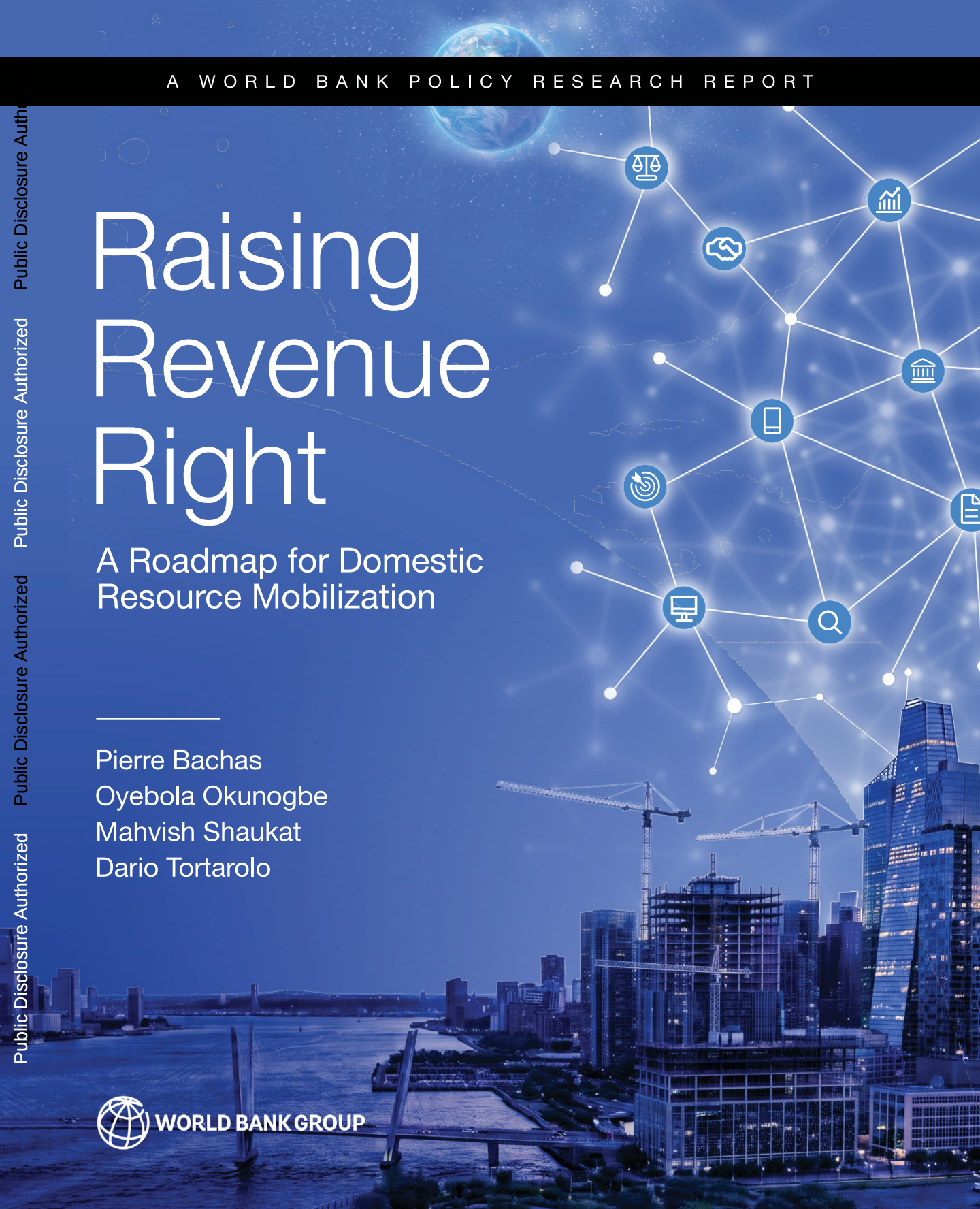
Public Disclosure Authorized
Public Disclosure Authorized
Public Disclosure Authorized
Public Disclosure Authorized

A WORLD BANK POLICY RESEARCH REPORT

Raising Revenue Right

A Roadmap for Domestic Resource Mobilization

Pierre Bachas
Oyebola Okunogbe
Mahvish Shaukat
Dario Tortarolo



WORLD BANK GROUP

CHAPTER 8

Administrative Tax Data for Policy

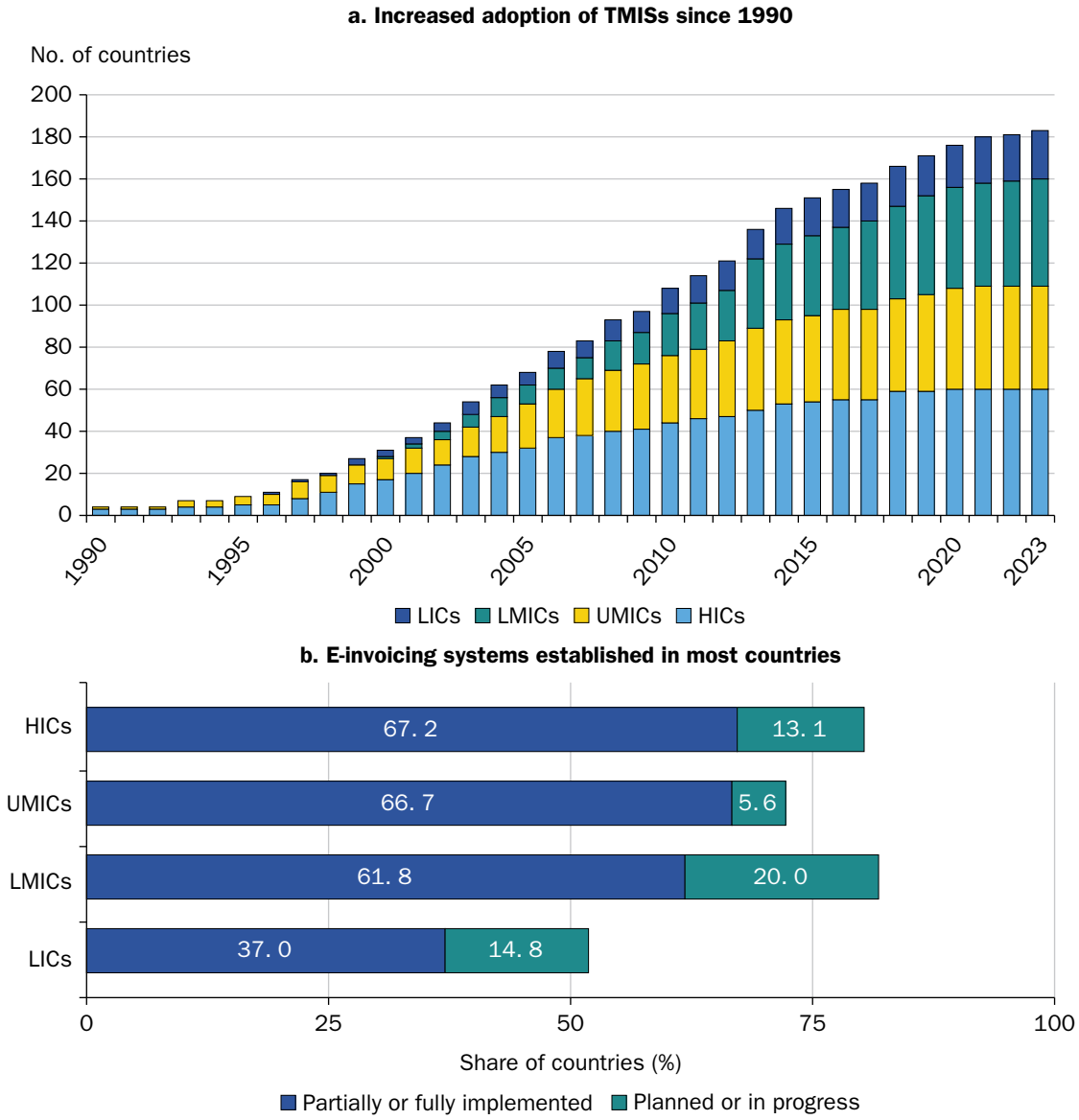
Prepared by Pierre Bachas, Gaston Bernheim, and Thiago Scot

Introduction

A digital tax (r)evolution. The digital age is reshaping how individuals, firms, and governments collect, analyze, and share information: Tax authorities are central to this data revolution. With the right investments in digital infrastructure, they can now collect vast amounts of high-quality, high-frequency data required to manage their tax and transfer systems. Integrated digital information systems, otherwise known as *tax management information systems* (TMISs), which automate and standardize processes such as filing and processing tax declarations, make such collection possible. Two decades ago, only high-income countries (HICs) implemented TMISs. Since then, the number of adopters has increased sixfold, and the practice has become nearly universal (refer to figure 8.1, panel a). On the advanced end of the TMIS technologies spectrum is electronic invoicing (e-invoicing), which transmits transaction data to tax administrations in real time. A majority of countries are implementing—or planning to implement—e-invoicing systems, with many lower-income and upper-middle-income countries (UMICs) leapfrogging their high-income counterparts (refer to figure 8.1, panel b).

The unique advantages of tax data. Compared to survey data, administrative tax data have several advantages. First, they record the economic activity of all registered firms and individuals, resulting in a much larger number of observations. They also capture outcomes beyond the scope of even the best surveys on, for example, firms' trade partners or employees' wages. Second, administrative tax data are collected from sources such as payroll and e-invoicing almost in real time. Third, administrative data collection is recurring, so taxpayers' activity can be followed over time. Fourth, taxpayer identifiers enable linking multiple datasets within the tax administration

Figure 8.1 Digital expansion in tax administrations, by income group



Sources: Original figure for this publication based on data from the World Bank's Group GovTech Dataset and the VatCalc E-Invoicing Global Tracker.

Note: The sample covers 183 countries in panel a (60 HICs, 49 UMICs, 51 LMICs, and 23 LICs) and 197 countries in panel b (61 HICs, 54 UMICs, 55 LMICs, and 27 LICs). e-invoicing = electronic invoicing; HICs = high-income countries; LICs = low-income countries; LMICs = lower-middle-income countries; TMISs = tax management information systems; UMICs = upper-middle-income countries.

and across other government bodies (for example, social security and procurement). Together, these features make tax data well-suited to policy evaluation and cutting-edge research (Bachas and Jensen 2023; Brockmeyer 2019; Card et al. 2010).

Putting data to use. Although many governments have an abundance of data, they face challenges in translating this information into timely, actionable insights. These challenges stem from limited institutional capacity, weak incentives, and inadequate data integration (Wani 2025). This chapter's objective is to showcase how tax data have the potential to inform public policy across a range of economic domains and to demonstrate how to overcome challenges in using tax data. This chapter also aims to guide policy makers, public officials, and researchers who are either new to or already working with administrative microdata. By the end, readers should have a solid grasp of how tax data can uniquely inform policy; how to build flexible, mutually beneficial partnerships with tax administrations; which analytical tools to use; and how to mitigate privacy and security risks.

Outline. This chapter presents practical insights and real-world examples to help readers with relevant use cases. By linking administrative datasets, one can answer questions such as, How do shocks propagate through supply chains? How do the wealthiest individuals earn their income? How often do workers change jobs? (refer to the section "Innovative Data to Learn About the Economy"). In addition to showcasing innovative examples, this chapter promotes collaboration among tax administrations, researchers, and civil society organizations by outlining partnership models that can be tailored to contexts such as data exchange, data labs, and remote workflows (refer to the section "Building Collaborations: Practical Models for Every Need"). This chapter then introduces analytical tools to work with large-scale administrative data, including descriptive methods, econometric techniques, and newer approaches that use artificial intelligence and machine learning (refer to the section "Analytical Tools for Big Administrative Data"). The chapter continues by discussing limitations and risks of administrative tax data use, such as gaps in coverage and privacy (refer to the section "Lurking in the Shadows: What Are We Missing?"). The chapter concludes with recommendations depending on a country's starting point with administrative data (refer to the section "From Tax Records to Data-driven Policy Making: A Roadmap for Strengthening Tax Data Use").

Link to the report. This final chapter builds on key themes presented in earlier chapters, especially technology (refer to chapter 3) and transparency (refer to chapter 4). Chapter 3 discusses how technological innovations are transforming tax collection by generating richer data. Chapter 4 describes global trends in transparency that provide tax administrations with information on top economic actors and offshore economic activity. This chapter builds on these foundations to present a practical approach to turning tax data into policy insights by illustrating concrete ways

tax administrations and researchers can harness this information to support evidence-based decision-making. On its own, this chapter serves as a reference regarding administrative tax data, identifying innovative uses, flexible collaboration models, and analytical tools.

Innovative data to learn about the economy

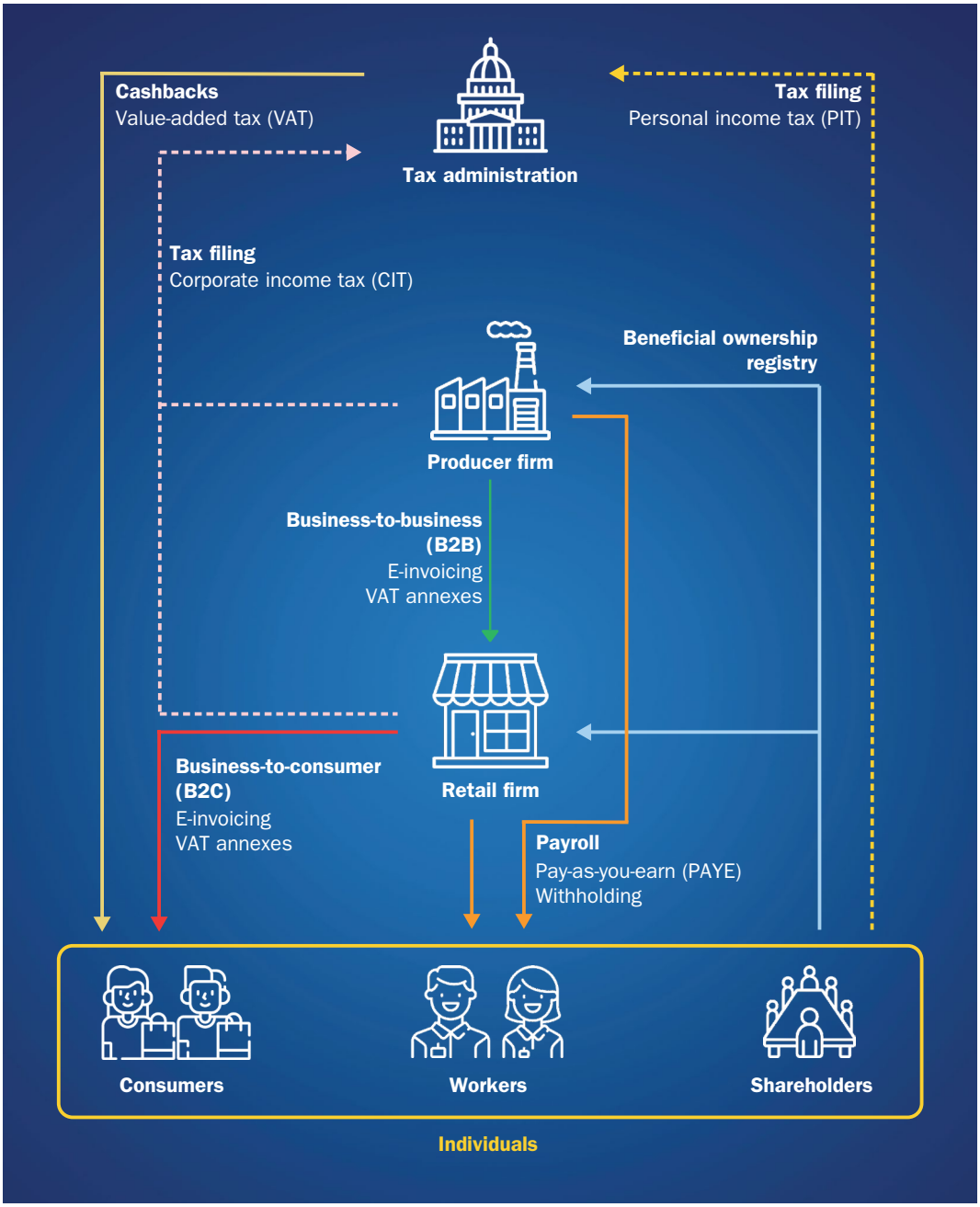
Administrative tax data. *Administrative tax data* refer to the various datasets that tax administrations generate by managing tax and transfer systems. These datasets can be classified as taxpayer registries, self-reported information, third-party reports, internal administrative records (for example, audit results), and data shared by other government entities (for example, customs and procurement). Tax administrations differ in how they integrate these different data types, due to differences in reporting requirements, technology adoption and digitization, and their collaboration with other government bodies (refer to the section "Building Collaborations: Practical Models for Every Need").

Data generation. One of the key functions of tax administrations is to tackle information asymmetries that arise from firms' and individuals' private knowledge of their own activity. As economic actors interact with each other and with the state, they signal their activity. For example, firms report both their own and their stakeholders' activities, including wages paid to employees, sales to clients, purchases from suppliers, and dividends to shareholders. Reporting is shaped by incentives that may distort the data (refer to the section "Lurking in the Shadows: What Are We Missing?"), yet tax data are increasingly valuable and often the only cost-effective and timely data source on economic activity.

Data linkages hold the key. Administrative data shine when multiple datasets can be linked together via taxpayer identifiers. Figure 8.2 displays the key actors and flows of information in the economy. At the top is tax administration, which mandates information reporting, receives taxes, and sends tax rebates. In the middle are the firms, such as retailers and upstream suppliers (producers), involved at different points of the supply chains. At the bottom are individuals, who play different economic roles: consumer, worker, or a shareholder of a firm. In this flow, there are four data links among employers and employees (in orange); producers and retailers (in green); firms and their owners (in blue); and retailers and individual consumers (in red) that provide critical insights about the economy.

For each of the four key links presented in figure 8.2, the following sections detail innovative use cases from collaborations between tax administrations and researchers. We order the topics from the least to most challenging based on each data type's difficulty and accessibility.

Figure 8.2 Administrative tax data generation process



Source: Original figure for this publication. Icons from Flaticon and licensed by CC-BY 3.0.

Note: E-invoicing = electronic invoicing.

Labor market outcomes. Because employers report wages to the tax administration, tax records can be used to observe labor market relationships. For instance, when workers change employers, tax records allow the tracking of their subsequent employment, enabling the study of job mobility and labor market dynamics in the formal sector. Labor force surveys are the only substitutes that follow a set of workers over the course of a few quarters (Donovan et al. 2023). However, these surveys usually include only limited information on employers. In contrast, tax data provide a much richer view of labor market demand. In this regard, administrative employer–employee matched datasets have become a staple of labor economics research in HICs and are increasingly being used elsewhere, given their longer-term panel dimensions. The section "Labor Market Transitions: Insights from Employer–Employee Linked Data" provides examples of recent studies that use such data to study job-switching dynamics and entrepreneurship in developing countries.

Supply chains and trading networks. The second link concerns firm-to-firm (also called business-to-business [B2B]) transactions). The recent expansion of electronic invoices has greatly facilitated the visibility of B2B transactions, as invoices often include identifiers for both partners. In most countries, data on B2B transactions are also available from value-added tax (VAT) forms, because firms are often mandated to report transactions with their suppliers and buyers to cross-validate claims. No other data source provides this information on intranational trade between firms. The section "Trade Networks Revealed: The Power of Firm-to-Firm Transaction Data" illustrates how these data are used to track domestic trade flows, to analyze how trade responds to weather shocks, and to examine discrepancies in how buyers and sellers report the same transaction.

Firm ownership. The third link concerns firms and their owners. Shareholder and beneficial ownership registries, often managed by tax administrations, make it possible to link firms to their owners (refer to chapter 4). The only substitute for such data are listings of publicly traded firms, of which few exist in lower-middle-income countries (LMICs), and some private databases which aggregate online sources (such as Orbis). Ownership registries are a new trove of valuable data with multiple applications. The section "The View from the Top: Linking Tax and Ownership Data" shows how these registries include business income and, thereby, provide a comprehensive view of wealthy individuals' income sources.

Prices and individual consumption patterns. The fourth data structure comes from e-invoicing, which records transactions in the formal sector, including prices, quantities, and products sold. Many countries implement programs that encourage consumers to request proof of purchase, typically by granting partial VAT rebates or awarding consumers lottery prizes for their transactions. To participate in these programs, consumers must include their national IDs at the

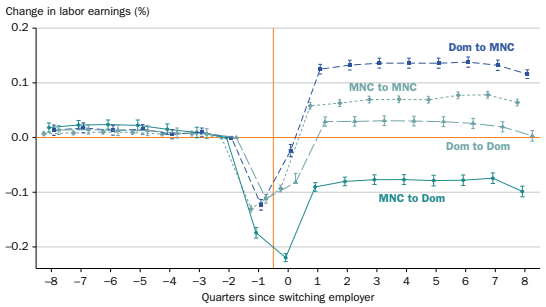
time of purchase. Connecting consumers to their purchases generates rich data on individuals' formal consumption patterns. Electronic invoices also offer high-quality, high-frequency data on prices, enabling real-time inflation measurement and a deeper analysis of the prices governments pay in public procurement. Building on these examples, the section "E-invoicing: Toward Granular Insights" presents frontier applications using e-invoicing data.

Labor market transitions: Insights from employer–employee linked data

The studies that follow illustrate how employer–employee linked data can be used to answer policy-relevant questions on labor mobility, firm dynamics, and entrepreneurship that are difficult to address with traditional surveys.

The Case of Costa Rica: How Do Multinational Corporations Impact Local Labor Markets?

Figure 8.3 Wage gains in Costa Rica, by type of move, 2006–17



Source: Adapted from figure 1, panel b, from Alfaro-Urena et al. (2021).

Note: Dom = domestic firm; MNC = multinational corporation.

Question: How do multinational corporations (MNCs) impact local workers' wages?

Data: Matched employer–employee data and corporate tax filings.

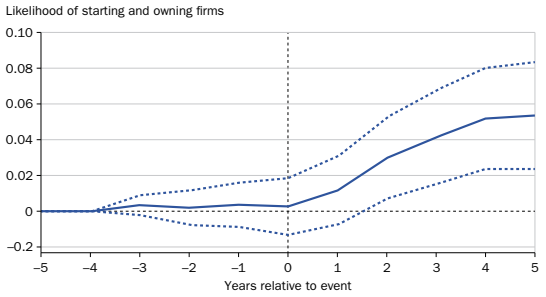
Findings:

- Workers taking a job at MNCs earn 9 percent more than those switching jobs between domestic firms.
- MNC expansion raises wages in local firms by improving outside options for workers (labor market exposure) and boosting performance of local suppliers (firm-level linkages).

Policy implications: Attracting MNCs has the potential to increase wages economywide, through direct and indirect channels (refer to figure 8.3).

The Case of the Dominican Republic: Does Managerial Experience Foster Entrepreneurship?

Figure 8.4 Firm ownership in the Dominican Republic, 2010–18



Source: Figure used with permission from Pecenco et al. (2026).

Question: How does managerial experience impact business creation?

Data: Matched employer–employee data, public procurement lottery records, and firm balance sheets.

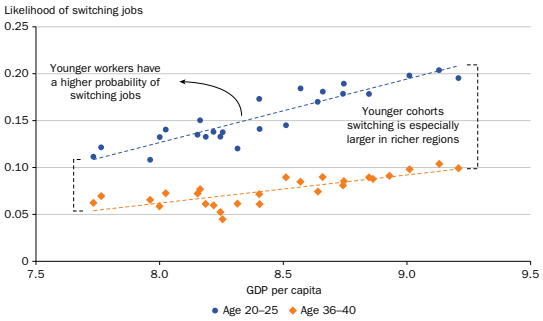
Findings:

- Access to temporary managerial roles boosts entrepreneurship, increasing the likelihood of starting a firm by 22 percent.
- This effect is driven by the gain in managerial experience and self-selection of high-potential individuals into these roles.

Policy implications: Managerial opportunities via public contracting can catalyze private business creation (refer to figure 8.4).

The Case of Brazil: How Does Worker Mobility Relate to Economic Development?

Figure 8.5 Likelihood of switching jobs in Brazil, by age group, 1994–2017



Source: Adapted from figure 4, panels c and d, from Brockmeyer et al. (2026).

Note: GDP = gross domestic product.

Question: How do formal employment patterns evolve with the income level of an economy?

Data: Matched employer–employee data across 8 countries and across more than 26 regions within 2 large countries (Brazil and Colombia).

Findings:

- Job transitions within the formal sector are more common in richer economies and correlate with higher lifecycle wage growth.
- Young workers are more mobile everywhere, particularly in richer economies.

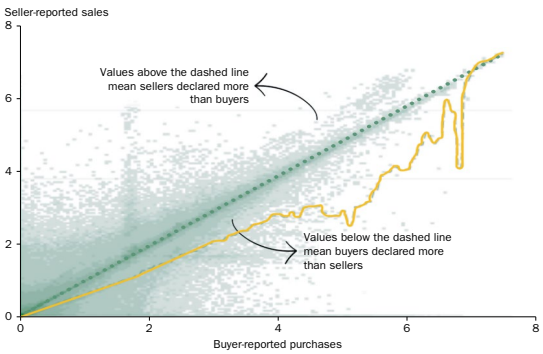
Policy implications: Enhancing labor mobility can lead to long-term wage gains (refer to figure 8.5).

Trade networks revealed: The power of firm-to-firm transaction data

The examples that follow illustrate how firm-to-firm transaction data can uncover the structure of domestic supply chains, firms' responses to shocks, and reporting frictions within the tax system.

The Case of Uganda: Are Firms Sophisticated Tax Evaders or Confused by the Tax System?

Figure 8.6 VAT amounts declared by sellers and buyers in Uganda, 2013–16



Source: Figure used with permission from Almunia et al. (2024).

Note: VAT = value-added tax.

Question: Do firms misreport VAT liabilities to strategically evade taxes, or do they make reporting mistakes?

Data: Firm-to-firm VAT transaction-level records.

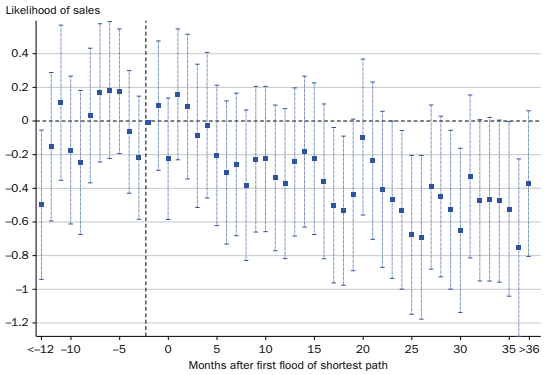
Findings: To pay less taxes, buyers want to overreport purchases (input VAT) and sellers want to underreport sales (output VAT).

- 75 percent of firms report in ways that lower their tax liability, but 25 percent misreport in ways that actually increase their tax liability.

Policy implications: Eliminating discrepancies between seller and buyer reports would increase total VAT collected by 28 percent. Yet, many firms are just confused by reporting requirements (refer to figure 8.6).

The Case of Pakistan: How Do Commercial Relationships Adapt to Extreme Climate Events?

Figure 8.7 Impact of road flooding on sales in Pakistan, 2011–18



Source: Figure used with permission from Balboni et al. (Forthcoming).

Question: How do firms adapt in response to weather shocks?

Data: Georeferenced VAT firm-to-firm monthly transactions and satellite-derived data on major flood events and flood risk.

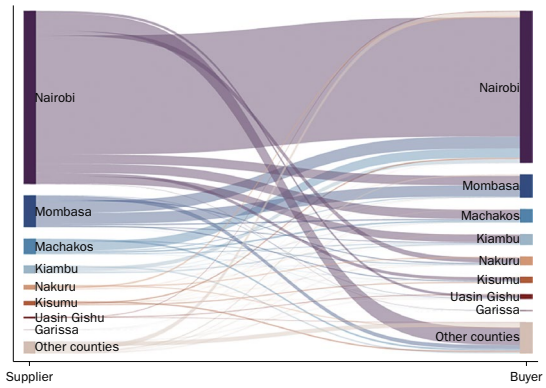
Findings:

- Firms adapt to flood events by relocating to safer areas and shifting input sourcing toward less flood-prone suppliers.
- These responses are persistent and consistent with firms learning about underlying climate risk.

Policy implications: Private sector adaptation is a significant factor mediating the impacts of climate events. Providing accurate information on risks would speed adaptation (refer to figure 8.7).

The Case of Kenya: Are Trade Flows in the Formal Sector Representative of the Full Economy?

Figure 8.8 County-level trade flows between formal firms in Kenya, 2015–19



Source: Figure used with permission from Wiedemann et al. (2024).

Question: Supply chains in developing countries combine formal and informal firms. What is missing when using transaction data only from formal firms?

Data: Firm-to-firm VAT transaction records; census and surveys recording activity in the informal sector.

Findings: Ignoring the informal sector implies that

- Trade seems more concentrated in large cities than it truly is and
- Exposure of domestic firms to international trade shocks is overestimated.

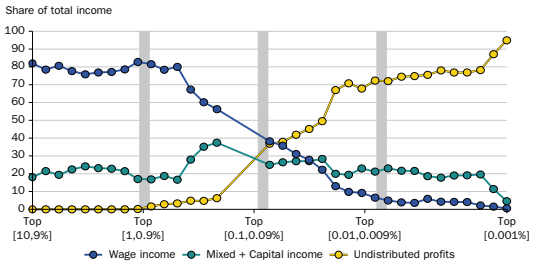
Policy implications: In the presence of a large informal sector, complementing tax records with additional data is key to correctly assessing macroeconomic impacts of policies (refer to figure 8.8).

The view from the top: Linking tax and ownership data

The studies that follow demonstrate how linking tax records with ownership information provides a more complete picture of top incomes, wealth concentration, and the distribution of the tax burden.

The Case of Honduras: From Wages to Profits—Tracking Changes in the Composition of Top Incomes

Figure 8.9 Composition of top incomes in Honduras, 2011–19



Source: Adapted from figure 3, panel a, from Del Carmen et al. (2025).

Question: What are the main income sources of high earners?

Data: Personal income tax (PIT) and corporate income tax (CIT) records, connected through a corporate shareholder registry.

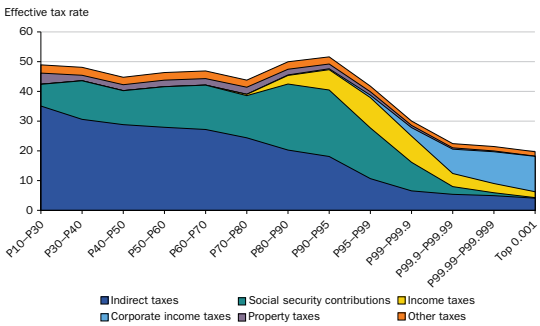
Findings:

- Capital income, mostly undistributed corporate profits, accounts for 80 percent of the income of the richest 0.01 percent of individuals.
- Honduras exhibits high and persistent inequality, with the top 1 percent receiving around 30 percent of total pre-tax income.

Policy implications: Taking a comprehensive view of income, including business income—profits retained in firms—is essential to understanding the income level and composition of top earners (refer to figure 8.9).

The Case of Brazil: Who Bears the Tax Burden? A Look across the Income Distribution

Figure 8.10 Effective tax rate across the income distribution in Brazil, 2019



Source: Adapted from figure 10 from Palomo et al. (2025).

Question: How equitable is the tax system when we consider all taxes paid, including taxes from capital income of firm owners?

Data: Firm ownership registry linking CIT records to their owners, PIT records, household surveys, and national accounts aggregates.

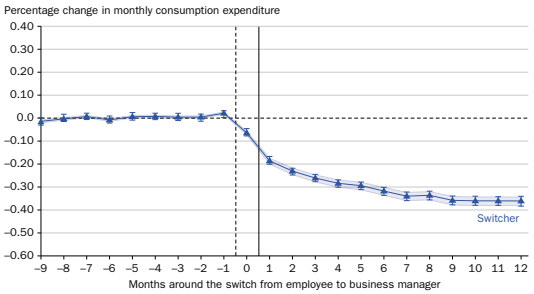
Findings:

- US$1 million earners (top 0.01 percent) pay 20 percent of their income in taxes, while the average tax rate in the economy is 42 percent.
- The tax system is regressive: Income taxes are slightly progressive but insufficient to offset the large regressive consumption taxes.

Policy implications: Designing equitable tax reforms requires considering the effects of indirect and direct taxes, including those collected at the corporate level (refer to figure 8.10).

The Case of Portugal: The Blurred Line between Business Inputs and Personal Consumption

Figure 8.11 Reported consumption drops after becoming a business manager in Portugal, 2016–19



Source: Adapted from figure 2 from Leite (2024).

Question: How does personal consumption via the firm affect tax fairness and income inequality?

Data: Expenditures from e-invoicing data, social security records, and PIT returns and firm accounting data.

Findings:

- Owner–managers of closely held firms shift one-third of their personal expenditures to the firm.
- This generates corporate tax deductions, leading to lower corporate tax payments and lower VAT: Revenue lost amounts to 1 percent of gross domestic product (GDP) annually.

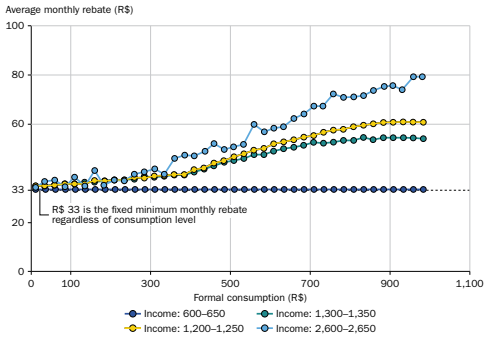
Policy implications: Consumption through the firm generates large tax losses. Because firm owners belong to the top of the income distribution, addressing this practice matters for tax equity (refer to figure 8.11).

E-invoicing: Toward granular insights

The examples that follow illustrate how transaction-level e-invoicing data can be repurposed to improve tax policy design, price monitoring, and public procurement analysis.

The Case of Brazil: Personalized and Targeted VAT Cashbacks

Figure 8.12 VAT rebates in Brazil, by income and consumption levels, 2023



Source: Adapted from figure 3 from Bachas, Flores, et al. (2025).

Question: How does a VAT cashback program, aimed at reducing consumption taxes for low-income households, work in practice?

Data: Administrative records on beneficiaries' formal consumption and cashback.

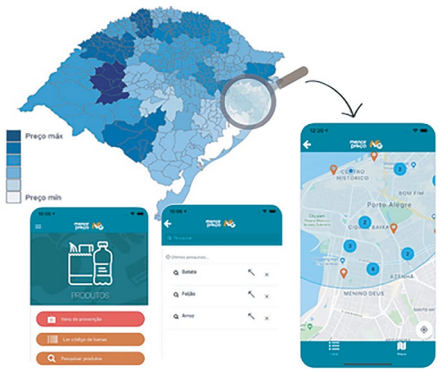
Findings:

- Over 90 percent of the lowest-income households include their national IDs sometimes when making formal purchases.
- The incentives to request formal invoices are muted: Due to a lump-sum part of the program, only 40 percent receive more than the minimum transfer.

Policy implications: VAT cashback programs that require formal identification at the point of sale can help low-income individuals, but program design is key (refer to figure 8.12).

The Case of Brazil: Tracking Consumer Prices in Near Real Time

Figure 8.13 Near-real-time information on consumer prices in Brazil



Sources: Figure based on Receita Estadual's Dynamic Prices Bulletin (*Boletim de Preços Dinâmicos*), State Secretariat of Finance of Rio Grande do Sul (SEFAZ-RS) (https://receitadoc.sefaz.rs.gov.br/), and images from the Menor Preço, Nota Gaúcha mobile application.

Question: How can real-time fiscal data be used to improve the transparency of consumer food prices across regions?

Data: E-invoicing data with detailed product classification.

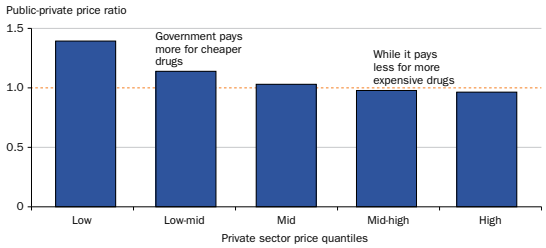
Findings:

- Prices for 65 food products are updated monthly, providing near-real-time information and improving both the timeliness and geographic coverage relative to traditional price surveys.
- Enables the creation of interactive dashboards that enhance transparency and reveal market dynamics across regions.

Policy implications: Electronic invoice data can be a powerful tool to reduce information asymmetries, empowering consumers with transparent price data and supporting real-time monitoring of inflation (refer to figure 8.13).

The Case of Brazil: Public Procurement Prices—Are Governments Getting a Good Deal?

Figure 8.14 When do governments pay more? Brazil, 2018–23



Source: Adapted from data from Best et al. (2024).

Question: Do governments pay more than private firms for the same goods?

Data: Firm-to-firm and firm-to-government transaction records from e-invoices and public procurement auction data from the health sector.

Findings:

- On average, the government pays 14 percent higher prices than the private sector for identical medications.
- The price wedge is largely driven by supplier selection—the government contracts with firms that charge higher prices.
- Governments perform best in markets where they are major buyers and for expensive goods.

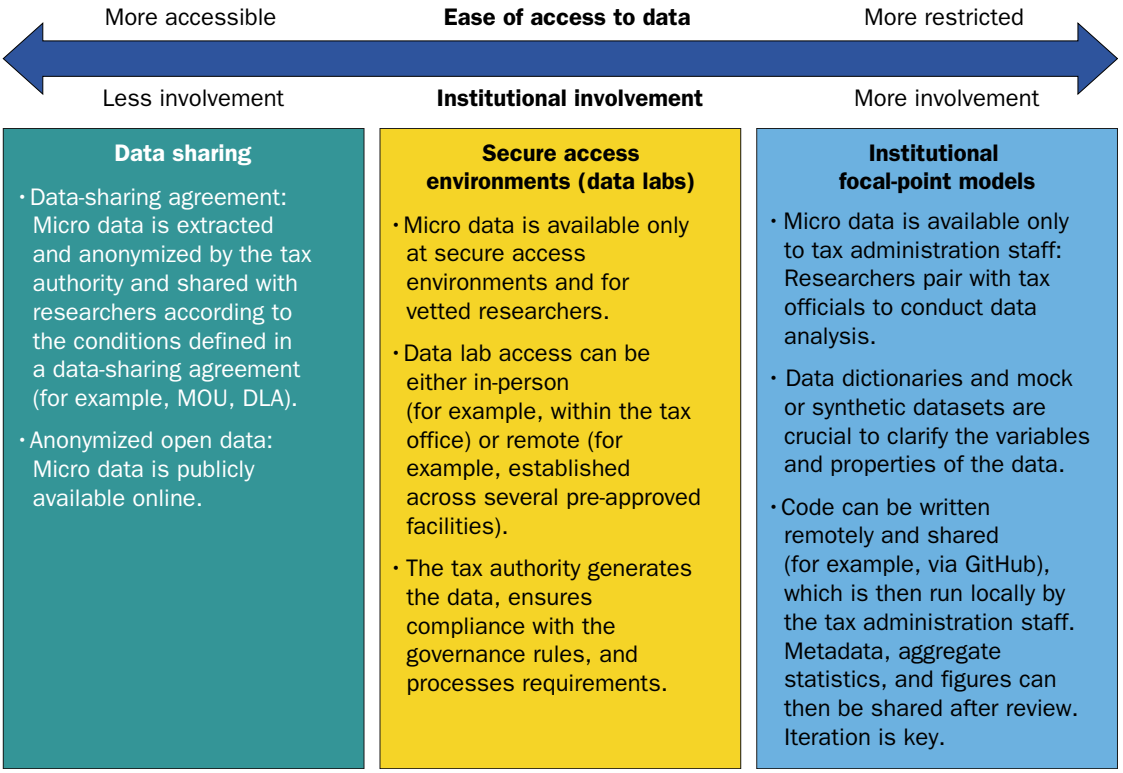
Policy implications: Improving procurement outcomes depends less on auction design and more on broadening access to lower-cost suppliers (refer to figure 8.14).

Building collaborations: Practical models for every need

Securing privacy while unlocking policy insight. Administrative data collected by tax authorities contains sensitive information, including names, addresses, income, and business transactions, which must be carefully protected. To ensure confidentiality, legal frameworks, internal protocols, and access restrictions should be in place to allow *only* authorized individuals to use the data for a predetermined purpose. Tax data hold immense potential to inform tax policy, compliance strategies, and our understanding of economic activity. Tax authorities often recognize this potential and are increasingly open to collaborations (Pomeranz and Vila-Belda 2019). Although a tradeoff between data usage and taxpayer privacy exists, several types of collaboration models enable meaningful policy analysis while fully complying with data protection (Bowen 2024; Burman et al. 2024; Slemrod 2025).

Practical collaboration models. This section presents several models for collaboration that tax authorities and researchers can adopt, ranging from secure data-sharing agreements to in-house data labs and remote analysis setups with code submission. Each model varies in terms of institutional involvement and technical setup, but all share two core principles: flexibility and security. The approaches described as follows can be adapted to meet diverse legal and capacity needs, allowing tax administrations to responsibly unlock the value of their data. These models are grouped into three broad categories that reflect increasing levels of institutional involvement: data sharing, secure access environments (data labs), and institutional focal-point workflows (refer to figure 8.15).

Figure 8.15 Types of collaboration models



Source: Original figure for this publication.

Note: DLA = data license agreement; MOU = memorandum of understanding.

Data-sharing models

Data sharing is the simplest form of collaboration between tax authorities and researchers. As such, it requires the least involvement from tax administration staff. This section discusses open-access tax data (publicly available) and arrangements governed by data-sharing agreements.

Public access microdata. In some contexts, tax authorities provide open online access to tax data. This practice has been implemented in Scandinavian countries; for example, in Norway, tax records have been available online since 2001, enabling public access to everyone's taxable income (Bø et al. 2015). The release of these records was highly popular among Norway's inhabitants and generated more views in the first week than all of YouTube during the same period (Perez-Truglia 2020). Similarly, Mexico's tax administration makes personal and corporate tax data available on its website (refer to https://www.sat.gob.mx/minisitio/DatosAbiertos/sat_mas_abierto.html).

These data are anonymized, with numerical values perturbed to make matching with external sources difficult. Although tax administrations release the data publicly to meet transparency mandates, allowing anyone to access them without a formal agreement, in many countries such public access requires tax administrations' collaboration with local universities. This requirement has promoted partnerships between the administrations and top researchers and helped build domestic research capacity.

Data sharing. Most governments are reluctant to release open-access data. More often, tax administrations provide anonymized tax records to researchers. In this model, researchers gain access to microdata on individual tax filings after all identifying information, such as tax identifiers, names, and addresses, is hidden or removed. From the researchers' perspective, this model allows direct access to the data from any location. For tax administrations, this is the simplest arrangement, because they need only prepare and anonymize the data without participating in any subsequent analysis.

Determining which data are shared. Determining precisely which data will be shared and carefully managing the anonymization process is key to this collaboration model. Suppose a researcher wants to investigate the impacts of a program that exempts specific firms from CIT, which is common in special economic zones. The researcher will need access to the list of eligible firms; their CIT filings reporting their profits and investments; and, possibly, their VAT filings, which include higher-frequency measures of revenues and inputs used. To protect privacy, tax administrations would need to generate "ID masks" for tax identifiers that are consistent across datasets and over time (that is, the same taxpayer is masked using the same unique randomly generated identifier). This enables the researcher to link datasets consistently, even when adding new data, while maintaining anonymity and upholding privacy. Tax administrations should also provide data documentation, such as variable dictionaries, for analysts to make appropriate use.

Formalizing access via data-sharing agreements. Data sharing is typically formalized by means of memorandums of understanding (MOUs) or data license agreements (DLAs). These agreements define the terms under which researchers can access administrative tax data. An *MOU* is a broad, flexible, and nonlegal agreement outlining mutual intentions and roles, while a *DLA* is a more specific and legally binding agreement that details the terms of data use. Both instruments play a central role in structuring a collaboration: They should clearly specify the data shared, for which purposes they can be used, how they will be stored, and the duration of their accessibility and by whom. These agreements protect the interests of all parties: the data provider; the data users; and, most importantly, the data subjects, the taxpayers whose information is contained in the data (Bjarkefur et al. 2021; O'Hara 2020).

What should a data-sharing agreement contain? Data-sharing agreements offer a flexible framework that can be tailored to tax administrations' and researchers' specific needs. While the content of the agreement may vary depending on the scope and complexity of the collaboration, it should contain certain key elements to facilitate effective coordination and promote responsible data use, handling, and protection. Table 8.1 presents a nonexhaustive list of recommended elements categorized as either critical (must have), desirable (should have), or optional (could have).

Table 8.1 Suggested elements of data-sharing agreements

Category	Element	Description
Critical	Parties involved	Identify participating institutions, and define their roles.
	Objectives	State collaboration goals and permitted data uses.
	Dataset details	Provide type, name, description, and metadata of dataset.
	Data handling	Define data format, how data will be accessed or delivered, usage rights, handling rules, and secure storage requirements.
	Personal data protection	Specify how identifiers are protected (for example, data anonymization).
	Confidentiality and intellectual property	Set rules for data ownership, confidentiality, and citations.
	Agreement terms	Define duration, renewal, and amendment rules.
Desirable	Points of contact	Assign responsible individuals or teams.
	Communication of results	Define the scope and process for publishing, sharing, or disseminating results.
	Quality standards	Set data validation checks and procedures to improve data quality.
Optional	Additional data requests	Define procedures for requesting additional datasets.
	Other types of collaboration	Specify other forms of collaboration, such as capacity building, training, and lessons learned.

Source: Original table for this publication.

Extending data sharing across government. Many of the same principles of data sharing between tax administrations and researchers apply to data sharing within government entities. Data sharing across public sector institutions (for example, ministries, agencies, and departments) allows governments to build integrated datasets that support policy design and evaluation. Box 8.1 discusses common approaches to interagency data sharing and highlights the potential to link tax data with other administrative sources.

Box 8.1 Breaking data silos for smarter policy

Fragmented data, limited insights. Government agencies often rely on partial datasets because interagency data sharing is limited. This siloing hampers the effective design and evaluation of public policies. Moving toward integrated, interoperable data systems can enable timely and evidence-based decision-making.

An example: Social mobility across generations. Governments often implement policies aimed at promoting social mobility. Public investments in health, education, and social transfers can enable children from low-income families to achieve higher living standards as adults. Estimating the extent of social mobility is daunting because it requires data on both parents' and children's incomes across decades. This is especially so in countries with high informality, where tax records provide a limited view of the income of poorer individuals. In Brazil, measuring intergenerational mobility precisely was possible by linking data from different agencies: tax records, social security data, public benefit registries, and surveys, which revealed that Brazil has much lower mobility than higher-income countries and that there are large disparities across regions and race (Britto et al. 2025).

Integrating data across institutions. Pooling data sources that are often fragmented across institutions, such as tax administrations, social security agencies, central banks, and business registries, requires strong interagency coordination. Countries can adopt several approaches to achieve integration:

- **Data-sharing agreements** allow agencies to share data for specific projects based on agreed conditions, frequency, and security (refer to the section "Data-sharing Models" in the chapter text).
- **Centralized microdata repositories** led by tax administrations, national statistical offices, or public-interest consortia link, anonymize, store, and provide secure access to administrative data (refer to Statistics Denmark, https://www.dst.dk/en/TilSalg/data-til-forskning, and France's Secure Data Access Center, https://www.casd.eu/en/).
- **Interoperable data-sharing platforms,** a key element of digital public infrastructure (DPI; refer to chapter 3 for its relevance in tax administration), take a systemic approach by establishing shared protocols and infrastructure to enable secure, real-time data exchange across agencies. Rather than building one-off connections, agencies can "plug into" a common platform that ensures scalability, security, and trust while continuing to operate their own

(Continued)

Box 8.1 Breaking data silos for smarter policy *(continued)*

systems (Clark et al. 2025). A touted example is Estonia's X-Road (https://e-estonia.com/solutions/interoperability-services/x-road/), which provides interoperability, data integrity, and privacy for secure data sharing across agencies.

Tax administrations can lead the way. Effective data integration requires standardized identifiers, secure access protocols, and strong governance. Tax administrations are well positioned to start with modest projects (for example, collaborating with one other agency on a specific project) and showcase the value generated for the entire government. Successful collaborations can then catalyze systemic change by promoting interoperable systems and data standards while reducing resistance.

Secure access environment models: Data labs

As administrative tax data become abundant, complex, and valuable, tax authorities are seeking more structured and secure mechanisms to support long-term collaborations. One such mechanism is the establishment of secure access environments known as *data labs.*

A lab to generate policy insights. Tax data labs, which are managed by tax administrations or in collaboration with other institutions, provide centralized and secure access to de-identified microdata. In some countries, legal frameworks restrict direct access to tax data exclusively to the tax administration, thereby limiting the integration of these data with other sources or sharing them for external analysis. Tax data labs, on the other hand, make it possible to handle all data within the tax administration's servers, which process, clean, and anonymize the data before making them accessible within a secure and controlled environment. This model protects data confidentiality while leveraging their access and use for policy insights.

The expansion of data labs. Secure data environments, such as the Federal Statistical Research Data Centers in the United States and the Restricted Data Access Lab (Sala de Acesso a Dados Restritos) in Brazil, have long provided researchers with access to microdata from national statistical offices. Building on similar principles, tax data labs have emerged more recently and gained traction in HICs; for example, the United Kingdom's His Majesty's Revenue and Customs Datalab and France's Secure Data Access Centre. Over time, the presence of these labs has increased in low-income countries (LICs) and middle-income countries (MICs) such as South Africa (refer to box 8.2), Uganda, and Zambia due to the support of the United Nations University–World Institute for Development Economics Research Domestic Revenue

Mobilization program, and Honduras recently launched the DATALAB SAR (refer to https://www.sar.gob.hn/laboratorio-de-datos/) with the support of the World Bank. These examples show that institutionalizing secure access to tax data is both feasible and valuable.

Designing the right lab. Setting up a tax data lab involves a series of strategic decisions shaped by policy objectives and by institutional, legal, and technical constraints. Table 8.2 summarizes the key elements to take into consideration when designing a tax data lab: technical infrastructure, governance rules, and operational processes:

- *Technical infrastructure* involves the lab setup (physical or virtual); the systems used to store and access data; and the protocols, if any, for managing remote access.
- *Governance rules* define the lab's management structure, the scope of data included, and user access policies and fees.
- *Operational processes* include determining the criteria for proposal evaluation, the protocols for user authorization, and privacy safeguards, such as output checking.

Box 8.2 South Africa's National Treasury Secure Data Facility

South Africa stands out as a leading example of how secure access to administrative tax data can fuel high-impact research and inform better policy making.

The origins. In 2014, South Africa created the National Treasury Secure Data Facility (NT–SDF), a collaboration among the National Treasury, the Revenue Service (SARS), and the United Nations University–World Institute for Development Economics Research (UNU–WIDER). Conceived as a secure access environment for administrative tax data, the NT–SDF initially operated out of National Treasury's offices in Pretoria with just three computers and a small team of data scientists. As demand for data access grew rapidly, policy makers recognized the value of these datasets, and the NT–SDF scaled up, acquiring more computers and adding new datasets in collaboration with users, making the NT–SDF a dynamic space where data infrastructure and research needs evolved hand-in-hand.

How does it work? The NT–SDF lab is structured around a secure workflow aligned with international best practices. Researchers submit project proposals, which are reviewed for policy relevance, feasibility, and legal compliance. Once a project is approved, users complete training

(Continued)

Box 8.2 South Africa's National Treasury Secure Data Facility *(continued)*

on legal obligations, data security, and compliance protocols before being granted onsite access to de-identified tax microdata. The secure environment is equipped with statistical software, and all outputs—tables, figures, and results—are subject to tax administration review to safeguard sensitive information.

Scaling up. By 2017, demand again outpaced the NT–SDF lab's capacity. In response, NT—supported by the Capacity Support Program and international donors—invested in expanding and institutionalizing its secure access model. This expansion took shape under the Southern Africa–Towards Inclusive Economic Development (SA–TIED) program; dedicated full-time staff were hired to manage daily operations and expand data availability. Today, the NT–SDF hosts more than 10 major administrative datasets, including corporate income tax, personal income tax, value-added tax, and payroll, and has supported the publication of over 175 studies and 60 papers on a wide range of topics.

A platform for national data integration. Building on the NT–SDF's success, South Africa is now developing an Integrated Data Lake, an ambitious initiative to combine administrative and survey data from across the government into a single standardized platform. The goal is to foster cross-agency data exchange, streamline data management, and provide a unified resource for data analysis. As part of this initiative, the NT–SDF was rebranded as South Africa's national secure data facility, positioning it at the heart of the country's data ecosystem to support government decision-making and respond to societal challenges.

Sources: This case study draws on information from NT–SDF (refer to https://www.treasury.gov.za/research/NT-SDF/NT-SDF%20Overview/Default.aspx), SA–TIED (refer to https://sa-tied.wider.unu.edu/about), and UNU–WIDER (https://www.wider.unu.edu/publication/building-tax-data-research), as well as on presentations delivered by Ayanda Hlatshwayo, Acting Chief Data Analytics Officer, at the National Treasury at the World Bank. Refer also to Hlatshwayo (2025).

Although the design choices presented in table 8.2 are binary, variations are often possible. For example, proposal evaluations may combine policy relevance and academic merit, and a single agency could host the tax data while collaborating with other institutions to incorporate other data.

One lab, one workflow. Tax data labs help establish a standardized and transparent workflow. While procedures vary across countries, most labs follow similar steps to examine a project. The process begins when a team interested in analyzing the tax data submits a proposal. The tax administration or designated scientific committee reviews the proposal based on its policy relevance, academic merit, and feasibility. If a proposal is approved, then the researchers are required to complete training modules on legal obligations, compliance protocols, and data security practices. From there,

Table 8.2 Institutional design elements of secure access environments

Pillar	Feature	Decision	Pros	Cons
Technical infrastructure	**Access environment**	On premises (physical lab)	• High security • Controlled environment • Less cyber risk	• Limited researcher access (subject to PC availability) • Costly to maintain and scale
		Remote access (virtual lab)	• Flexible for researchers • Scalable	• Cybersecurity risks • Requires robust monitoring • Complex setup
	IT infrastructure	Local servers	• Complete data control • High data security	• Requires physical space and ongoing maintenance
		Cloud based or hybrid	• Scalability • Flexibility • Potentially lower costs	• Sensitive data management complexities (for example, legislation might not allow it)
Governance rules	**Managing institution and scope of data**	Single-agency governance	• Clear accountability • Centralized oversight and streamlined decision-making and processes • Focused data scope (that is, tax, employment, social protection, and education)	• Potential for narrow focus • Limited cross-sector collaboration • Reduced stakeholder engagement
		Multi-agency governance	• Shared responsibilities • Improved buy-in • Cross-sector integration	• Complex coordination • Legislation or legal agreements between institutions perhaps necessary
	Access fees	Free access	• Equity • Promotes research	• Limited cost recovery
		Fee-based access	• Helps sustain the lab • Covers maintenance costs	• Barriers for underfunded researchers
Operational processes	**Proposal evaluation criteria**	Policy relevance	• Ensures alignment with institutional goals and increases policy impact	• May exclude important but less immediately relevant research
		Academic merit	• Promotes academic rigor and innovation	• Risk of limited institutional buy-in or misalignment with tax authority priorities
	Privacy safeguards (output checking)	Manual review	• Ensures no disclosure	• Time-consuming • Delays researcher outputs
		Automated tools	• Faster processing • Scalable	• May miss complex disclosure risks

Source: Original table for this publication.

Note: IT = information technology; PC = personal computer.

the researchers are granted access to the relevant datasets within a secure environment. In-person labs provide onsite access, while virtual labs enable remote access via secure terminals. These environments are equipped with standard statistical software and configured to meet a project's technical requirements. Throughout the project, the tax administration reviews the outputs to ensure that no sensitive information is disclosed. Any statistic should be built with a minimum number of observations (for example, at least 10) to avoid de-anonymization. In some cases, analysts are expected to collaborate with tax officials, which can improve understanding of the data and ensure that a project remains aligned with policy goals.

Focal-point models: Workflows without direct data access

When government officials alone can access the data. In some countries, legal restrictions prevent outside analysts from accessing tax records, even in an anonymized form, reserving access only to tax officials. In such cases, researchers can collaborate via an institutional focal point: a government official authorized to access and analyze the data. This model demands substantial engagement from the tax administration by requiring one or more officials to work closely with researchers throughout the project. This dynamic, however, also yields large benefits: Researchers and government analysts learn from each other, fostering knowledge transfer in policy, data, and analytical techniques.

Conducting empirical research without seeing the data. When analysts do not have direct access to data, tools like data dictionaries and synthetic ("mock") datasets are essential for developing a functional workflow. A *data dictionary,* which lists available variables along with their definitions and formats, is indispensable for writing code appropriately. Even more useful is access to *synthetic data:* a simulated dataset that mimics the structure of the real data without revealing any sensitive information. Although synthetic data do not contain actual taxpayer records, they preserve key features of the original data such as variable names, types, and formats. Researchers can use the mock data to write, test, and refine code before it is run on the confidential dataset.

Iteration is key. Once researchers understand the structure of the synthetic data, they may conduct their entire analysis without ever directly accessing the confidential data. In other words, they can share code—via platforms like GitHub—with their designated tax administration contact who, in turn, can run the code to generate descriptive statistics, such as the number of observations, mean values, or tabulations of key variables. The code may not run perfectly at first, prompting government partners to troubleshoot and adjust it. Based on the initial outputs, researchers request further statistics or investigate unexpected patterns, initiating an iterative process that continues until the final outputs—tables, figures, or regressions—are complete.

Which information can be extracted? An important element of such collaborations is agreeing on which types of aggregated data may be disclosed and setting minimum observation thresholds for statistical calculations. For instance, disclosing the total profits of all firms is typically considered acceptable, whereas publishing the average profits of the three largest taxpayers might be problematic. Although such statistics are not at the individual level, their high granularity could risk exposing sensitive information. To protect taxpayer privacy, tax administrations often impose disclosure rules—for example, requiring that any publicly released statistic be based on a minimum of 5–10 underlying observations. These safeguards allow research using detailed information while protecting taxpayer confidentiality.

Analytical tools for big administrative data

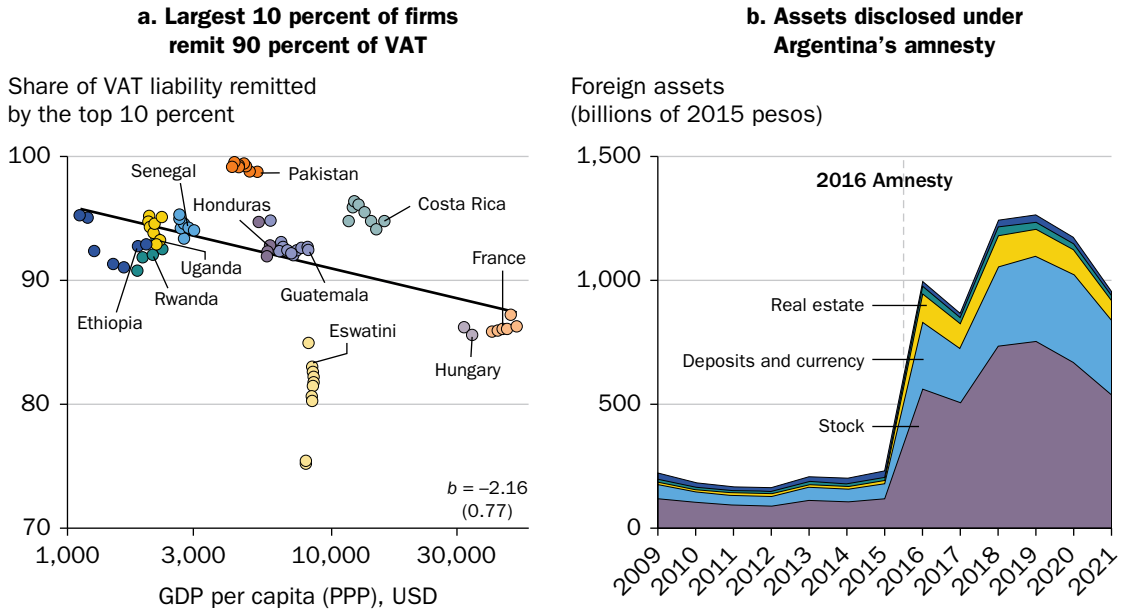
Using the right analytical tools. Certain administrative tax records are better suited to specific types of analyses. Data features such as comprehensive coverage, temporal depth, and granularity facilitate the use of specialized analytical tools to address policy questions. This section provides an overview of these methods along with illustrative examples.

Descriptive analysis: Learning from the data

Starting with the basics. Tax administrations often publish aggregate data, such as total revenue by tax type or overall tax expenditures. Aggregates are valuable to monitor public finances and the macroeconomy. For example, monthly VAT revenues may signal shifts in economic activity. Individual-level tax records allow for richer descriptive analysis and insights into taxpayer behavior, distributional patterns, and policy impacts. Despite their potential, microdata remain underused, even for simple descriptive statistics.

Learning from your neighbors. Cross-country comparisons using microdata can shed light on the quality of tax policy implementation and tax administration practices. For instance, consider VAT; while figures like total VAT collected or tax expenditures can be compared across countries using simple aggregates, access to microdata can provide more granular diagnostics. Figure 8.16, panel a, shows that, across countries with vastly different GDP per capita, the largest 10 percent of firms by revenue always report more than 85 percent of total VAT liability (Brockmeyer et al. 2024), indicating that, although VAT is a broad-based tax on all formal consumption, a small number of firms are responsible for the majority of the reported value added and tax payments. The use of firm-level declarations made this insight possible. As microdata become increasingly available, conducting cross-country studies should become increasingly feasible.[1]

Figure 8.16 Simple descriptive analyses can be very informative



Sources: Panel a is used with permission from Brockmeyer et al. (2024), and panel b is used with permission from Londoño-Vélez and Tortarolo (2025).

Note: GDP = gross domestic product; PPP = purchasing power parity; USD = US dollars; VAT = value-added tax.

Sometimes plotting the microdata is all that is necessary. Descriptive work based on tax records can also provide a detailed assessment of tax reforms, particularly when the effects are large. Consider Argentina's 2016 voluntary disclosure program, which exempted taxpayers from paying tax penalties for previously undisclosed assets. Detailed tabulation from tax records revealed that taxpayers disclosed new assets equivalent to 21 percent of the GDP (Londoño-Vélez and Tortarolo 2025). Over 80 percent of these newly disclosed assets were held abroad, with the majority held in offshore stocks (refer to figure 8.16, panel b). The same study shows that tax evasion was concentrated at the top: 0.1 percent of individuals with the largest wealth revealed most of the undisclosed assets.

Learning from both data and theory. While simple descriptive analysis using microdata can reveal key insights, it also has limitations. To answer the question, "What is the causal impact of a policy on the economic outcomes we care about?" researchers will often rely on a combination of knowledge from theory, insights from data, and specific statistical and econometric tools, which we discuss as follows.

Discontinuities and thresholds

The world is not smooth. Real-world policies often display discontinuities: Firms that exceed a certain number of employees become subject to stricter labor regulations, workers who make a specific number of quarterly social contributions qualify for retirement benefits, and businesses with an annual revenue above a threshold must register for VAT. These conditions are implemented for practical reasons: to reduce compliance costs and administrative burdens, especially for smaller economic actors. At these thresholds, taxpayers' incentives change abruptly, creating opportunities to study behavioral responses and estimate policy parameters. For example, one can assess tax elasticities: What is the change in taxpayers' reported base when their tax rate changes by 1 percent? While policy discontinuities and thresholds have long existed, they have become far more useful for studying behavior as administrative data and modern econometric methods have expanded.

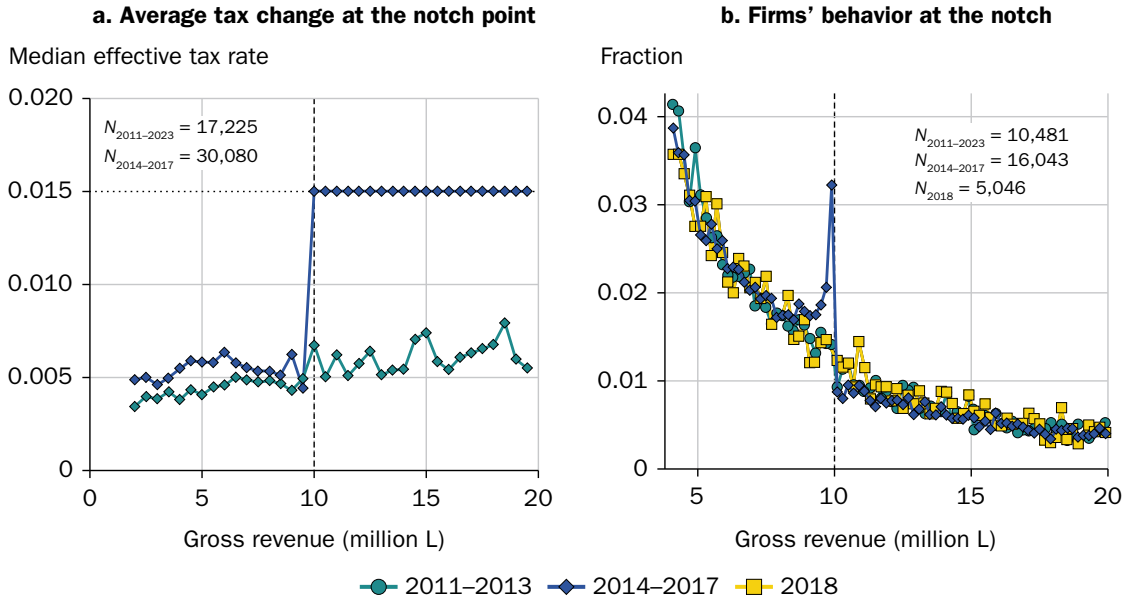
Everyone wants big data—For discontinuities, it is a necessity. Tax data are particularly useful for analyzing policy discontinuities for several reasons. First, they cover the full formal population, offering large sample sizes that are essential for econometric techniques that rely on variation around thresholds. Second, although administrative records may be strategically manipulated, they are recorded at high frequency and with less noise than self-reported or recall-based survey data. This combination of scale and precision makes tax data uniquely powerful for detecting sharp behavioral responses at policy cutoffs.

Manipulation experts: Bunching. The first econometric technique well-suited to administrative data is bunching. *Bunching* is used to estimate how people or firms respond to policies above or below a threshold. For example, a firm might face a higher tax rate (the policy parameter) when its revenue (the running variable) exceeds the threshold. Firms wanting to escape the higher tax rate might report a revenue just below the threshold. If this were the case, then an analyst might observe an excess of taxpayers with revenue at the threshold, or "a bunch." By combining the size of the bunch with the change in incentives at the threshold, the analyst could estimate the responsiveness of firms' taxable revenue to the tax rate (tax elasticity). In bunching estimation, a key hypothesis is constructing a credible counterfactual distribution of taxpayers along the running variable. In other words, if the tax rate had remained constant, what would have been the shape of the (smooth) distribution of taxpayers along the running variable?

Bunching varieties: Kinks and notches. There are two main bunching designs: kinks and notches. *Kinks* correspond to thresholds that generate a change in the marginal tax rate:

The excess revenue above the threshold is taxed at a higher rate. Kinks generate a bunching mass around the threshold (both below and above). *Notches* correspond to thresholds that generate a change in the average tax rate: The full revenue past the threshold is taxed at a higher rate; therefore, notches only generate a bunching mass below the threshold and too little mass above it. Figure 8.17, panel a, shows an example of a tax notch in Honduras: As the firms' revenue surpasses 10 million, their effective tax rate triples. As a result, firms face strong incentives to report revenue below the threshold. Indeed, figure 8.17, panel b, shows that the fraction of taxpayers with revenue below the threshold is three times larger than the fraction of taxpayers reporting revenue just above. Compared to 2011–13, when there was no notch, the taxpayer distribution was continuous, providing a credible counterfactual (Lobel et al. 2024). Because these techniques plot distributions, they require large and precisely recorded data, key attributes of administrative tax data, whose increased availability has led to the development and refinement of the econometrics of bunching (for detailed guides, refer to Bertanha et al. [2020] and Kleven [2016]).

Figure 8.17 Bunching at a tax notch: Incentives and responses in Honduras



Source: Figure used with permission from Lobel et al. (2024).

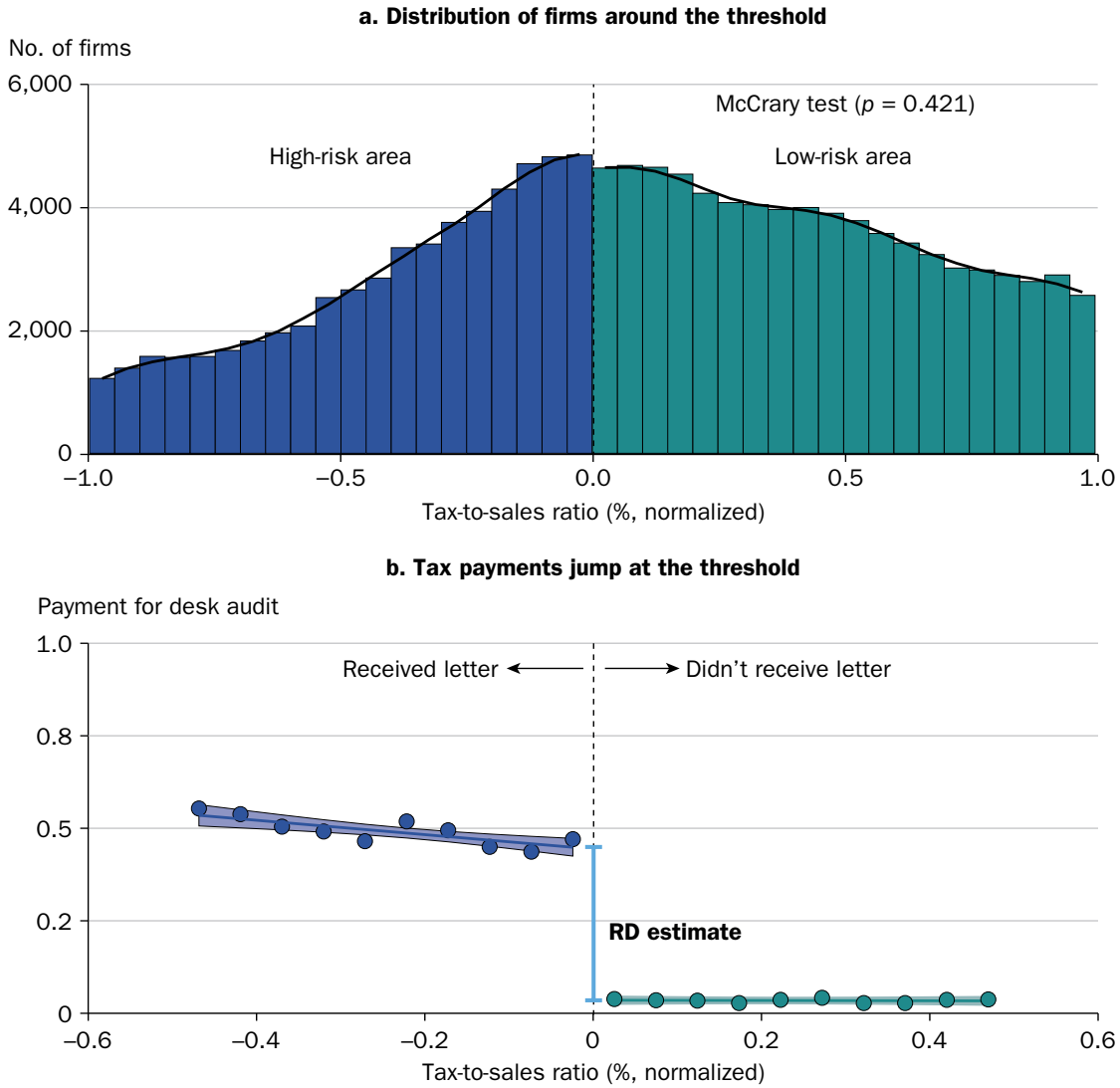
When there is no manipulation: Regression discontinuity. Econometrically, regression discontinuity is a close cousin of bunching that also relies on analyzing data around a policy threshold. *Regression discontinuity* is used when a policy changes at a threshold, but the running variable is not manipulated, which means that no excess density at the threshold occurs. If the distribution of the running variable is smooth around the threshold, then taxpayers just below the threshold can be credibly compared to those just above.

Consider an example from China, where firms with a VAT-to-sales ratio below a certain threshold receive a notification from the tax administration requesting justification for their low effective tax payments (Li and Wang 2025). This threshold is computed using past tax records and is reset annually. As a result, at the time of reporting, firms cannot manipulate their tax returns to stay above an unknown threshold, so the distribution of firms around the threshold remains smooth (that is, there is no bunching), as shown in figure 8.18, panel a. This lack of manipulation makes it possible to use a regression discontinuity design to estimate causal effects. In this case, the outcome of interest is the tax revenue recovered through audits. Figure 8.18, panel b, demonstrates that recovered revenue is significantly and discontinuously higher just to the left of the threshold (where firms are notified) than to the right (where firms are not). Because firms near the threshold are otherwise similar, assignment to treatment (audit) is effectively random within a narrow interval, allowing any observed difference in outcomes to be attributed to the policy. As with bunching, regression discontinuity estimations require a large volume of data near the threshold to obtain precise estimates (for methodological guidance, refer to Cattaneo et al. [2020, 2024]).

Artificial intelligence, machine learning, and high-frequency data

Bring on the machines. Policy makers involved in predictions in high-dimensional settings—that is, who use a large number of variables to forecast outcomes of interest—face some of the most-pressing challenges. A prime example is audit selection. To predict which taxpayers are most likely to evade taxes, tax administrations must rely on vast amounts of information from individual taxpayers, firms, suppliers, and so on. For tasks like this based on complex, data-rich problems, modern machine learning (ML) and artificial intelligence (AI) techniques—combined with administrative records—offer powerful tools to improve predictive accuracy and support policy decisions (Aslett 2024; Junquera-Varela et al. 2022).

Figure 8.18 Regression discontinuity: Tax audit risk and recovered revenue in China



Source: Panels a and panel b adapted from figures 2 and 5, respectively, from Li and Wang (2025).

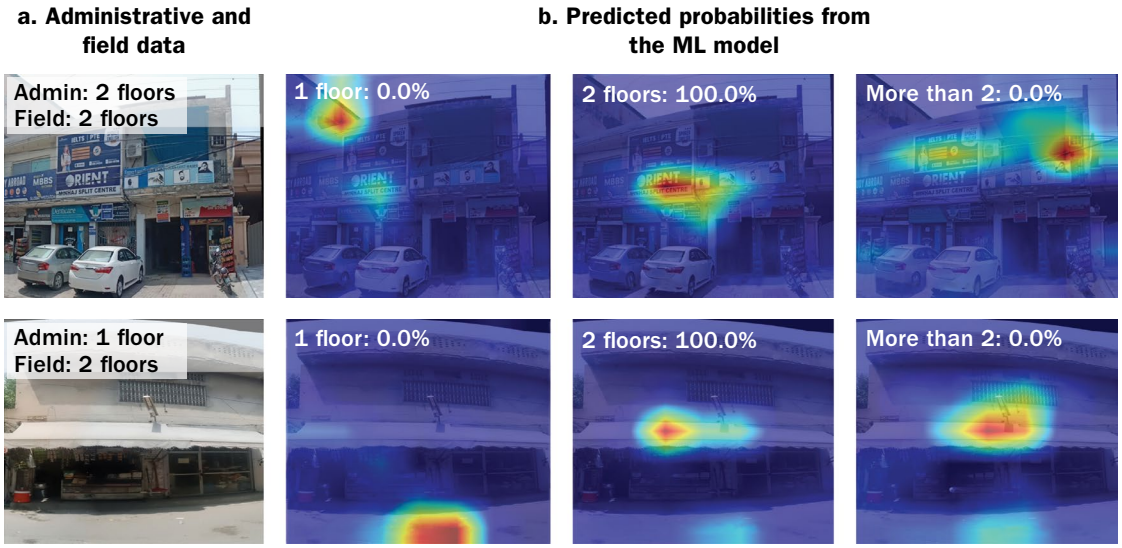
Note: RD = regression discontinuity.

Targeting scarce resources. How should policy makers allocate scarce resources? Although tax audits, for example, can uncover evasion, they are costly to implement and intrusive. To address this, tax administrations are increasingly developing risk models—often powered by ML—to prioritize enforcement actions (Boning et al. 2024; Caspi et al. 2024). ML models are well-suited to this task, especially when combined with random audits that help establish "ground truth." By auditing a random subset of taxpayers and observing their compliance, tax authorities can train models to predict other taxpayers' evasion. However, accurate prediction alone does not guarantee improved compliance or revenue collection. In India, for example, a highly accurate model for detecting fraudulent firms failed to raise revenue due to its limited integration into day-to-day practices (Barwahwala et al. 2024). Moreover, algorithms do not necessarily outperform humans: In Senegal, inspector-selected audits were sometimes better targeted than algorithm-selected audits, particularly when the data available for model training were limited (Bachas, Brockmeyer, Ferreira, et al. 2025).

How much does that property cost? Another way in which ML models can be used to mobilize revenue is to improve property taxation. The biggest challenge that many LICs and MICs face when taxing properties is determining correct property valuation. While computer-assisted mass appraisal (CAMA) tools have become more common, many countries still rely on tax officers to conduct manual appraisals. These two approaches can differ markedly, as a tailored CAMA in Dakar, Senegal, provides more accurate property valuation, particularly at the top of the distribution, than bureaucrats (Knebelmann et al. 2026). Another common challenge is that property tax rolls are often incomplete or outdated. ML models trained on property images and administrative data offer new opportunities to address this problem by helping reconstruct missing records and fill gaps in survey data (Bergeron et al. 2023). In Punjab, Pakistan, image-based ML models are being used to improve property tax assessments by estimating key attributes, such as the number of floors, that can update or complement existing administrative data, potentially leading to more accurate property valuations (Asad et al. 2025). Figure 8.19 shows how these models can work in practice by highlighting the parts of the image that influence the prediction of the number of floors.

Better data for better targeting. Another application of ML and AI is in *social assistance targeting,* or determining who should receive support from social programs. Classical approaches rely on self-reporting of income, community targeting, or asset-based welfare measures (Grosh et al. 2022). Novel literature documents the benefits of using satellite

Figure 8.19 Using image-based ML to identify property characteristics in Pakistan



Source: Adapted from Asad et al. (2025).

Note: ML = machine learning.

imagery and mobile phone data, combined with ML models, to improve targeting (Aiken et al. 2022; Lawson et al. 2023). While the use of tax data for similar purposes is underexplored, electronic invoices are a potential data source for better targeting because formal consumption can be directly linked to individuals, providing detailed measures of expenditures that could be leveraged for targeting.

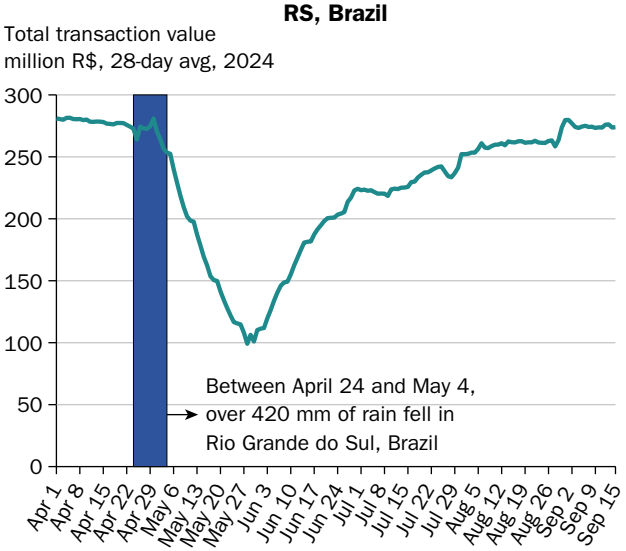
Seeing the world in real time. The high-frequency nature of tax records makes them an invaluable resource to document close to real-time changes in the economy. Consider first the task of measuring GDP, a core indicator of economic activity. Recent research has shown that transaction data from banks in Spain can be used to determine estimates of private consumption consistent with those used to measure GDP in national accounts (Buda et al. 2023). This not only allows for a much higher-frequency measurement of consumption, because data are available in real time, but also gives a glimpse of the distribution of consumption and inequalities across individuals. While this strand of research (refer to Andersen et al. [2026] as well) uses banking transactions, the same application can be made to electronic invoice data, which covers the universe of formal transactions in many countries.

Providing insights in response to crises. Providing near-real-time data on the state of the economy in times of crises is a particularly attractive feature of tax records. Consider, for example, the devastating floods in the state of Rio Grande do Sul, Brazil, in late April to early May 2024, which affected over 2.4 million people. The state government leveraged electronic invoice data to estimate not only how many businesses had to stop operations but also how economic activity evolved throughout the disruption (SEFAZ, RS 2024; refer to figure 8.20). This information supported the design of programs aimed at providing support to affected firms as well as of a VAT cashback program to encourage families to replace household appliances. High-frequency data on firms were also a useful tool at the onset of the COVID-19 pandemic. B2B transaction data, for example, shed light on the impacts of the crisis on production chains in Kenya (Chacha et al. 2024). Harmonized CIT data were also used to estimate the likely impact of the pandemic and responses to it, such as lockdowns, on firms' activity (Bachas, Brockmeyer, Garriga, et al. 2025). In retrospect, firms turned out to be more resilient than initially predicted, and monthly VAT data on firms' revenues were an early indicator of that resilience.

Figure 8.20 Tracking crises with e-invoicing data in Brazil

a. Drone view of the flooded neighborhood of Mathias Velho in Canoas, RS, Brazil

b. Daily economic activity of businesses in affected areas of RS, Brazil



Source: Panel a is used with permission from Reuters (2024), and panel b is adapted from data from SEFAZ, RS (2024).

Note: Panel b shows the evolution of total transaction value from April 1 to September 15, 2024. Total transaction value includes B2B and B2C operations. B2B = business to business; B2C = business to consumer; RS = Rio Grande do Sul.

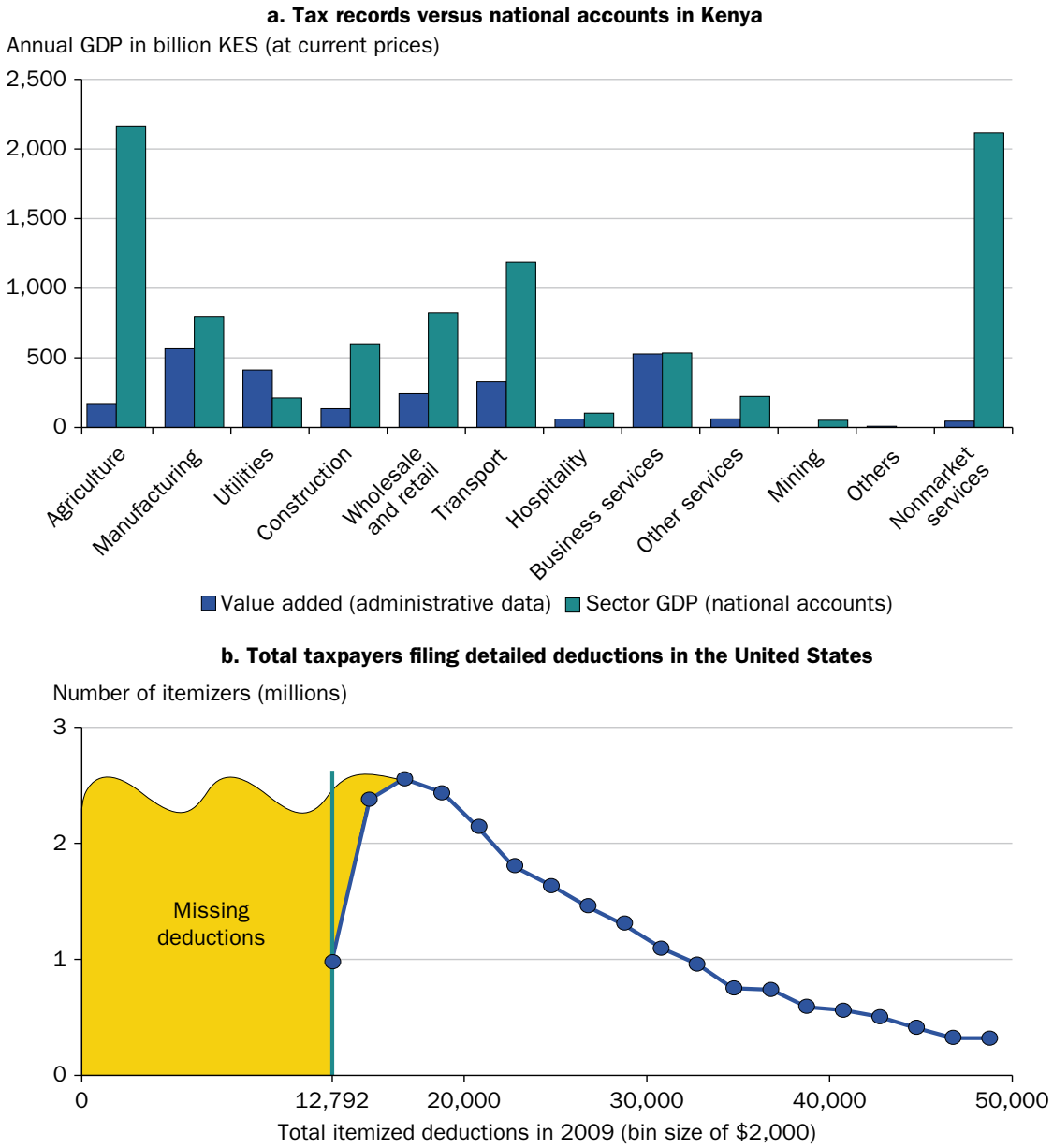
Lurking in the shadows: What are we missing?

So far, we have highlighted the strengths of tax records and how they illuminate otherwise difficult to observe dimensions of the economy. However, like any data source, tax records are not without their limitations. In this section, we discuss key shortcomings of administrative tax data, along with the ethical risks associated with their use, with the hope that researchers and policy makers will be aware of the challenges they may face when using tax data and devise thoughtful strategies to mitigate them.

All the economic activity we cannot see. Unlike surveys, which are designed to produce representative indicators of economic outcomes, tax records are not intended to capture all types of activity. Often, they are a (very useful) by-product of the government's core function to collect taxes. Activities that fall outside of tax records are often referred to as *informal* and include firms not registering as taxpayers, making cash purchases without an invoice, and receiving payments *under the table,* payments that do not appear on official payrolls. There are many reasons why economic agents might choose not to formally register or report certain transactions to tax authorities, among which the primary motive is to avoid taxation. The section "Trade Networks Revealed: The Power of Firm-to-Firm Transaction Data" includes an example of unobserved activities within the context of formal and informal trade flows in Kenya (Wiedemann et al. 2024).

Figure 8.21, panel a, illustrates that tax records miss substantial value added in both agriculture and construction but for different reasons: In agriculture, VAT data capture only large-scale commercial farming, while subsistence agriculture, which is responsible for most of the sector's GDP, remains invisible. In construction, many registered firms underreport their activity. In contrast, tax data perform much better in sectors where production is largely formal, such as manufacturing and business services. Many labor markets exhibit high levels of informality: 60 percent of all workers in the world are informal, with a high prevalence in LICs and MICs. Tax records, therefore, often offer an unprecedented view of *formal* economic activities, but, in highly *informal* settings, they fail to capture a large share of transactions, which should be taken into account when interpreting tax data findings.

Figure 8.21 Limitations of administrative tax data



Sources: Panel a is used with permission from Wiedemann et al. (2024), and panel b is adapted from figure 6, panel a, from Benzarti (2020).

Note: GDP = gross domestic product.

The rules that govern tax recording. Tax laws also limit the potential of tax records. For example, not all registered firms are required to file declarations and report comprehensive information. Firms with revenue below a threshold are often exempt from VAT registration, to reduce compliance burdens, which can be high for small and less sophisticated businesses. As a result, small firms are not included in VAT records. A similar rationale explains why detailed deductions are observable only just above (and not below) a certain threshold in PIT filings—most PIT systems include a simplified deduction that taxpayers can use without itemizing each deduction separately. Figure 8.21, panel b, shows that few US income taxpayers will choose to face the hassle of providing detailed itemizations only to save a few dollars (Benzarti 2020).

A related but distinct limitation arises when taxpayers are required to submit only partial data to the tax administration, which is the case with small firms that are subject to simplified tax regimes. Unlike the standard CIT regime, which mandates detailed financial reporting to calculate taxable profits, small firms are required to report only total sales. This reduces the administrative burden but also limits data availability and precludes the disclosure of key variables such as cost structure and investments, making parts of firms' activities invisible to researchers.

Tax me if you can. One additional limitation of tax records comes neither from fully informal agents nor from those who do not need to report information, but from taxpayers who change their behavior or misreport information in response to incentives. The bunching phenomenon discussed earlier illustrates such a response: If firms pay much higher taxes by declaring to be slightly larger, they will then resort to misreporting their true revenue. While this behavior can help economists estimate key policy parameters, it also distorts the data: A naive analyst might take the data at face value and be befuddled by this odd behavior. Evidence shows that intrinsic motivation to pay taxes can drive compliance at times (for example, if individuals believe their taxes will be used to provide public goods, then they will pay them), but some degree of avoidance and evasion is part of the tax enforcement business and, therefore, of tax records. This does not mean, however, that these behaviors are unchangeable—on the contrary. One data-related example is the increasing use of third-party records to enforce compliance, which lies at the heart of the functioning of VAT systems, where buyers and suppliers must report the same information about purchases or sales, which are checked against each other for accuracy (refer to figure 8.6).

Beyond these limitations, working with tax data also raises important questions about privacy, responsible use, and governance. Box 8.3 discusses how tax administrations and researchers can manage these risks while preserving the value of tax data for policy and research.

Box 8.3 Ethical governance of tax data

Balancing privacy and data use. Privacy is a fundamental right, and intrusions into privacy impose real costs. Concerns about privacy have shaped tax policy for centuries. In 17th-century England, for example, the government replaced a levy on fireplaces with a tax on windows, enabling assessments from outside the home rather than through intrusive household inspections (Oates and Schwab 2015). Modern tax systems rely primarily on self-reporting and third-party information, both of which involve the collection of sensitive personal and financial data. Governments must, therefore, consider privacy when deciding what information to collect, how it may be used, and who should have access to it.

Tax privacy has two dimensions. The first is external: the government's responsibility to protect tax data from unauthorized access or disclosure. The second is internal: the individual's right not to disclose certain information to the government (Slemrod 2025).

External privacy: Safeguards for secure data access. Tax administrations can use a range of safeguards when conducting analytical work. These include de-identification and anonymization, which remove or mask taxpayer identifiers and other personal information, as well as secure analytical environments that allow approved users to access microdata under strict protocols. Advances in privacy-preserving methods have expanded the available toolkit. A prominent example is *differential privacy,* which introduces calibrated statistical noise into outputs while preserving aggregate patterns and distributions (Wood et al. 2022). Other tools include *synthetic public-use datasets,* which reproduce the statistical properties of confidential data without revealing individual records, and verification and validation servers, which allow researchers to analyze sensitive data while limiting disclosure risks. *Verification servers* compare results obtained from synthetic and confidential data, while *validation servers* execute code developed on synthetic data against the confidential dataset and return only disclosure-protected aggregate outputs (Bowen 2024; Burman et al. 2024).

Internal privacy: Tax data governance. As the scope and granularity of tax data expand, so do public expectations regarding their responsible use. The misuse of administrative data for political surveillance, discriminatory targeting, or purposes unrelated to legitimate public policy objectives can erode trust and discourage individuals and firms from sharing information. The 2021 World Development Report, *Data for Better Lives,* provides a framework for managing these risks (World Bank 2021, particularly chapter 6). Among its recommendations are independent-oversight mechanisms, such as ethics review committees, to monitor data collection and use.

(Continued)

Box 8.3 Ethical governance of tax data *(continued)*

Similar to institutional review boards in academic research involving human subjects, these bodies can help ensure that government data are collected and used only for legitimate, well-defined, and proportionate purposes. Transparency about what data are collected, why they are needed, and how they are protected is essential for maintaining public trust. Ultimately, governments should weigh the public benefits of collecting additional taxpayer information against the associated privacy costs.

From tax records to data-driven policy making: A roadmap for strengthening tax data use

Different starting points. The previous sections discussed how tax records can inform policy making, how governments can build the collaborations and capabilities needed to use them effectively, and the challenges involved. In practice, however, countries differ markedly in their readiness to leverage tax data. Some still rely on fragmented or paper-based systems and inconsistent taxpayer registries. Others generate rich digital data but lack the analytical capacity, governance arrangements, or institutional processes needed to systematically embed evidence into policy design and implementation.

The path forward. Table 8.3 translates this chapter's lessons into a practical roadmap for strengthening the use of tax data across different country circumstances using the reportwide framework of structural fundamentals and distance from the fiscal frontier. Rather than prescribing a single path toward greater data use, these recommendations help countries identify priorities based on their starting point and the constraints they face. Where fundamentals are weak, priorities range from establishing basic digital and data infrastructure to building analytical capabilities, depending on how effectively existing capacity is being used. Where fundamentals are stronger, the emphasis shifts toward breaking down data silos, strengthening governance and interoperability, institutionalizing analytics, and adopting more sophisticated approaches to secure data access and sharing.

Table 8.3 Policy priorities for strengthening tax data use

	Far from Fiscal Frontier	Close to Fiscal Frontier
Weak fundamentals	***Constrained and underperforming:*** **Fragmented systems, limited data use** **Current situation** • Some systems remain paper based. • Taxpayer registries are inconsistent. • No data-sharing rules are in place. **Actions** • Assign unique taxpayer identifiers, and harmonize taxpayer registries. • Standardize main datasets. • Plan to digitize core processes (registration, filing, and payments). **Data sharing:** Simple data exchange for pilot collaborations under MOUs.	***Constrained but functional:*** **Strong rules, limited analytical capacity** **Current situation** • Digital processes generate rich data. • Protocols for data access and exchange are in place. • Analytical capacity is the bottleneck. **Actions** • Assemble an analytical unit with statistics and economics skills. • Determine policy-relevant analytical products and priority use cases. • Promote a culture of data-driven decisions across the organization. **Data sharing:** Promote systematic collaboration with external partners with a strong capacity-building component.
Strong fundamentals	***Capable but underperforming:*** **Digitized systems, siloed data** **Current situation** • Data analysis occurs in silos. • Data governance is undetermined. • Interoperability and privacy safeguards lag best practice. **Actions** • Institutionalize analytics teams. • Define data governance (for example, roles, approvals, and workflows), and strengthen privacy-by-design. • Maintain a data lake for the analytics team. **Data sharing:** Adopt governance rules that define access, responsibilities, and oversight to enable collaboration across government and with external partners.	***Capable and high-performing:*** **Integrated and data-driven** **Current situation** • Interoperable systems use tax data that are embedded in digital processes (for example, IDs, payments, and business registries). • Analytical models inform policy and operational decision-making. • Data safeguards are in place. **Actions** • Expand interoperability. • Adopt privacy-preserving data access and analytics (for example, synthetic data, differential privacy, and validation servers). • Formalize ethics and data reviews. **Data sharing:** Support data analytics as a public good (for example, data labs and trusted research environments), ensuring data security and reproducibility.

Source: Original table for this publication.

Note: MOUs = memoranda of understanding.

Note

1. Another example of leveraging microdata for cross-country insights is Bachas, Brockmeyer, Dom, et al. (2026), which harmonizes CIT returns in 16 countries to document that the largest firms in each country pay lower effective tax rates than medium-sized firms due to a higher take-up of tax incentives.

References

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. 2022. "Machine Learning and Phone Data Can Improve Targeting of Humanitarian Aid." *Nature* 603: 864–70.

Alfaro-Urena, Alonso, Isabela Manelici, and Jose P. Vasquez. 2021. "The Effects of Multinationals on Workers: Evidence from Costa Rican Microdata." Working Paper 285. Princeton University, Department of Economics, Center for Economic Policy Studies.

Almunia, Miguel, Jonas Hjort, Justine Knebelmann, and Lin Tian. 2024. "Strategic or Confused Firms? Evidence from 'Missing' Transactions in Uganda." *Review of Economics and Statistics* 106 (1): 256–65.

Andersen, Asger L., Kilian Huber, Niels Johannesen, Ludwig Straub, and Emil Toft Vestergaard. 2026. "Disaggregated Economic Accounts." *Quarterly Journal of Economics* 141 (2): 1005–1075.

Asad, Sher Afghan, Adnan Khan, Ben Olken, Mahvish Shaukat, and Jie Zhou. 2025. "Increasing Revenue Collection with Computer Vision: Experiments in Pakistan." Abdul Latif Jameel Poverty Action Lab. https://www.povertyactionlab.org/initiative-project/increasing-revenue-collection-computer-vision-experiments-pakistan.

Aslett, Joshua. 2024. "Understanding Artificial Intelligence in Tax and Customs Administration." *Technical Notes and Manuals* 2024 (6): 1.

Bachas, Pierre, Anne Brockmeyer, Roel Dom, and Camille Semelet. 2026. "Effective Tax Rates, Firm Size and the Global Minimum Tax." *Journal of Public Economics* 261: 105721.

Bachas, Pierre, Anne Brockmeyer, Alipio Ferreira, and Bassirou Sarr. 2025. "Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal." Policy Research Working Paper 11205. World Bank.

Bachas, Pierre, Anne Brockmeyer, Pablo Garriga, and Camille Semelet. 2025. "The Impact of COVID-19 on Formal Firms: Lessons from Administrative Tax Data." *Journal of Development Economics* 175: 103461.

Bachas, Pierre, Tatiana Flores, Gabriel Lara Ibarra, et al. 2025. "VAT Cashback Programs in Practice: TheCase of Devolve–ICMS in Rio Grande do Sul, Brazil." Prosperity Insight Series. World Bank.

Bachas, Pierre, and Anders Jensen. 2023. *Administrative Data and Methodologies for Tax Policy.* International Growth Centre.

Balboni, Clare, Johannes Boehm, and Mazhar Waseem. Forthcoming. "Firm Adaptation in Production Networks: Evidence from Extreme Weather Events in Pakistan." *American Economic Review*.

Barwahwala, Taha, Aprajit Mahajan, Shekhar Mittal, and Ofir Reich. 2024. "Is Model Accuracy Enough? A Field Evaluation of a Machine Learning Model to Catch Bogus Firms." NBER Working Paper 32705. National Bureau of Economic Research.

Benzarti, Youssef. 2020. "How Taxing Is Tax Filing? Using Revealed Preferences to Estimate Compliance Costs." *American Economic Journal: Economic Policy* 12 (4): 38–57.

Bergeron, Augustin, Arnaud Fournier, John Kabeya Kabeya, Gabriel Tourek, and Jonathan L. Weigel. 2023. *Using Machine Learning to Create a Property Tax Roll: Evidence from the City of Kananga, D.R. Congo.* ICTD Working Paper 176. Institute of Development Studies.

Bertanha, Marinho, Carolina Caetano, Hugo Jales, and Nathan Seegert. 2020. "Bunching Estimation Methods." In *Handbook of Labor, Human Resources and Population Economics,* edited by Klaus F. Zimmermann, 1–44. Springer International Publishing.

Best, Michael, Augustin Chaintreau, Joana Naritomi, and Dimitri Szerman. 2024. "When and Why Do Governments Pay More? Evidence from Brazil." Working Paper.

Bjarkefur, Kristoffer Gustav, Luiza Cardoso De Andrade, Benjamin B. Daniels, and Maria Ruth Jones. 2021. *Development Research in Practice: The DIME Analytics Data Handbook.* World Bank.

Bø, Erlend E., Joel Slemrod, and Thor O. Thoresen. 2015. "Taxes on the Internet: Deterrence Effects of Public Disclosure." *American Economic Journal: Economic Policy* 7 (1): 36–62.

Boning, William C., Nathaniel Hendren, Ben Sprung-Keyser, and Ellen Stuart. 2024. "A Welfare Analysis of Tax Audits Across the Income Distribution." *Quarterly Journal of Economics* 140 (1): 63–112.

Bowen, Claire McKay. 2024. "Government Data of the People, by the People, for the People: Navigating Citizen Privacy Concerns." *Journal of Economic Perspectives* 38 (2): 181–200.

Britto, Diogo G. C., Alexandre Fonseca, Paolo Pinotti, Breno Sampaio, and Lucas Warwar. 2025. "Intergenerational Mobility in the Land of Inequality." Working Paper.

Brockmeyer, Anne. 2019. *Working with Administrative Tax Data: A How-to-Get-Started Guide.* MTI Practice Note 7. World Bank.

Brockmeyer, Anne, François Gerard, Gabriel Ulyssea, et al. 2026. "Formal Labor Market Dynamics and Development." CEPR Discussion Paper DP21614. Centre for Economic Policy Research.

Brockmeyer, Anne, Giulia Mascagni, Vedanth Nair, Mazhar Waseem, and Miguel Almunia. 2024. "Does the Value-Added Tax Add Value? Lessons Using Administrative Data from a Diverse Set of Countries." *Journal of Economic Perspectives* 38 (1): 107–32.

Buda, Gergely, Stephen Hansen, Tomasa Rodrigo, Vasco M. Carvalho, Alvaro Ortiz, and José V. Rodríguez Mora. 2023. "National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption." Working Paper 2244. Cambridge Working Papers in Economics.

Burman, Leonard, Barry Johnson, Victoria L. Bryant, Graham MacDonald, and Robert McClelland. 2024. "Protecting Privacy and Expanding Access in a Modern Administrative Tax Data System." *National Tax Journal* 77 (4): 927–47.

Card, David, Raj Chetty, Martin S. Feldstein, and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." In *Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*. American Economic Association. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1888586.

Caspi, Aviv, Jacob Goldin, Daniel Reck, and Daniel E. Ho. 2024. "Optimal Tax Audits Using Predictions." Working Paper.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations.* Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2024. *A Practical Introduction to Regression Discontinuity Designs: Extensions.* Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Chacha, Peter Wankuru, Benard Kipyegon Kirui, and Verena Wiedemann. 2024. "Supply Chains in Times of Crisis: Evidence from Kenya's Production Network." *World Development* 173: 106363.

Clark, Julia, Georgina Marin, Oya Pinar Ardic Alper, and Guillermo Alfonso Galicia Rabadan. 2025. "Digital Public Infrastructure and Development: A World Bank Group Approach." Digital Transformation White Paper. World Bank.

Del Carmen, Giselle, Santiago Garriga, Wilman Ponce, and Thiago Scot. 2025. "Two Decades of Top Income Shares in Honduras." *Journal of Public Economics* 246: 105362.

Donovan, Kevin, Will Jianyu Lu, and Todd Schoellman. 2023. "Labor Market Dynamics and Development." *Quarterly Journal of Economics* 138 (4): 2287–325.

Grosh, Margaret, Phillippe Leite, Matthew Wai-Poi, and Emil Tesliuc, eds. 2022. *Revisiting Targeting in Social Assistance: A New Look at Old Dilemmas.* Human Development Perspectives. World Bank.

Hlatshwayo, Ayanda. 2025. "Introduction to the National Treasury Secure Data Facility." Unpublished presentation at the World Bank, March.

Junquera-Varela, Raúl Félix, Cristian Óliver Lucas-Mas, Ivan Krsul, Vladimir Calderón, and Paola Arce. 2022. "Digital Transformation of Tax and Customs Administrations." Equitable Growth, Finance and Institutions Insight. World Bank.

Kleven, Henrik Jacobsen. 2016. "Bunching." *Annual Review of Economics* 8: 435–64.

Knebelmann, Justine, Victor Pouliquen, and Bassirou Sarr. 2026. "Discretion versus Algorithms: Bureaucrats, Tax Equity, and Acceptability." IFS Working Paper 26/34. Institute for Fiscal Studies.

Lawson, Cina, Morlé Koudeka, Ana Lucía Cárdenas Martínez, Luis Iñaki Alberro Encinas, and Tina George Karippacheril. 2023. "Novissi Togo—Harnessing Artificial Intelligence to Deliver Shock-Responsive Social Protection." Social Protection and Jobs Discussion Paper 2306. World Bank.

Leite, David. 2024. "The Firm as Tax Shelter Micro Evidence and Aggregate Implications of Consumption Through the Firm." Working Paper 25. World Inequality Lab.

Li, Haonan, and Xuan Wang. 2025. "Firm Responses to Tax Audits: Regression Discontinuity Evidence from a Threshold-Based Audit Program." *Economic Journal* 136 (674): 560–601.

Lobel, Felipe, Thiago Scot, and Pedro Zúniga. 2024. "Corporate Taxation and Evasion Responses: Evidence from a Minimum Tax in Honduras." *American Economic Journal: Economic Policy* 16 (1): 482–517.

Londoño-Vélez, Juliana, and Dario Tortarolo. 2025. "Revealing 21% of GDP in Hidden Assets: Evidence from Argentina." *Journal of Public Economics* 250: 105471.

Oates, Wallace E., and Robert M. Schwab. 2015. "The Window Tax: A Case Study in Excess Burden." *Journal of Economic Perspectives* 29 (1): 163–80.

O'Hara, Amy. 2020. "Model Data Use Agreements: A Practical Guide." In *Handbook on Using Administrative Data for Research and Evidence-Based Policy,* edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Abdul Latif Jameel Poverty Action Lab. https://admindatahandbook.mit.edu/book/v1.0/dua.html.

Palomo, Theo, Davi Bhering, Thiago Scot, et al. 2025. *Tax Progressivity and Inequality in Brazil: Evidence from Integrated Administrative Data.* Report 9. EU Tax Observatory.

Pecenco, Matthew, Carlos Schmidt-Padilla, and Hamilton Taveras. 2026. "High Responsibility Work and Firm Creation: Experimental Evidence from the Dominican Republic." Working Paper.

Perez-Truglia, Ricardo. 2020. "The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment." *American Economic Review* 110 (4): 1019–54.

Pomeranz, Dina, and José Vila-Belda. 2019. "Taking State-Capacity Research to the Field: Insights from Collaborations with Tax Authorities." *Annual Review of Economics* 11: 755–81.

Reuters. 2024. "Stunning Photos Show Devastating Scale of Flooding in Brazil." May 10, 2024. https://www.reuters.com/pictures/stunning-photos-show-devastating-scale-flooding-brazil-2024-05-10/.

SEFAZ, RS (Secretaria da Fazenda do Estado de Rio Grande Do Sul). 2024. *Boletim Econômico Tributário: Edição No. 12.* SEFAZ, RS.

Slemrod, Joel. 2025. "Tax Privacy." *Journal of Economic Perspectives* 39 (1): 205–24.

Wani, Shahrukh. 2025. "From Data to Decisions: Improving Data Use in Governments." https://www.theigc.org/blogs/data-in-government-developing-countries.

Wiedemann, Verena, Benard K. Kirui, Vatsal Khandelwal, and Peter W. Chacha. 2024. "Spatial Inequality and Informality in Kenya's Firm Network." Policy Research Working Paper 10932. World Bank.

Wood, Alexandra, Micah Altman, Kobbi Nissim, and Salil Vadhan. 2022. "Designing Access with Differential Privacy." In *Handbook on Using Administrative Data for Research and Evidence-Based Policy,* edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Abdul Latif Jameel Poverty Action Lab. https://admindatahandbook.mit.edu/book/v1.0/diffpriv.html.

World Bank. 2021. *World Development Report 2021: Data for Better Lives.* World Bank.