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.

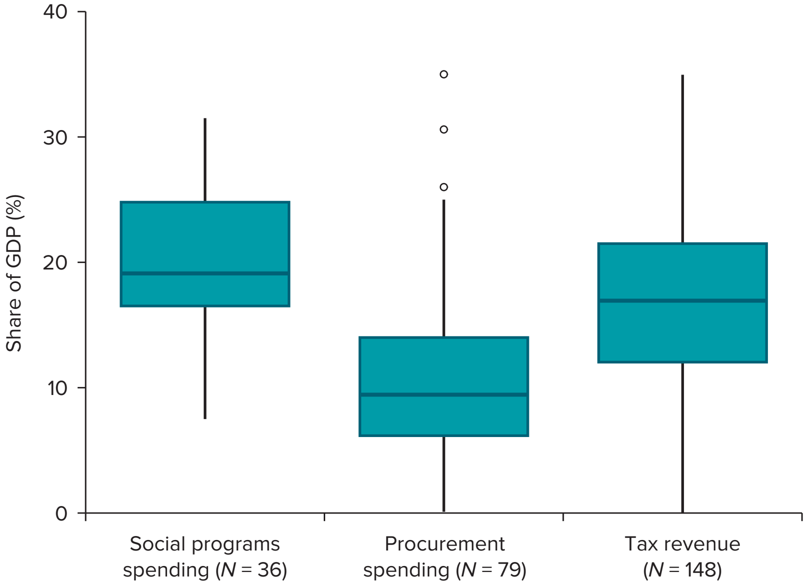
A large share of government operations involve the processing of case files or cases. Case data are the records of responses by public officials to requests for public services or the fulfillment of public responsibilities. A case file is typically a collection of records regarding an application. The 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.

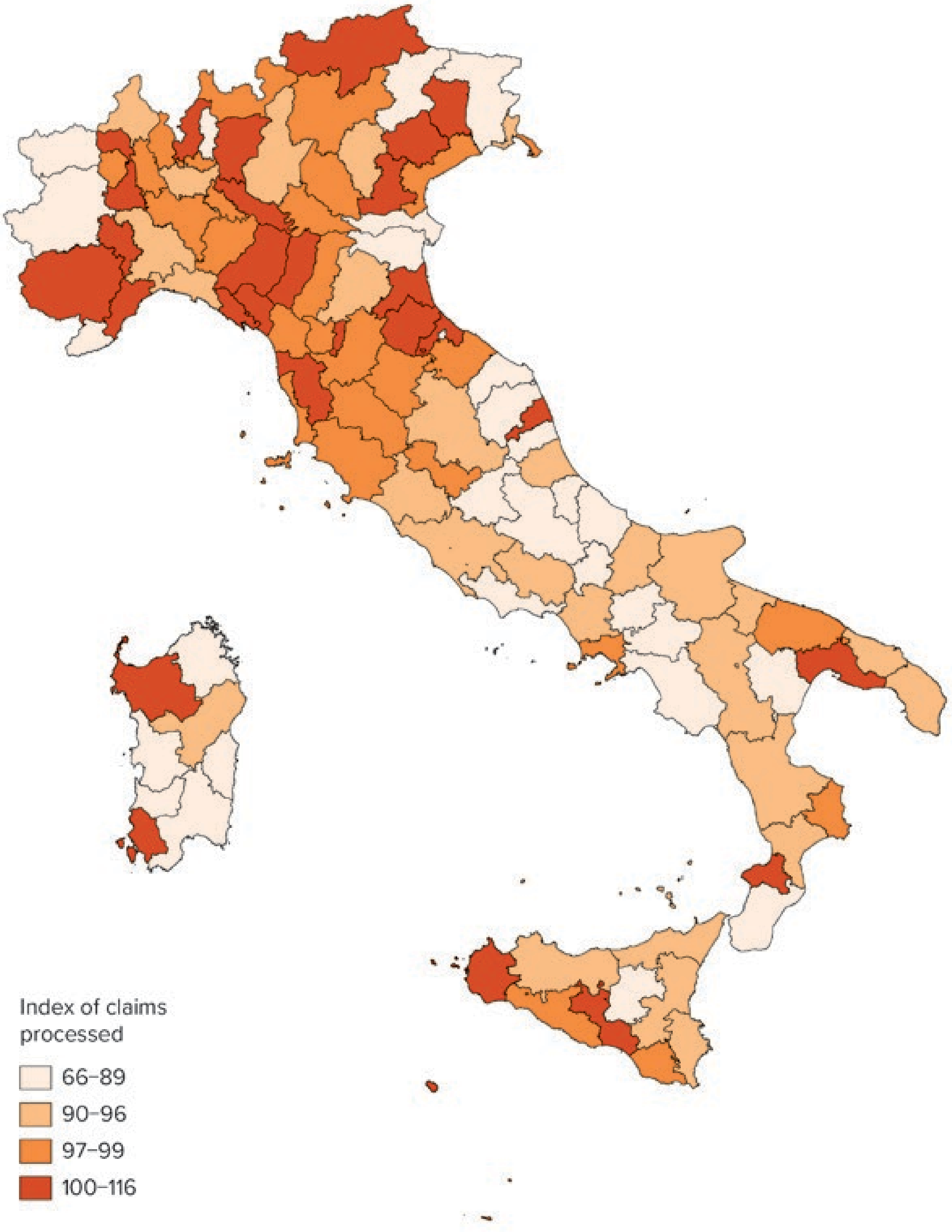
A naive solution might involve counting the number of claims processed by each office. Despite being simple and transparent, this measure suffers from a major draw-back: it does not take into account task complexity. Some claims might be very quick to process, while others might require a lot of time 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.

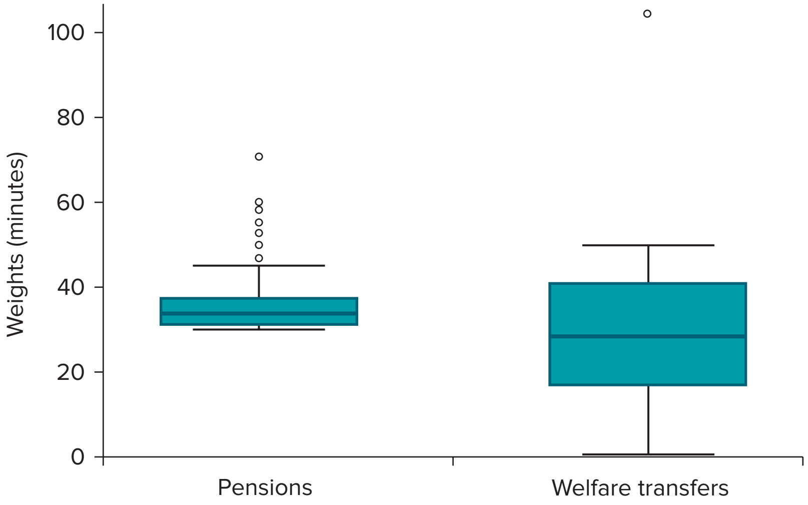
Importantly, the ISSA complexity-adjustment formula uses objective weights as opposed to subjective scores. As part of the ISSA quality control department, there is a team devoted to measuring weights and keeping them up-to-date. To construct the weight for product *v*, this team selects an excellent, 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:

$$\text{output}_i = \sum_{\upsilon=1}^{V} c_{vi} \times w_\upsilon \qquad (15.1)$$

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):

$$y_i = \tfrac{1}{6}\sum_{k=1}^{6}(y_i^k - \overline{y}^k)/\sigma^k. \tag{15.2}$$

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.

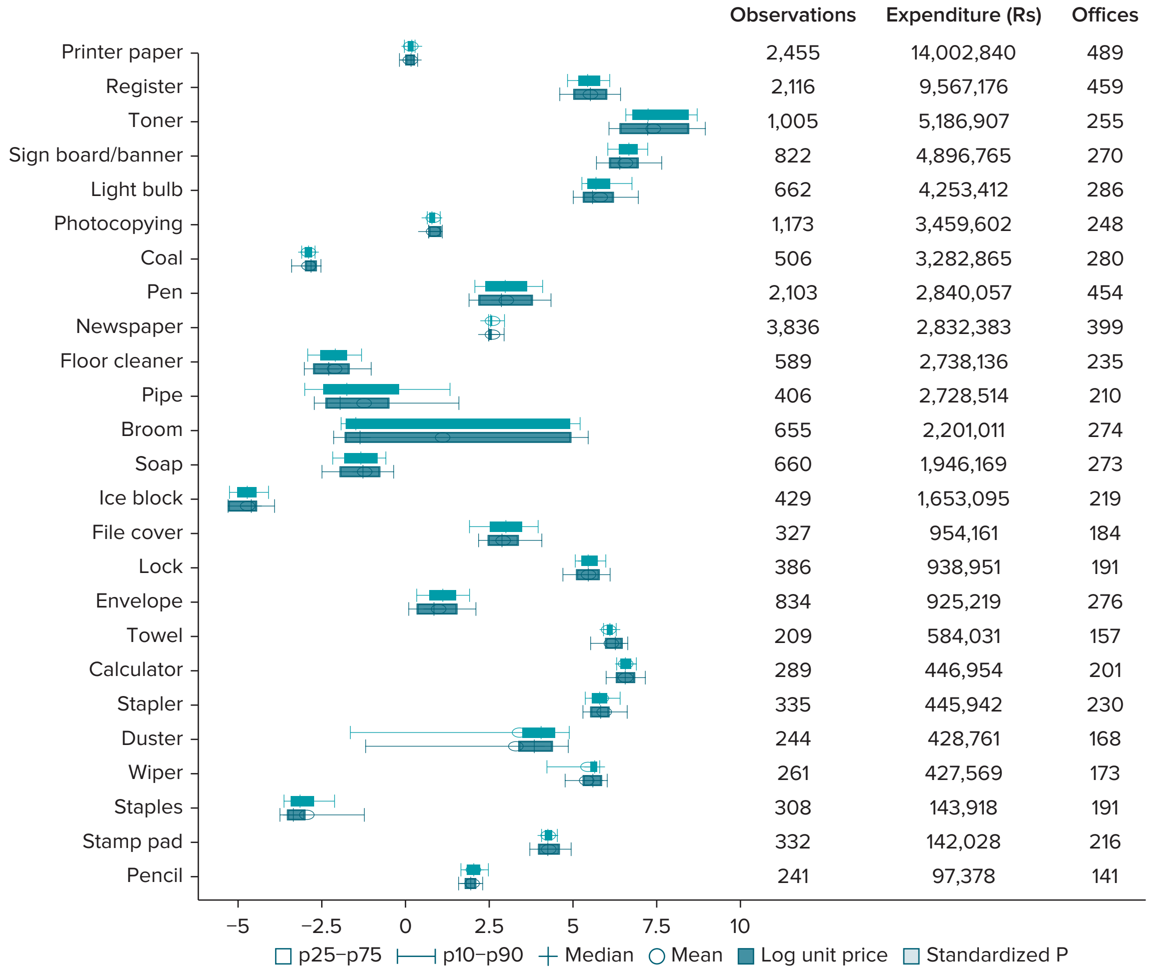
The government of Punjab considers the primary purpose of public procurement to be ensuring that "the object of procurement brings value for money to the procuring agency" (PPRA 2014). In line with this, they developed a measure of bureaucratic performance that seeks to measure value for money in the 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 $v_{igto} = \sum_{j\in A(g)} \hat{\lambda}_j X_j$, 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.