fieldwork (sometimes in combination with one another) to generate insights into the determinants of government functioning.

These papers measure task completion in a series of different ways that all aim to capture the underlying phenomenon of what share of the intended outputs are completed. But there are important commonalities to their approaches. First, the definition of a task or project is determined by a common, or consensus, engineering judgment that crosses institutional boundaries. Thus, though a ministry of urban development may bundle the creation of multiple water distribution points, the building of a health center, and road repaving into a single "slum upgrading" project, the analysts discussed above split these groupings into individual components that would be recognizable across settings, and thus across government. A water distribution point will be conceived as a discrete task whether it is a component of a project in an agriculture, education, health, or water infrastructure project. The wider point is that an external conception of what makes up a discrete activity, such as the common engineering conception of a water distribution point, provides discipline on the boundaries of what is conceived as a single task for any analytical exercise.

Second, within these conceptions of projects, an externally valid notion of completion and progress can be applied. For example, the threshold for a water distribution point is that it produces a sufficient flow of water over a sustained period for it to be considered "completed." Williams (2017) uses the engineering assessments included in administrative data to categorize projects into bins of "complete" (for values such as "complete" or "installed and in use") or "incomplete" (for values such as "ongoing" or "lintel level"). Rasul and Rogger (2018) use engineering documents specific to each project to define a percentage scale of completion for each project allowing for a more granular measure of task progress, mapping them along a 0–1 continuum. Thus, highly varied project designs are mapped into a common scale of progress by consideration of the underlying production function for that class of infrastructure. What constitutes a halfway point in the development of a water distribution point and a dam will differ, but both can be feasibly assessed as having a halfway point.

Third, notions of scale or complexity can be determined from project documentation, providing a basis for improving the credibility of comparisons across tasks. As will be discussed in section three, there is little consensus about how to proxy such complexity across tasks. The literature on complexity in project management and engineering emphasizes the multiple dimensions of complexity (Remington and Pollack 2007). This can be seen as a strength, in that a common framework for coding complexity can be flexibly adapted to the particular environment or analytical question. In the above examples, planned (rather than expended) budget is frequently used as one way to proxy scale and complexity. The challenge is that the planned budget may already be determined by features related to task completion. For example, the history of task completion at an agency may influence contemporary budget allocations.

For this reason, physical infrastructure tasks can be conceptualized and judged by external conceptions and scales that discipline the analysis. A strength of these measurement options is that they offer a relatively clear, unambiguous measure of task completion. Fundamentally, generating a sensible binary completion value requires understanding how progress maps onto public benefit (for example, an 80 percent finished water distribution point is of zero public value). With this basic knowledge across project types, task completion indicators can be computed for the full range of physical outputs produced by government.[3]

However, this type of task completion framework measurement also comes with limitations. It is easier to measure completion than quality with these types of measures. Quality is typically multifaceted, such that it is more demanding to collect and harmonize into an indicator that can be applied across project types. In Rasul and Rogger (2018), assessors evaluate the quality of infrastructure projects on a coarse scale related to broad indicators that implementation is of "satisfactory" quality relative to professional engineering norms. Analysis can then be defined by whether tasks are, first, completed, and second, completed to a satisfactory level of quality. Administrative progress reports vary in their information content but tend to assume quality and focus on the technical fulfillment of different stages in the completion process.

One way to gain information on quality is to undertake independent audits or checks, though these tend to be highly resource intensive relative to the use of administrative data. For example, Olken (2007, 203) relies on a team of engineers and surveyors to assess the quality of road infrastructure, who "after the projects were completed, dug core samples in each road to estimate the quantity of materials used, surveyed local suppliers to estimate prices, and interviewed villagers to determine the wages paid on the project." From these data, Olken constructs an independent estimate of the quality of each road project.

Some conceptions of quality go as far as the citizen experience of the good or service or how durable or well managed it is. Rasul and Rogger (2018) also include assessments of citizen satisfaction with the project overall as determined by civil society assessors, but such data are almost never available in administrative records and have to be collected independently.

There are also issues pertaining to the reliability and interpretation of task completion that are worth highlighting. First, doubts may be raised when the progress reports that act as the foundation for task completion assessments are provided by the same public organizations that undertake the projects themselves (see the discussion in chapter 4). For this reason, they may not constitute reliable measures of progress, or at least may be perceived as unreliable. The problem is whether organizations can be considered reliable in their assessments of their own work. Measures of task progress sourced from administrative data must thus be used with care and, ideally, validated against a separate (independent) measure of progress. A good example of this comes from Rasul, Rogger, and Williams (2021), who match a subsample of tasks from government-produced progress reports to task audits conducted by external auditors in a separate process.[4] Such validation exercises can be very helpful in providing evidence that the measures produced by government organizations on their own performance are credible, thus salvaging an important source of data that might otherwise be deemed unusable.

Additionally, *noncompletion* can mean different things depending on how the timeline of infrastructure procurement, construction, and operation is organized. This is especially clear in the case described by Bancalari (2022), where it is hard to establish whether the effect uncovered is an effect of noncompletion or delays and cost overruns in delivery.[5] It can be hard to distinguish noncompletion (a project will remain unfinished) from delays (a project will be completed but is running over schedule). Here, the point in time when one decides to measure completion and the initial time frame set for a given task become important and can affect how one interprets task noncompletion.

Finally, a separate issue pertains to whether tasks are completed as planned, not simply whether they are completed. The existing literature from management studies has mostly focused on overruns, delays, and over-estimated benefits rather than on noncompletion per se (Bertelli, Mele, and Whitford 2020; Post 2014). This body of literature tends to focus on the service and goods delivery side of government rather than on the full range of government activities. However, it is an important complement to the task completion framework precisely because it focuses on whether the tasks governments undertake are being completed *and* are being completed in the time frame and up to the standard that they were planned for. For example, a vast body of literature emphasizes the value-for-money or cost calculations of infrastructure projects rather than the efficiency or effectiveness of the processes via which they are delivered (for example, Engel, Fischer, and Galetovic 2013). Scholars such as Flyvbjerg (2009, 344) have argued that the "worst" infrastructure gets built because "ex ante estimates of costs and benefits are often very different from actual ex post costs and benefits. For large infrastructure projects the consequences are cost overruns, benefit shortfalls, and the systematic underestimation of risks."

Nonphysical Outputs

Now we turn to the task completion framework as it applies to the production of nonphysical outputs. Examples of nonphysical outputs are auditing activities, identifying localities where infrastructure is required, raising awareness about a given social benefit scheme, or planning for management meetings. These types of task, in short, involve government activities that pertain to the less visible side of government: not delivery in the form of physical goods or services but the planning, monitoring, information sharing, reviewing, and organizational tasks of government.

Rasul, Rogger, and Williams (2021) use administrative data on the roughly 3,600 tasks that civil servants undertook in the Ghanaian civil service in 2015. The data on these tasks are extracted from quarterly progress reports and represent the full spectrum of government activities. As can be seen from figure 17.1, a large proportion of these tasks are related to nonphysical outputs. For each type of task, in relation to both physical and nonphysical outputs, the researchers identify a scheme by which to judge task completion by allocating a threshold of progress to represent completion for each task type.

Rasul, Rogger, and Williams (2021) also collect data on the management practices under which these tasks are undertaken via in-person surveys with managers covering six dimensions of management: roles, flexibility, incentives, monitoring, staffing, and targets. Together, the task and management data allow for an assessment of how public sector management impacts task completion, allowing for the comparison of the effect of management practices on the same tasks across different organizations. Their data demonstrate, first, that there is substantial variation in task completion across types of task and across civil service organizations. Second, there is also substantial variation in the types of management practice that public servants are subject to across organizations, and the nature of management correlates significantly with task completion rates.

Integrating the analysis of tasks related to both physical and nonphysical outputs allows for a broad assessment of government functioning, encompassing the many interactions between tasks of different natures. Such a holistic approach also enables the assessment of tasks with different underlying characteristics, which has long been identified as a core determinant of government performance.

Rasul, Rogger, and Williams (2021) are interested in exploring whether different management techniques are differentially effective, depending on the clarity of the task in project documents. They build on the literature arguing that where settings involve intensive multitasking, coordination, or instability, management techniques using monitoring and incentive systems are likely to backfire. The question, as they put it, harking back to the Friedrich vs. Finer debate (Finer 1941; Friedrich 1940), is "to what extent should [civil servants] be managed with the carrot and the stick, and to what extent should they be empowered with the discretion associated with other professions?" (Rasul, Rogger, and Williams 2021, 262). Their central finding is that there are "positive conditional associations between task completion and organizational practices related to autonomy and discretion, but negative conditional associations with management practices related to incentives and monitoring" (Rasul, Rogger, and Williams 2021, 274).[6] The authors distinguish between government tasks with high and low ex ante and ex post clarity. Incentives and monitoring-intensive management approaches are hypothesized (and found) to be more effective when ex ante task clarity is high (and ex post task clarity is low), whereas autonomy and discretion-intensive management approaches are relatively more effective when ex ante task clarity is low (and ex post task clarity is high).

The main contribution of Rasul, Rogger, and Williams (2021) to the discussion of this chapter is providing a holistic, output-based organizational performance metric. However, their approach also takes a holistic account of the multifarious nature of management practices in government and showcases the value of combining such data. The authors "conceptualize management in public organizations as a portfolio of practices that correspond to different aspects of management, each of which may be implemented more or less well. Bureaucracies may differ in their intended management styles, that is, what bundle of management practices they are aiming to implement, and may also differ in how well they are executing these practices" (262). That is, there is a combination of both intent and implementation when it comes to management practices that may affect the effectiveness of an organization. The task completion framework, with its focus on both the breadth of activities that government bodies undertake and on the detail of the characteristics of government tasks, represents an important stepping stone toward a more holistic and realistic understanding of government work and effectiveness.

A separate body of literature that brings together tasks and projects of distinct types into a single analytical framework is the literature on donor projects. For example, using data on the development projects of international development organizations (IDOs)—specifically, eight agencies—including project outcome ratings of holistic project performance, Honig (2019) investigates the success of IDO projects according to internal administrative evaluations. The success ratings are undertaken by IDO administrators, who employ a consistent underlying construct across different IDOs, with an OECD-wide standard in place. These ratings are combined with a host of other variables capturing various features of the projects (for example, their start and end dates, whether there was an IDO office presence in situ, what the sector of the project is, etc.).

Honig (2019, 172, 196) uses "variation in recipient-country environments as a source of exogenous variation in the net effects of tight principal control" to find that "less politically constrained IDOs see systematically lower performance declines in more unpredictable contexts than do their more-constrained peers." That is to say that monitoring comes with costs in terms of reducing the ability of agents to adapt, particularly in less predictable environments.

Similarly, Denizer, Kaufmann, and Kraay (2013, 288) leverage a data set of over 6,000 World Bank projects (over 130 developing countries) to "simultaneously investigate the relative importance of country-level 'macro' factors and project-level 'micro' factors in driving project level outcomes." The authors leverage Implementation Status Results Reports completed by task team leaders at the World Bank, which report on the status of the projects, as well as Implementation Completion Reports, which include a "subjective assessment of the degree to which the project was successful in meeting its development objective" (290), plus more detailed ex post evaluations of about 25 percent of projects, in order to assess project outcomes. They find that roughly 80 percent of the variation in project outputs occurs across projects within countries, rather than between countries, and that a large set of project-level variables influence aid project outputs.

A related but separate body of literature considers nonphysical task completion by frontline delivery agents. For example, using the case of the Department of Health in Pakistan, Khan (2021) undertakes an experiment in which he randomly emphasizes the department's public health mission to community health workers, provides performance-linked financial incentives, or does both. He measures task completion through a combination of internal administrative data on service delivery and outputs, gathered as part of routine monitoring processes, and household surveys of beneficiaries. Mansoor, Genicot, and Mansuri (2021), instead, use the case of the agriculture extension department in Punjab, Pakistan, to measure both objective task completion and supervisors' subjective perception of performance. They measure this through a combination of household surveys and data from a mobile phone tracking app that frontline providers use to guide and record their work.

Analogous to the physical outputs case, then, to apply a task completion framework to tasks related to nonphysical outputs, we require common definitions of tasks that cross institutional boundaries, externally valid notions of completion and progress, and notions of scale or complexity. Such external standards for what completion and quality look like across institutions are rare, but they do exist in some fields, such as health care (see the example of the joint health inspection checklist in Bedoya, Das, and

Dolinger [forthcoming]). Creating an analogous approach to these issues for tasks related to nonphysical outputs ensures comparability with tasks related to physical outputs. However, they are also valid pillars for analysis even within the set of tasks related to nonphysical outputs only.

For many tasks related to nonphysical outputs, there are, in fact, natural conceptions of task and output. For example, a curriculum development project is only complete once the curriculum is signed off on by all stakeholders, and an infrastructure monitoring program is only complete when a census of the relevant infrastructure has been completed. Similarly, such an approach can be developed for measures of progress. The curriculum development will typically be broken down into substantive stages in planning documents, and each of these stages can be assigned a proportion of progress. In the infrastructure monitoring case, a simple proportion of infrastructure projects assessed, perhaps weighted by scale or distance measures, seems fitting. Not all cases will be so clear-cut. To identify a consensus definition of task by task type that could apply across institutional boundaries, Rasul, Rogger, and Williams (2021) employ public servants at a central analytics office (in the Ghanaian case, this was the Management Services Department) to agree on relevant definitions using data from across government. As will be seen below, this team also defines measures of complexity relevant across the full set of tasks, including (as mentioned above) clarity of design. Decisions as to how to define task completion will be influenced by, but then very much influence, the approach to data collection. Table 17.1 summarizes the approaches analysts have taken to measuring task completion for physical and nonphysical outputs.

While we have focused our discussion mainly on research-oriented examples of measuring task completion, there are also examples of government organizations' use of task completion measures for tasks related to physical and nonphysical outputs—with varying degrees of formality. For example, the United Kingdom Infrastructure and Projects Authority conducts in-depth annual monitoring of all large-scale projects across UK government departments—235 as of 2022—and publishes an annual report with a red/amber/green project outlook rating (IPA 2022). At the other end of the formality and resource-intensiveness spectrum, in their engagement with the government of Ghana in 2015–16 in the course of conducting fieldwork, Rasul, Rogger, and Williams (2021) found that Ghana's Environmental Protection Agency tallied the percentage of outputs completed by each unit in their quarterly and annual reports for internal monitoring purposes. In between these two examples, the Uganda Ministry of Finance and the International Growth Centre (IGC) have partnered to apply Rasul, Rogger, and Williams's (2021) coding methodology (supplemented with qualitative interviews) to monitor the implementation progress of 153 priority policy actions across government and examine the determinants of their completion (Kaddu, Aguilera, and Carson n.d.). And of course, as argued above, many if not most government organizations do some form of task or output completion measurement in the course of their own routine reporting—despite most not taking the next step of using these data for formal analytical purposes.

TABLE 17.1 Selected Measures of Task Completion

Task type	Potential data sources and measurement methods	Selected examples
Physical tasks	• Site visits by expert teams • Site visits by survey teams • Compilation from other secondary sources (for example, media or project reports) • Administrative data from periodic reports	Olken (2007); Rasul and Rogger (2018) Khwaja (2009) Flyvbjerg, Skamris Holm, and Buhl (2002); Williams (2017) Bancalari (2022)
Nonphysical tasks	• Surveys of beneficiaries or citizens • Tracking app used by frontline personnel • Administrative data from periodic reports • Administrative data from internal management monitoring sources • International donor project evaluation reports	Khan (2021) Mansoor, Genicot, and Mansuri (2021) Rasul, Rogger, and Williams (2021) Mansoor, Genicot, and Mansuri (2021), Khan (2021) Denizer, Kaufmann, and Kraay (2013), Honig (2019)

Source: Original table for this publication.

Applying the task completion framework to nonphysical outputs comes with its challenges and limitations, building on those noted above for physical outputs. The issues pertaining to assessing the quality of the implementation of tasks related to nonphysical outputs are twofold. First, establishing how to assess quality is not straightforward, and second, the nature of a task can render the difficulty of assessing quality differentially complex. For instance, if the task one is measuring is the completion of a bridge, one first has to establish the criteria that dictate whether it can be considered a high- or low-quality bridge, whereas if one is also considering nonphysical outputs, such as the development of an education strategy, then one faces a potentially even greater challenge in defining what “high-quality” means for such a project (see Bertelli et al. [2021] for a discussion of this).

There are certain types of task, in short, for which establishing objective benchmarks is more difficult than for others. It does not seem like too much of a leap, for example, to hypothesize that the nonphysical tasks we have considered in this section might frequently be more complex to benchmark in terms of quality than the physical outputs we described earlier.

This difficulty creates discontinuity in measurement quality across physical and nonphysical goods, which, in turn, raises the issue of the potential endogeneity of task and output selection. That is to say, out of the universe of possible government tasks, the types of tasks we are best able to measure may be correlated with particular outputs. This could provide us with a distorted image of the types of tasks that are conducive to producing certain outputs.

MEASURING TASK CHARACTERISTICS

As we outline in the introduction to this chapter, a task completion framework is helpful to analysts in two main senses. First, it pushes analysts to better encapsulate the breadth of work undertaken by public administration across government. Second, it encourages them to think carefully about the characteristics of the tasks themselves. In this section, we will focus on the latter feature of a task completion framework: how to measure task characteristics.

There are, naturally, a plethora of government task characteristics on which one could focus. Here, we will focus on several of the most relevant characteristics from the perspective of implementation. We concentrate on implementation because it has been the focus of the literature on task completion and because it is of direct relevance to the work of practitioners, the intended audience of this chapter.

We start by considering task complexity. When examining government outputs and their relationship to phenomena such as management practices, government turnover, or risk environment, it is often important to understand their relationship with project or task complexity (Prendergast 2002). This is because the complexity of the task will frequently be strongly correlated with variables such as time to completion, total cost, the likelihood of delays, and customer satisfaction, which might be of interest to scholars or practitioners interested in task completion. Table 17.2 summarizes how the analysts described in this paper have attempted to implement measurement of complexity, as well as how authors have measured two further important features of government tasks to which we will turn next, visibility and clarity.

Rasul and Rogger (2018, 12), in their study of public services in the Nigerian civil service, create complexity indicators that capture “the number of inputs and methods needed for the project, the ease with which the relevant labour and capital inputs can be obtained, ambiguities in design and project implementation, and the overall difficulty in managing the project.” They are thus able to condition on the complexity of projects along these margins when exploring the relationship between managerial practices and project completion rates. However, such an approach does not account for the fact that worse-performing agencies may be assigned easier (less complex) tasks in a dynamic process over time. So in background work for the study, Rasul and Rogger assess the extent to which there was sorting of projects across agencies by their level of complexity, a task only feasible with appropriate measures. They do not find any evidence of such sorting.

TABLE 17.2 Selected Measures of Task Characteristics

Task or project characteristic	Potential data sources and measurement methods	Selected examples
Complexity	• Expert data coding from site visits • Semi-expert data coding from administrative reports • International donor project evaluation reports	Khwaja (2009); Rasul and Rogger (2018); Rasul, Rogger, and Williams (2021); Denizer, Kaufmann, and Kraay (2013)
Visibility	• Project-level data from infrastructure database assembled from governmental and financial sources	Woodhouse (2022)
Clarity (ex ante and ex post)	• Semi-expert data coding from administrative reports	Rasul, Rogger, and Williams (2021)

Source: Original table for this publication.

Khwaja (2009, 915), instead, captures project complexity by creating an index that measures whether "the project has greater cash (for outside labor and materials) versus noncash (local labor and materials) maintenance requirements, . . . the community has had little experience with such a project, and . . . the project requires greater skilled labor or spare parts relative to unskilled labor for project maintenance." In this way, he is able to distinguish group-specific features—such as social capital—from features of task design—such as degree of complexity—in order to better understand their relative importance to one another.

Denizer, Kaufmann, and Kraay (2013) also consider complexity in their study of how micro (project-level) or macro (country-level) factors are correlated with aid project performance, albeit as a secondary focus. Using three proxies for project complexity (the extent to which a project spans multiple sectors, a project's novelty, and the size of the project), they find "only some evidence that larger—and so possibly more complex—projects are less likely to be successful. On the other hand, greater dispersion of a project across sectors is in fact significantly associated with better project outcomes, and whether a project is a 'repeater' project or not does not seem to matter much for outcomes" (Denizer, Kaufmann, and Kraay 2013, 302).

Given, then, that the issue of accounting for complexity is widespread and often relies upon assessments that are not anchored to an external concept or measure of what complexity is, what are some of the ways that analysts can validate their measures of complexity? Rasul, Rogger, and Williams (2021), in their construction of a measure of the complexity of the tasks being undertaken by Ghanaian civil servants, ensure that coding is undertaken by two independent coders because the variables they measure require coders to make judgment calls about the information reported by government agencies. They also implement reconciliation by managers in cases where there are differences between coders. Discussion between coders and managers about how they see different categories or levels of complexity can be a good way to iron out differences in the measurement of complexity.

Another way to ensure consistency in measuring complexity can be to randomly reinsert particular tasks into the set of tasks being assessed by the coders to check whether they award the same complexity score to identical tasks. This is something that Rasul and Rogger (2018) do in their construction of a measure of task complexity completed by the Nigerian civil service. Rasul and Rogger (2018) also assess the similarity of scores between their two coders and leverage the passing of time to get one of the coders to recode a subsample of projects from scratch (without prompting) to assess the consistency of coding in an additional way.

In a similar spirit, audits of coding can be an effective way to validate a measure of complexity, albeit a costly one. For example, Rasul, Rogger, and Williams (2021, 265) use an auditing technique to check the validity of their measure of task completion; they "matched a subsample of 14% of tasks from progress reports to task audits conducted by external auditors through a separate exercise." Although this technique was applied to task completion, a similar method could easily be used to validate a complexity measure in many contexts; if there are data available on the technical complexity of a task (for example, from engineers or other field specialists), such assessments could be used to check a subsample of the analyst's own evaluations of complexity. Rasul and Rogger (2018), for example, work with a pair of Nigerian engineers to get them to assess the complexity of government tasks according to five dimensions.

Another salient feature of government tasks is how easy it is to define a given task and to evaluate whether and when it has been completed. This feature is related to, but conceptually separate from, the *complexity* of the task. Rasul, Rogger, and Williams (2021) call this feature ex ante and expost *task clarity*. According to their definition, bureaucratic tasks are "ex ante clear when the task can be defined in such a way as to create little uncertainty about what is required to complete the task, and are ex post clear when a report of the actual action undertaken leaves little uncertainty about whether the task was effectively completed" (Rasul, Rogger, and Williams 2021, 260).

Task clarity is an important characteristic to consider, especially in relation to management practices, because the types of management strategy that one wishes to implement may be heavily influenced by the types of task that they govern. Indeed, Rasul, Rogger, and Williams (2021, 260) hypothesize, and find evidence, that "top-down control strategies of incentives and monitoring should be relatively more effective when tasks are easy to define ex ante because it is easier to specify what should be done and construct an appropriate monitoring scheme." On the other hand, they also theorize (and, again, find evidence) that "empowering staff with autonomy and discretion should be relatively more effective when tasks are unclear ex ante, as well as when the actual achievement of the task is clear ex post" (260).

The clarity of task definition is thus also important to take into consideration when exploring questions pertaining to the management of public administration. The degree to which a task is easy to describe and evaluate has a significant bearing on the types of management strategy that make sense to employ when undertaking that task. Task clarity can also impact a number of other features of government work, such as the level of political and citizen support it enjoys—with simpler, more visible projects tending to garner more interest from politicians and support from citizens (Mani and Mukand 2007; Woodhouse 2022)—or the degree to which a task is subject to measurement or performance-pay mechanisms.

Task clarity is important to measure for its potential interactions with the concepts of effort substitution and gaming (Kelman and Friedman 2009). If performance measures are applied only to those tasks that are ex ante and ex post clear, such tasks may be prioritized to the detriment of others because they are subject to measurement or because bureaucrats seek to "game" the system by focusing their attention on improving statistics relating to their performance but not their actual performance. As we have seen in the work of Honig (2019) and Khan (2021), it is especially in complex, multidimensional task environments where granting autonomy or discretion to bureaucrats can have beneficial results. In short, thinking about the nature of the task at hand and its interaction with features such as the management practices being adopted and individual behavioral responses on the part of public servants and politicians is highly important if one wants to get to the bottom of "what works" in government.

DISCUSSION: KEY CHALLENGES

The previous sections have reviewed the scattered and relatively young literature on the systematic measurement of task and project completion in government organizations. The measurement methods and data sources identified hold great promise for practitioners and researchers but also present a number of conceptual and practical challenges. While we have discussed some of these above in relation to specific papers or measurement methods, in this section, we briefly highlight some cross-cutting issues for measurement and analysis as well as for integration into management practice and decision-making.

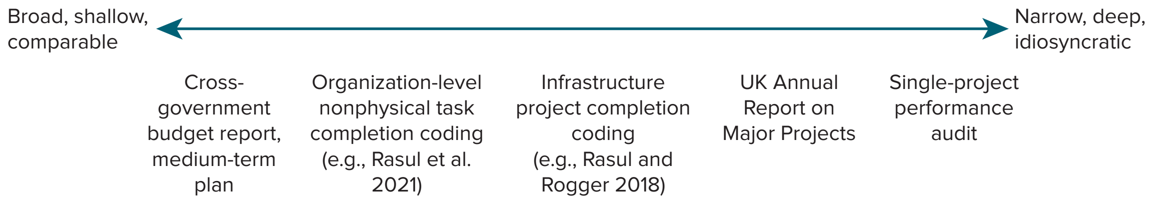
The first challenge is determining what a *task* is. At the beginning of this chapter, we defined outputs as the final products delivered by government organizations to society and tasks as the intermediate steps taken by individuals or teams within government to produce those outputs. We characterized both as discrete, bounded, and clearly linked to each other. While this is conceptually useful and can serve as a guide for measurement, it is also a profound simplification of the messy, interlinked, and uncertain reality of work inside most government organizations. Indeed, the research insights produced by several of the studies we have discussed emphasize that the ambiguity, complexity, and interconnection of tasks and bureaucratic actions often mean that simplistic management efforts do not produce their anticipated effects. Analysts interested in measuring task completion must thus strike a difficult balance between identifying distinct tasks, projects, and outputs in order to measure their completion and simultaneously calibrating their analysis and inference to capture the nuances of the effective performance of these tasks.

A second and related challenge is drawing appropriate inferences from measures of task completion, which, in itself, is just a descriptive fact of the level of task performance. On its own, measuring task completion does not diagnose the causes of task (in)completion, predict future levels of performance, pinpoint needs for improvement, or measure the performance of the individual personnel responsible for a task (since factors outside their control may also matter). It does, however, provide a foundation upon which to conduct further analysis along these lines. Indeed, for most of the studies cited above, the measurement of task completion simply provides a dependent variable for analysis of a diverse range of potential factors and mechanisms. This chapter has focused mainly on the measurement of this dependent variable; linking it to causes and consequences requires additional analysis, which will differ in its aims and methods depending on an analyst's purposes.

A third challenge relates to integrating the measurement of task completion into practice and management—that is, taking action based on it. One main challenge relates to the well-known potential for gaming and distorting effort across multiple tasks (Dixit 2002; Propper and Wilson 2003), exemplified by "Goodhart's Law": "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart 1984). In other words, it may well be possible to accurately measure task completion in government organizations, but using these measures for the purpose of management—particularly if it involves benefits or consequences for the actors involved—risks undermining the validity of the measures and their linkage to bureaucratic performance. See the discussion in chapter 4. While some strategies can be put in place to mitigate such effects (for example, data quality audits or measuring multiple dimensions of bureaucratic performance), these are nearly always imperfect. Analysts should thus seek to innovate in measuring task completion as a means of improving understanding while being cautious and selective in how they use it to guide management actions.

A final consideration in deciding what tasks to measure and how is the trade-off between prioritizing breadth and comparability, on the one hand, and specificity and depth, on the other. Figure 17.2 illustrates this trade-off. In general, task completion measures that are widely applicable across the whole of government will naturally tend to be less specific to (and hence less informative of) the performance of any given unit or task. An example of this might be the type of data contained in a government's annual report, budget execution report, or multiyear plan, which usually cover the whole of government activity but do so at a relatively shallow level. At the other extreme, researchers or practitioners can gather a great deal of information about the completion of a specific task, as a performance audit might do. This gives a very informative picture of the completion of that particular task but permits little comparison across tasks or units. In between, one can locate the various measurement options we have discussed in this chapter. For example, Rasul and Rogger's (2018) project completion data set focuses on physical infrastructure projects, which are likely to be more comparable to each other and across organizations than Rasul, Rogger, and Williams's (2021) data set of both physical and nonphysical outputs—but at the cost of less comprehensive coverage of government activity. The optimal place on this spectrum for any given measure of task completion naturally depends on

FIGURE 17.2 A Spectrum of Task Completion Measures, with Selected Examples


Broad, shallow, comparable
Narrow, deep, idiosyncratic
Cross-government budget report, medium-term plan
Organization-level nonphysical task completion coding (e.g., Rasul et al. 2021)
Infrastructure project completion coding (e.g., Rasul and Rogger 2018)
UK Annual Report on Major Projects
Single-project performance audit

Source: Original figure for this publication.

the analytical purpose for which it is being created. From the standpoint of advancing measurement, the aim is to find ways to surmount this trade-off by increasing both the comparability and the rigor of task completion measures.

CONCLUSION

We conclude by returning to the question with which we opened: how do we know if governments are performing their functions well? In this chapter, we have sought to describe and demonstrate how to apply the task completion framework in order to answer precisely this question. The framework conceives government activity in such a way as to allow analysts to assess public performance in a standardized manner across organizations and types of activity. As such, it gives us a fuller and more accurate picture of government work, forces us to think more carefully about the characteristics of the tasks that different agencies perform, and facilitates comparison of performance on a large sample that spans many types of organizations.

We have applied the framework to different categories of tasks in order to illustrate both its strengths and its limitations. In the case of tasks related to physical outputs, we have shown how data such as engineering assessments, annual progress reports, and budget reports can be merged with other data, such as management or user surveys, to provide a hitherto-inaccessible vision of the extent of project implementation and the quality of the work undertaken.

Much of this work relies, at least partly, on data that already exist but have to be digitized or rendered usable in some other way. The existence of objective, external benchmarks—produced, for example, by experts such as infrastructure engineers—means that the development of projects of many different types can be mapped onto a comparable continuum. The strength of the evaluation of physical outputs is that analysts can produce a meaningful measure of completion that gives the user some sense of how task completion maps onto public benefit. However, the weakness of the approach, as applied to physical outputs, is that the quality of task completion is often overlooked because it rests upon more complex, multifaceted assessments that are difficult to harmonize into a single indicator. Moreover, the reliability of such measures may be called into question where completion rates are reported by the same organizations that undertake the tasks themselves (although this can be counteracted to some degree if external audits of task reports are available to validate the measure).

In the case of nonphysical outputs (such as auditing, planning, or awareness-raising activities), we have demonstrated how data may come from existing sources, such as progress reports, that need to be digitized or processed to be used for analysis. The strength of extending task completion assessments to nonphysical outputs is that this provides a much richer and fuller picture of the activities that governments engage in and allows for meaningful comparisons across departments. However, the task completion framework as applied to nonphysical outputs also suffers from the same potential misreporting concern associated with physical outputs and comes with additional challenges in terms of how to measure the quality of the tasks being completed. The challenges of measuring quality are distinct from those for physical outputs, in that quality is not necessarily overlooked but is more difficult to define. For example, how do you assess the quality of a health strategy objectively and in such a way that it is comparable with, for example, education strategies or fiscal strategies?

The task completion framework, in short, moves us in the right direction when it comes to measuring the performance of governments in a way that takes into account the full breadth of government activity. However, there is much room for improvement when it comes to the measurement of the quality of the provision of both physical and nonphysical outputs. For physical outputs, expert benchmarks are often taken at face value without critical engagement with what the index or evaluation actually captures; whereas, for nonphysical outputs, benchmarks are often nonexistent, with no way to anchor quality assessments that makes them comparable across organizations. This is where we see the frontier in terms of the measurement of government performance; we need to expand the application of the task completion framework and complement this with greater attention to how technical benchmarks are used in the measurement of physical outputs and the development of workable benchmarks for the measurement of nonphysical outputs.

NOTES

The authors gratefully acknowledge funding from the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice. We are grateful to Galileu Kim and Robert Lipinski for helpful comments.

1. See, for instance, the World Bank's World Governance Indicators, available at https://info.wordlbank.org/governance/wgi, and the Millennium Challenge Corporation scorecards—for example, on the website of the Millennium Challenge Coordinating Unit for Sierra Leone, http://www.mccu-sl.gov.sl/scorecards.html.
2. *Outputs* are not to be confused with *outcomes*, or "the impacts on social, economic, or other indicators arising from the delivery of outputs (e.g., student learning, social equity)." *OECD Glossary of Statistical Terms*, s.vv. "output," "outcome" (Paris: OECD Publishing, 2022), http://stats.oecd.org/glossary.
3. Such indicators do not rely upon subjective citizen-survey responses, which are limited by their reliance on human judgment and prey to multiple biases and recall issues (Golden 1992), both from the researcher designing the study and the experts or citizen respondents evaluating the government.
4. No evidence was found that completion levels differed significantly across auditors and agencies, with 94 percent of completion rates being corroborated across coding groups (Rasul, Rogger, and Williams 2021, 265).
5. The measure of unfinished projects is a "combination of projects still underway (on time or delays) and abandoned (temporarily or indefinitely) in a given district" (Bancalari 2022, 10).
6. It is important to note that their findings are relative to one another—that is, "organizations appear to be overbalancing their management practice portfolios toward top-down control measures at the expense of entrusting and empowering the professionalism of their staff" (Rasul, Rogger, and Williams 2021, 261).

REFERENCES

Ammons, David N. 2014. *Municipal Benchmarks: Assessing Local Performance and Establishing Community Standards*. 3rd ed. London: Routledge.

Andrews, Rhys, George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole Jr., and Richard M. Walker. 2005. "Representative Bureaucracy, Organizational Strategy, and Public Service Performance: An Empirical Analysis of English Local Government." *Journal of Public Administration Research and Theory* 15 (4): 489–504. https://doi.org/10.1093/jopart/mui032.

Bancalari, Antonella. 2022. "Can White Elephants Kill? Unintended Consequences of Infrastructure Development in Peru." IFS Working Paper 202227, Institute for Fiscal Studies, London. https://ifs.org.uk/publications/can-white-elephants-kill-unintended-consequences-infrastructure-development.

Bedoya, Guadalupe, Jishnu Das, and Amy Dolinger. Forthcoming. "Randomized Regulation: The Impact of Minimum Quality Standards on Health Markets." Working paper, World Bank, Washington, DC.

Bertelli, Anthony Michael, Eleanor Florence Woodhouse, Michele Castiglioni, and Paolo Belardinelli. 2021. *Partnership Communities.* Cambridge Elements: Public and Nonprofit Administration. Cambridge: Cambridge University Press.

Bertelli, Anthony Michael, Valentina Mele, and Andrew B. Whitford. 2020. "When New Public Management Fails: Infrastructure Public-Private Partnerships and Political Constraints in Developing and Transitional Economies." *Governance: An International Journal of Policy, Administration, and Institutions* 33 (3): 477–93. https://doi.org/10.1111/gove.12428.

Boyne, George A. 2003. "What Is Public Service Improvement?" *Public Administration* 81 (2): 211–27. https://doi.org/10.1111/1467-9299.00343.

Brown, Karin, and Philip B. Coulter. 1983. "Subjective and Objective Measures of Police Service Delivery." *Public Administration Review* 43 (1): 50–58. https://doi.org/10.2307/975299.

Carter, Neil, Rudolf Klein, and Patricia Day. 1992. *How Organisations Measure Success: The Use of Performance Indicators in Government*. London: Routledge.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay. 2013. "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance." *Journal of Development Economics* 105: 288–302. https://doi.org/10.1016/j.jdeveco.2013.06.003.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *Journal of Human Resources* 37 (4): 696–727. https://doi.org/10.2307/3069614.

Engel, Eduardo, Ronald Fischer, and Alexander Galetovic. 2013. "The Basic Public Finance of Public-Private Partnerships." *Journal of the European Economic Association* 11 (1): 83–111. https://www.jstor.org/stable/23355049.

Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. https://doi.org/10.3982/ECTA19 244.

Finer, Herman. 1941. "Administrative Responsibility in Democratic Government." *Public Administration Review* 1 (4): 335–50. https://doi.org/10.2307/972907.

Flyvbjerg, Bent. 2009. "Survival of the Unfittest: Why the Worst Infrastructure Gets Built—And What We Can Do about It." *Oxford Review of Economic Policy* 25 (3): 344–67. https://doi.org/10.1093/oxrep/grp024.

Flyvbjerg, Bent, Mette Skamris Holm, and Soren Buhl. 2002. "Underestimating Costs in Public Works Projects: Error or Lie?" *Journal of the American Planning Association* 68 (3): 279–95. https://doi.org/10.1080/01944360208976273.

Friedrich, Carl J. 1940. "Public Policy and the Nature of Administrative Responsibility." In *Public Policy: A Yearbook of the Graduate School of Public Administration, Harvard University* 1: 1–20.

Golden, Brian R. 1992. "The Past Is the Past—Or Is It? The Use of Retrospective Accounts as Indicators of Past Strategy." *Academy of Management Journal* 35 (4): 848–60. https://doi.org/10.2307/256318.

Goodhart, Charles A. E. 1984. "Problems of Monetary Management: The UK Experience." In *Monetary Theory and Practice: The UK Experience*, 91–121. London: Red Globe Press. https://doi.org/10.1007/978-1-349-17295-5.

Hefetz, Amir, and Mildred E. Warner. 2012. "Contracting or Public Delivery? The Importance of Service, Market, and Management Characteristics." *Journal of Public Administration Research and Theory* 22 (2): 289–317. https://doi.org/10.1093/jopart/mur006.

Ho, Alfred Tat-Kei, and Wonhyuk Cho. 2017. "Government Communication Effectiveness and Satisfaction with Police Performance: A Large-Scale Survey Study." *Public Administration Review* 77 (2): 228–39. https://doi.org/10.1111/puar.12563.

Honig, Dan. 2019. "When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation." *International Organization* 73 (1): 171–201. https://doi.org/10.1017/S002081831800036X.

IPA (Infrastructure and Projects Authority). 2022. *Annual Report on Major Projects 2021–22.* Reporting to Cabinet Office and HM Treasury, United Kingdom Government. London: IPA. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachmentdata/file/1092181/IPAAR2022.pdf.

Kaddu, M., J. Aguilera, and L. Carson. n.d. *Challenges to Policy Implementation in Uganda (Review of Policy Implementation in Uganda)*. London: International Growth Centre, London School of Economics and Political Science.

Kelman, Steven, and John N. Friedman. 2009. "Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service." *Journal of Public Administration Research and Theory* 19 (4): 917–46. https://doi.org/10.1093/jopart/mun028.

Khan, Muhammad Yasir. 2021. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." Working paper delivered at 25th Annual Conference of the Society for Institutional Organizational Economics, June 24–26, 2021 (accessed February 8, 2023). https://y-khan.github.io/yasirkhan.org/muhammadyasirkhanjmp.pdf.

Khwaja, Asim Ijaz. 2009. "Can Good Projects Succeed in Bad Communities?" *Journal of Public Economics* 93 (7–8): 899–916. https://doi.org/10.1016/j.jpubeco.2009.02.010.

Lee, Soo-Young, and Andrew B. Whitford. 2009. "Government Effectiveness in Comparative Perspective." *Journal of Comparative Policy Analysis* 11 (2): 249–81. https://doi.org/10.1080/13876980902888111.

Lewis, David E. 2007. "Testing Pendleton's Premise: Do Political Appointees Make Worse Bureaucrats?" *The Journal of Politics* 69 (4): 1073–88. https://doi.org/10.1111/j.1468-2508.2007.00608.x.

Lu, Jiahuan. 2016. "The Performance of Performance-Based Contracting in Human Services: A Quasi-Experiment." *Journal of Public Administration Research and Theory* 26 (2): 277–93. https://doi.org/10.1093/jopart/muv002.

Mani, Anandi, and Sharun Mukand. 2007. "Democracy, Visibility and Public Good Provision." *Journal of Development Economics* 83 (2): 506–29. https://doi.org/10.1016/j.jdeveco.2005.06.008.

Mansoor, Zahra, Garance Genicot, and Ghazala Mansuri. 2021. "Rules versus Discretion: Experimental Evidence on Incentives for Agriculture Extension Staff." Unpublished manuscript.

Nistotskaya, Marina, and Luciana Cingolani. 2016. "Bureaucratic Structure, Regulatory Quality, and Entrepreneurship in a Comparative Perspective: Cross-Sectional and Panel Data Evidence." *Journal of Public Administration Research and Theory* 26 (3): 519–34. https://doi.org/10.1093/jopart/muv026.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49. https://doi.org/10.1086/517935.

Poister, Theodore H., and Gregory Streib. 1999. "Performance Measurement in Municipal Government: Assessing the State of the Practice." *Public Administration Review* 59 (4): 325–35. https://doi.org/10.2307/3110115.

Post, Alison E. 2014. *Foreign and Domestic Investment in Argentina: The Politics of Privatized Infrastructure.* Cambridge, UK: Cambridge University Press.

Prendergast, Canice. 2002. "The Tenuous Trade-Off between Risk and Incentives." *Journal of Political Economy* 110 (5): 1071–102. https://doi.org/10.1086/341874.

Propper, Carol, and Deborah Wilson. 2003. "The Use and Usefulness of Performance Measures in the Public Sector." *Oxford Review of Economic Policy* 19 (2): 250–67. https://doi.org/10.1093/oxrep/19.2.250.

Rainey, Hal G. 2009. *Understanding and Managing Public Organizations.* 4th ed. New York: John Wiley & Sons.

Rainey, Hal G., and Paula Steinbauer. 1999. "Galloping Elephants: Developing Elements of a Theory of Effective Government Organizations." *Journal of Public Administration Research and Theory* 9 (1): 1–32. https://doi.org/10.1093/oxfordjournals.jpart.a024401.

Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal* 128 (608): 413–46. https://doi.org/10.1111/ecoj.12418.

Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2): 259–77. https://doi.org/10.1093/jopart/muaa034.

Rauch, James E., and Peter B. Evans. 2000. "Bureaucratic Structure and Bureaucratic Performance in Less Developed Countries." *Journal of Public Economics* 75 (1): 49–71. https://doi.org/10.1016/S0047-2727(99)00044-4.

Remington, Kaye, and Julien Pollack. 2007. *Tools for Complex Projects.* Aldershot, UK: Gower.

Talbot, Colin. 2010. *Theories of Performance: Organizational and Service Improvement in the Public Domain.* Oxford: Oxford University Press.

Thomas, John Clayton, Theodore H. Poister, and Nevbahar Ertas. 2009. "Customer, Partner, Principal: Local Government Perspectives on State Agency Performance in Georgia." *Journal of Public Administration Research and Theory* 20 (4): 779–99. https://doi.org/10.1093/jopart/mup024.

Walker, Richard M., M. Jin Lee, Oliver James, and Samuel M. Y. Ho. 2018. "Analyzing the Complexity of Performance Information Use: Experiments with Stakeholders to Disaggregate Dimensions of Performance, Data Sources, and Data Types." *Public Administration Review* 78 (6): 852–63. https://doi.org/10.1111/puar.12920.

Williams, Martin J. 2017. "The Political Economy of Unfinished Development Projects: Corruption, Clientelism, or Collective Choice?" *American Political Science Review* 111 (4): 705–23. https://doi.org/10.1017/S0003055417000351.

Woodhouse, Eleanor Florence. 2022. "The Distributive Politics of Privately Financed Infrastructure Agreements." Unpublished manuscript.