ECONOMIC DEVELOPMENT & INSTITUTIONS

EDI WORKING PAPER SERIES

HOW TO TARGET ENFORCEMENT AT SCALE? EVIDENCE FROM TAX AUDITS IN SENEGAL

Pierre Bachas

World Bank & IFS

Anne Brockmeyer

IFS, UCL, World Bank & CEPR

Alipio Ferreira

TSE & IFS

Bassirou Sarr

EHESS

September 2021

Abstract

Developing economies are characterized by limited compliance with government regulation, such as taxation. Resources for enforcement are scarce, but the increasing availability of digitized data and data processing technologies have the potential to improve the targeting of enforcement. Levering an experiment at scale in Senegal, we compare the yield of tax audit cases selected by a risk-scoring algorithm to cases selected by tax inspectors based on a traditional discretionary procedure. Discretionary methods select larger firms on average and uncover equivalent evasion rates as the algorithm, thus outperforming it in terms of fines.

*Pierre Bachas: World Bank Research, pbachas@worldbank.org.; Anne Brockmeyer: Institute for Fiscal Studies, anne.brockmeyer@ifs.org.uk; Alipio Ferreira: Toulouse School of Economics, alipio.ferreira@tse-fr.eu; Bassirou Sarr: Ecole des Hautes Etudes en Sciences Sociales (EHESS). We thank the Senegal Tax Administration (Direction G_en_erale des Imp^ots et des Domaines, DGID) for an oustanding colllaboration. In particular, we are indebted to Bassirou Samba Niasse, Amadou Abdoulaye Badiane, Oumar Diop Diagne, Hady Dieye, Mor Fall, Serigne Mabaye Fall and Mathiam Thioub for helping us access the data, providing comments on the design of the study and facilitating its implementation. We thank Denis Cogneau, Laurent Corthay, Janet Jiang, Nicola Limodio, Jan Loeprick, Markus Kitzmuller and Dan Rogger for helpful comments and discussions. We are grateful to Oumy Thiandoum for excellent research assistance. We are also thankful to the Paris School of Economics and CEPREMAP for their support over the course of the research project. Financial support from UKAID from the UK government through the Economic Development and Institutions (EDI) Initiative and the Centre for Tax Analysis in Developing Countries (TaxDev) is gratefully acknowledged. The _ndings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its a_liated organizations, or those of the Executive Directors of the World Bank or the governments they represent, nor do they reflect the views of the Senegal Tax Administration.

About Economic Development & Institutions

1 Introduction

To ensure compliance with regulations, governments allocate scarce resources towards enforcement activities. Bureaucrats have traditionally held a large degree of autonomy on how to target enforcement interventions in low-income countries. This may be optimal if bureaucrats hold valuable experience and soft information, but this information advantage likely erodes as the spread of new technologies makes hard information readily available, allowing data driven targeting of enforcement. Although digitized data is increasingly available, many administrations do not take advantage of it systematically, but rather on an ad-hoc basis. Can the systematic use of data in low capacity settings improve enforcement activities?

In this paper, we analyse the tax revenue yields and deterrence effects from the at scale implementation of a risk-score based selection of firm audits, in Senegal, and compare it to the discretionary selection method, in use. To understand the context, all audit cases in Senegal were selected with a discretionary procedure until 2017. At the beginning of each year, the different tax units select a set of full (in-person) audits to be conducted in teams, and each inspector selects desk audits, conducted individually.

Working with the tax administration, the research team combined a large dataset of self-reported tax declarations across different taxes with third-party reported information, from customs, procurement contracts and transacting partners. This dataset was then used to construct a compliance risk profile, based on discrepancies across tax declarations and third-party data, and outlier behavior. Firms were then ranked within their peer groups based on their risk-scores, and the highest risk scores were assigned to the audit program.

Starting in 2018, the risk-score selection complemented the discretionary selection. Each tax unit selected half of the cases planned for the *full (in-person) audit* program , and the remaining half was assigned by the risk-score. Moreover, each inspector selected 45% of her *desk audit* program, 45% came from the risk-score, and the remaining 10% were selected randomly. Desk audits across the three selection methods were cross-randomized into an information treatment: a subset of cases received information on the largest compliance risks detected by the risk-score, and detailed data from third parties regarding that taxpayer. The information treatment thus pinpoints the specific risks and facilitates data access, thus potentially easing inspectors' work. The experiment, hypothesis and specifications used were submitted to the AEA registry.

How does the audit selection, completion and tax evasion uncovered varies across selection methods? What role, did information on compliance risks play? We find four key results. First, we observe that the average size of selected firms varies substantially across selection methods. On average, firms chosen by tax inspectors report 50% more revenues than firms selected by the risk-score, and higher profits. Second, we find that cases were started less often when selected by risk-scoring. Moreover, conditional on starting an audit, cases were finished less often for the risk-scoring cases. Third, conditional on an audit starting, the tax evasion rate uncovered is similar across the discretionary and risk-based selection methods. However, cases recovered larger amounts for discretionary selection, since they selected larger firms. Fourth, the information treatment yields no significant improvement in terms of audit yield, but it increased the probability of the audit being carried out to the end.

To interpret these null results, it is important to note that the intervention was conducted at scale. Selected firms represented 24% of corporate tax revenue of the tax centers in the experiment[1]), and implemented directly by the audit planning and intelligence division of the tax administration. Two design features should be further highlighted. First, the risk-score applied best international practices and was constructed after ample consultation; it however did not rely on fancy machine-learning tools and fine-tuned parametrization. Indeed, the tax administration decided that the risk-score should be guided by transparency, such that underlying compliance risks could be understood by tax inspectors, and explained to taxpayers, and because the available data on historical audit results was sparse, which limited the scope for model training. Thus, although the risk-scoring method can be improved over time, it represents an accurate counterfactual for low income countries considering introducing transparent risk-based selection of audits at scale. Second, it is likely that tax inspectors efforts were asymmetric across selection methods despite the pressure from their hierarchy to devote equal efforts; inspectors receive a share of the audits fine collected and have career incentives to perform but they also hold office for life, and enjoy significant autonomy. Changing the monetary incentives is not allowed legally. These set of constraints are likely to bind in many countries, especially in West Africa which often looks at Senegal for administrative innovations.

To our knowledge, no prior study has rigorously explored the implications of audit selection mechanism in an environment with widespread evasion. Recognizing the administrative

[1] The total amount of corporate tax liability (VAT and CIT) over the years 2015-2018 for the tax centers used was around 315 billion FCFA, and the selected firms in the 2019 program accounted for 75 billion FCFA.

constraints of developing countries, recent work has shown that optimal tax policy might differ from textbook models, which often assume perfect enforcement (Best et al. 2015). The importance of third-party information to detect and deter evasion is emphasized in a growing literature (Pomeranz 2015, Kleven et al. 2011, Kleven et al. 2016 Naritomi 2019).[2] However, despite growing availability of hard data, the optimal audit selection in low income countries may differ from the standard prescription of machine-based selection, if inspectors' private information is more valuable than the limited third-party information available, or if this hard information is hard to access and analyse. Our intervention thus varies the third-party trails provided and the ease of access to such information for tax auditors. As such it provides a test of the value of changing fiscal capacity, which has been argued to be a key determinant of governments' effectiveness in tax collection (Besley and Persson 2014 , Jensen 2016 ,Xu 2019).

Second, we extend the literature on audit selection, which focused on the United States. Murray 1995 and Alm et al. 2004 have examined the effect of sales tax audit selection on the compliance behaviour of firms in different US states. Both papers model audit selection as a strategic interaction in which the tax authority first signals the probability of audits and announces penalties. In a second step, taxpayers report their tax liabilities. Based on these reports and its resources, the tax authority decides which firms to select. Both papers use a two-stage selection model to test the theory and find suggestive evidence that Tennessee and New Mexico use informal selection rules. Yet both studies suffer from selection bias, as the authors do not have full information about the rules used to select the audits. Compared to these studies, Senegal's tax authority has not yet implemented a fully-fledged risk-scoring audit strategy: this provides a unique opportunity to rigorously explore how audit selection methods impact tax collection. The IMF and World Bank have long advocated the use of risk-based algorithms for audit selection, but have not published impact evaluations. In the context of environmental regulation in India, Duflo et al. 2018 compare the audit reports of randomly selected firms with the reports of audits selected by the enforcement agency. They find that the randomly selected audits perform worse than the discretionary audits, suggesting that the value of discretion compensates for the risks of corruption or human errors. In our setting, we compare the discretionary method to random audits and to a risk-score algorithm aimed at using all available information in a systematic way to select firms. Finally, we hope to contribute to a nascent literature (Gerardino et al. 2017) on the cost incurred by firms and their medium-term outcomes following tax audits.

[2] Recent papers study firms' behaviour as they get exposed to new third-party information trails and show that taxpayers substitute evasion to less verifiable margins (Carrillo et al. 2017, Slemrod et al. 2015).

This paper provides evidence on taxpayer audit selection mechanisms as an enforcement tool to augment fiscal capacity. Central to the literature on public finance and taxation, the notion of fiscal capacity tends to include many concepts that current research is yet to disentangle. In a series of publications(Besley and Persson 2009, Besley and Persson 2010,Besley and Persson 2013, Besley and Persson 2014), Besley and Persson argue that, in the past, in trying to shed light on the concept of fiscal capacity, the public finance literature primarily focused on incentive constraints, information asymmetries and political institutions (Acemoglu et al. 2005). Yet, in the context of developing countries, fiscal capacity tends to be inevitably linked to administrative capacity. Thus, it is worthwhile providing a definition, which allows us to discuss audit capacity as part of a broader set of tax administration challenges. So, what is fiscal capacity?

Fiscal capacity is a set of strategic and forward-looking decisions aimed at increasing investments in institutions, particularly public administrations, to bolster tax revenue. Thus, consistent with the exposition in Besley and Persson 2013, fiscal capacity is a product of investments in state structures such as monitoring, administration and compliance, which, then, determine the level of revenue a state could mobilize, given the parameters of its tax system and its enforcement powers. Fiscal capacity, thus, encompasses both rate and non-rate tax enforcement instruments [3]. Within the latter category, withholding and remittance responsibilities, human resources investment, organizational structure and *audit policies* are all elements of fiscal capacity. Hence, to increase its fiscal capacity, a state could enhance its tax audit capacity. In this regard, this paper makes important contributions by introducing a tax audit programming reform and documenting bureaucrats' work in Senegal through a tight collaboration with the tax authority over a three-year period.

Ensuring that firms and physical persons comply with laws and regulations is a fundamental challenge for state institutions in developing countries (Acemoglu et al. 2005). Bureaucratic enforcement is often a two-pronged allocation problem with the necessity to i) first decide on the human, financial and technical resources dedicated to bodies in charge of enforcement (Xu 2019, Bertrand et al. 2018; Finan et al. 2017) and ii) second decide on which firms to scrutinize. For the audit functions of many revenue and expenditure institutions such as tax and customs authorities, procurement regulatory bodies or budget control de-

[3] Recall that rate instruments are mainly measures that seek to bolster revenue through legislative adjustments on tax rates and bases. Non rate-instruments refer to administrative and structural changes that seek to improve revenue performance.

partments, optimal firm selection policies are essential in enforcing compliance. Thus, given weak institutional environments, corruption and interference which can spare some firms from scrutiny, how much human discretion should there be in selection processes for compliance checks? Institutions with enhanced capacity to exploit modern information systems to identify risks could use analytics for rule-based selection. Nevertheless, with limited data, strong information asymmetries, can the state rely solely on bureaucrats' expert judgments? Allowing discretion in audit selection could leave room for collusion with those subject to audits or limit selection to the set of information that human beings can reasonably consume while, today, data is often complex and rich. Despite their centrality to revenue mobilization, state expenditure and economic development in general, evidence on the impact of selection methods on compliance outcomes is not well documented.

Tax law enforcement through well-targeted audits can boost revenue and contribute to the reduction of inequalities because with weak audit capacity, owners of capital are more likely to escape taxes. Slemrod et al. 2001, for instance, claims that taxpayers with higher earnings have access to advisors, preparers and view the audit process as a bargaining one. The deterrence effect of audits through optimal targeting can also increase aggregate efficiency by reducing distortions that may arise because of unequal audit probabilities between firms that are similar in compliance behavior. When evasion is widespread and audits selection excludes firms for non-objective reasons, the state ends up selecting winners who display higher productivity because of distortions introduced by the audit process (Hsieh and Klenow 2009, Restuccia and Rogerson 2008, Monitor 2017). Still on the revenue side, with limited personnel, scanners or dedicated areas in ports for inspections, customs administrations often rely on risk scoring to select containers subject to inspection at major ports of entry. On the spending side, the state in developing countries is the largest single procurer of goods and services from firms. Administrations in charge of state procurement need to ensure that suppliers meet the specifications of goods and services they contracted to provide to the state. Thus, in the budget control process, bureaucrats conduct *ex ante* audits of the quantity and quality of goods of all procurement contracts before making payments. This process is often lengthy, reduces firms' economic efficiency. Poor targeting of expenditure audits also reduces overall procurement and expenditure efficiency. When they expect payment delays, firms bid at a significant mark-up when they sell to the state.

Khwaja et al. 2011 review tax audit selection practices across countries at different income levels. Table 4 summarizes policies in selected countries. Selection methods can be classified into three broad categories, namely random selection programs, decentralized discretionary proposals by tax inspectors who are familiar with the tax returns and the past behavior of firms and risk scoring or machine learning techniques, which rely on flags based on deviations noticed in information reports as well as the taxpayers' compliance history. Most countries combine two of the three listed methods while only the United Kingdom uses all three. For the U.K., 55% of all cases are based on discretionary selection while 35% and 10% of cases are respectively selected via a risk-scoring technique and a simple random sample. This approach is closest to the policy reform we introduce in Senegal. In other sub-Saharan African countries, Kenya uses a risk for all large taxpayers and discretionary selection for all others. Tanzania and Lesotho constitute examples on the extreme, respectively relying only on risk-scoring and random selection to audit all taxpayers.

However, the impact of such approaches to audit selection of taxpayers is yet to be documented in the public finance literature on developing countries. For the United States, Troiano 2017 reveals that Audit Exchange Information Agreements on income tax audit plans and techniques between states and the federal government raised state revenue collections by about 15%. Crosschecks on different sets of information reports, third-party data and analytics are touted as crucial dimensions of state capacity in developing countries as computing technologies have and will expend information sets the state can use to scrutinize firm activity. In this respect, our paper fills an important gap in a critical dimension of fiscal capacity, reflected in its capacity to audit taxpayers.

2 Institutional Setting: Senegal's Revenue Administration

Tax revenue represented on average 16.71 % of GDP in Senegal between 2013 and 2019. These revenue collection levels are below the West African Economic and Monetary Union (WAEMU) target of 20%, and fall short of goals set in Senegal's own medium term expenditure strategy. The tax gap analysis indicates that 23% of the theoretical VAT revenue is not collected (a shortfall of 2% of GDP) and that close to 63% of theoretical receipts from income taxes are missing (approximately 7% of GDP).

Similar to other developing countries (Slemrod et al. 2001), most taxes in Senegal are remitted by 1911 large and . In particular, firms remit the Value Added Tax (VAT) and income taxes (Corporate income tax, personal income tax and dividend withholding taxes), which account respectively for 36% and 29% percentage of total tax revenue in 2019. Firms also withhold income taxes on wages of their employees (Pay-as-You-Earn), which is often the only source of reporting on salaried income, given the incompleteness of self-reported personal income taxes. Other important sources are customs duties (15%) and specific taxes on petroleum consumption, which are not covered by our study.

The Direction Générale des Impôts et des Domaines (DGID) is the administrative body in charge of domestic tax collection and enforcement (IRS), and reports to the Ministry of Finance. Figure 1 displays DGID's organizational chart. The large taxpayer directorate oversees firms whose turnover is greater than equal to 3 billion CFA francs and has four units, which are specialized by economic sectors.[4] The medium taxpayer directorate oversees firms with less than with turnover between 100 million CFA francs and 3 billion CFA francs, and has two units. A third unit is in charge of the regulated liberal professions such as lawyers, notaries and medical practitioners. Taxpayers which do not belong to these seven strategic units are assigned to one of 19 regional tax offices.

To enforce taxes, the Senegal's Tax Code provides two main types of audit procedures: desk audits and full audits.[5]. These audits differ in the information that can be requested, the type of contact with the taxpayer, and the number of tax inspectors involved

One inspector within the tax authority's premises, using only tax returns and , conducts them and other data submitted to the tax authority by any other party. Unless, documentation is missing in a tax return, inspectors are not allowed to communicate with taxpayers. When data is missing, auditors can issue an information request notice. Finally, we note that, although there is no stated upper bound on the number of audits completed by staff, inspectors have quarterly completion targets for desk audits.

Full and and surprise audits are conducted by a team of inspectors at the taxpayer's premises. Full audits are announced at least five days before the audit starting date with an information request notice to the taxpayer. Such information includes any documentation, contracts or payment proofs related to the taxpayer's activities. As a general rule, full audits cannot last longer than 12 months and for firms with turnover less than 1 billion CFA francs (about 2 million USD), it cannot last longer than 4 months with the exception of cases with highly suspicious activity or when there is a delay in the transmittal of requested information to auditors . Surprise audits are similar to full audits, except that, as their name indicates,

[4] Unit 1 is in charge of the mining and energy sectors. Unit 2 deals with financial services and the telecommunications industry. Unit 3 covers real estate and firms. Unit 4 is a generalist one with broad competence covering all other sectors.

[5] there are also surprise audits which can take place either based on information that DGID receives either internally or from whistle-blowers.

they are unannounced.

Figure 2 illustrates the steps in the audit process. Upon issuing an audit notice and reviewing the case, auditors can issue an initial notice to the taxpayer with an indication of discrepancies per tax, as well as assessed penalties. They can also request additional information from the taxpayer. Upon receiving the initial notice, taxpayers have up to 30 days to provide a response confirming or contesting the inspector's findings. Without a response within 30 days, the taxpayer is deemed to agree with the initial findings. Once the taxpayer responds, the auditor examines its response and prepares a written final notice (henceforth called confirmation) with amounts due and penalties within 60 days. This steps marks the end of the audit process. The inspector creates a revenue order for the collection unit which begins its own process with payments required within 10 business days, unless a moratorium or installations are granted. Taxpayers can lodge an administrative appeal with the Minister of Finance or a judicial one in court. Neither appeal suspends the collection process or taxpayers' payment obligations.

If there is no discrepancy in full audits, auditors issue a notice which indicates that all declarations are correct. However, in desk audits which are the focus of this paper, no information is sent to the taxpayer who, in general, is not aware of the process until an initial discrepancy notice is issued.

3 Data

Our study draws on three sets of administrative data sources and two surveys. The administrative data contains self-assessment declarations filed by taxpayers, third-party data used to cross-check the tax declarations, and data generated as a result of the enforcement process. All of these data are accessible to both the research team and to DGID, and all observations carry a unique taxpayer ID which allows us to match across datasets. We discuss details of the matching process and match rates in Appendix C. We further match the administrative data with a taxpayer survey and a tax inspector survey. These data are not available to the senegalese tax authorities. We now discuss each dataset in turn.

Tax Declarations. Table 6, Panel A, provides an overview of the available tax declarations. The CIT is paid annually, at a rate of 30% profits or a 0.5% of turnover, whichever is larger. The CIT data covers about 4 thousand firms per years, and the available data series covers years 2014-2019. The VAT is a paid at a monthly basis, at a standard rate of

18% and a reduced rate of 10% for tourism businesses and hotels. The VAT data contains around 8 thousand firms every year and covers the period 2014-2018. There are many more firms declaring VAT than CIT, because self-employed individuals and unincorporated firms file VAT but not CIT. A small number of financial sector firms pay the financial services tax instead of the VAT, also at a rate of 18%. They are mostly concentrated in the Large Taxpayers Unit, which is not included in our analysis. We further match these data with monthly Pay-As-You-Earn data, which refers to the withheld progressive personal income tax, for all formal employees with an employment contract. This allows us to calculate the number of employees and the aggregate wage bill for each firm. Small firms with a yearly turnover of less than 50 million CFA Francs (about 100,000 USD) are eligible for a simplified tax (*Contribution globale unique*, CGU), which replaces all other taxes. The CGU is levied on turnover, at rates varying from 1% to 8%, where rates vary across sectors and increase in turnover.

Third-Party Data. Table Table 6, Panel B, describes the third-party data available to cross-check taxpayers' self-assessment declarations. Through inter-agency data sharing agreements, the tax agency regularly obtains import and export data from customs and procurement data on firms' sales to other state institutions. Both are transaction-level datasets which we aggregate at the firm-year level to merge with the tax data. As the last two columns in Table 6 indicate, a non-negligible share of firms captured in the third-party data are non-filers in the sense that they cannot be found in any of the tax datasets for the corresponding year. The share of taxpayers for whom third-party data is available hovers around 28%, with the share increasing over time and in firm size. Starting in 2020, large taxpayers were required to electronically file VAT annexes, listing transaction amounts and transaction partner tax ID for all sales and purchases.

Enforcement process and results data. We collect audit results data for fiscal years 2018 and 2019 by digitizing the content of the key audit-related communications between the tax agency and the taxpayer: audit announcement, notification, confirmation and payment request. Importantly, the audit data covers all audits, including those which inspectors initiated independently of the audit program set in the beginning of each year. The main variables include audit case identification (IDs for the firm and the inspector(s) conducting the audit), years and taxes verified in the audit, infractions detected, evaded amounts, applicable penalties, and the issuance dates of all notices, from which can calculate the audit length. We calculate the detected evasion rate by dividing the amount evaded in the final notice (or the initial notice, if the final notice is not available) by the taxpayer's turnover.

We also ask inspectors to report qualitative information on each audit case in an excel file which inspectors submit quarterly. These qualitative information cover the reasons for abandoning an audit case for those cases which did not lead to a notification, the perceived difficulty of the audit, and indicators for various dimensions of difficulty, for instance whether the taxpayer was uncooperative, the business activities were complex, or information was unavailable.

In their quarterly reports, inspectors also record all audit outcomes that we observe in the digitized data. Note that the current version of the paper temporarily uses the quarterly reports to analyze the results of our interview, but we ultimately plan to rely primarily on the digitized data which is of higher quality.

Tax Inspector Survey. Prior to our intervention, we conducted a detailed survey among all participating tax inspectors, capturing information about their demographics, employment history, perceptions of the audit function, methods for audit selection, and use of different sources of information. The survey data contain 97 inspectors, which covers 73% of the 132 inspectors involved in conducting audits in 2018 and 2019. The discrepancy is due to a partial re-assignment of inspectors across teams within the tax agency after the beginning of our intervention.

4 Audit Selection and Experimental Design

4.1 Discretionary Selection

Up until fiscal year 2018, all audit cases were selected with a discretionary procedure. At the beginning of the year, the Director general publishes a note requesting personnel in each tax unit to propose firms for the annual audit program. The unit suggests a set of full audits to be conducted in teams of at least two inspectors, and each inspector suggests a set of desk audits, to be conducted individually.[6]

Tax inspectors use a standardized form to motivate their full audit choices. This form provides basic information on the identity of the selected firm, past audit outcomes as well

[6] Since desk audits are selected individually, it could happen that two inspectors select the same taxpayer; in practice this is rare as inspectors specialize by economic sectors and/or geographical areas. When this happens the manager presumably rules which inspector is in charge of the case.

as a summary of relevant information from non-prescribed tax turnover and profit margin. Once the tax unit's manager approves the form, a selection committee in the Director general's office finalizes the list of firms for the full audit program. Based on interviews with members of this committee, most cases are accepted, although in rare instances the committee requests additional information or rejects a proposal. The selection committee could also add firms to the list based for example on denunciations. The committee then returns the names of approved audits to tax units.

Desk audits are also selected at the beginning of the fiscal year by individual inspectors, and are signed off by the director. Based on discussions with DGID's staff, there are no guidelines on how one should select desk audits. Each staff member follows her own rule, which ranges from randomly picking files to selecting cases based on private information, or from their own data cross-checks.

4.2 Risk-Score Method

Government agencies increasingly rely on rule-based methods to select firms suspected of regulatory non-compliance. For tax administrations, risk analyses are conducted through information cross-checks between tax returns and third-party reports, from other taxpayers, other government agencies, and large private actors. Following this trend, we assisted the Senegalese tax administration (DGID) to design a tool which assesses firms' tax evasion risks. Starting in 2017, the team held consultations with DGID leadership and former tax inspectors to map the compliance risks of Senegalese firms and to exploit all available data sources to assess this risk. Moreover, we discussed with experts in the field of taxation and risk management, who worked on tax evasion risk assessment in middle-income countries. With these inputs, we designed a risk-scoring tool, following best international practice.[7]

Although the use of advanced machine-learning tools for prediction has exploded in economics, it was decided with DGID, that the risk-score would be guided by simple indicators which should logically predict evasion risk. The simplicity of the design is motivated by several factors. First, the need for transparency, such that underlying compliance risks could be understood by tax inspectors, and explained to taxpayers when required. Second, the available data on historical audit results was sparse and not digitized, which limited the scope of our model calibration and model selection exercises. Finally, all cases concluded by

[7] We designed the risk-score following best practices, drawing on work by the World Bank (tax administration projects in Pakistan and Turkey), SKAT in Denmark, and the IMF's recommendations to DGID.

2017 had been selected in a discretionary manner, thus leading to possible spurious relations.[8]

Thus, the risk-scoring tool should be viewed as a transparent best-practice risk assessment, which takes into account the administrative capacity of DGID, rather than a fined-tuned optimized algorithm. We note that the constraints faced by DGID are likely to bind in many low-income countries, especially in West Africa which often looks at Senegal for administrative innovations.

Table 1 summarizes the seven key steps in the design of the risk-score. Step (1) corresponds to the construction of a database covering all tax declarations across years and merged with third-party reported sources, discussed in section 3. Steps (2) and (3) determine specific risk indicators, based respectively on discrepancies across data sources, and on behavioral outliers, examples of each cases are discussed below. Step (4) defines the peer-group comparison clusters which regroup firms by economic activity and size or geographical zones, depending on the organization of each tax center. Step (5) assigns a numerical value to each risk indicator, depending on the size of the deviation or anomaly (higher scores when larger discrepancies), while step (6) assigns weights to each indicator reflecting beliefs about their relative importance. Finally, step (7) aggregates the weighted indicators in each of the past four fiscal year, and then sums up the yearly scores to form a total risk score.

As hinted previously, the risk-score relies on two types of risk indicators: discrepancies and anomalies. Discrepancies flag taxpayers whose self-reported information according to their tax returns differs from information from third parties,including customs, state procurement and transaction network. For instance, a discrepancy indicator is logged when a taxpayer's reported turnover is lower than a lower bound third party cost estimate, which sums up its imports, wage bill and purchases from suppliers. The larger the (normalized) deviation the higher the score. Anomalies use sector benchmarking to flag firms with unusual behavior relative to their peers. An example is a firm with abnormally low profit margin compared to its peers. In total 4 discrepancy indicators and 6 anomaly indicators are used to construct the risk-score. Discrepancies are over-weighted compared to anomalies to reflect the higher confidence that discrepancies reflect non-compliance, while anomalies might only reflect temporary economic problems or poor management.

[8] Despite these limitations we attempted to check whether risk indicators predict the outcomes from previous DGID audit results, using machine learning techniques. This exercise was inconclusive due to the small number of observations and noise in the historical audit return data.

Table 1: Steps of risk-score design

Step	Description
(1) Prepare database	The tax declarations of each taxpayer are merged across type of taxes (VAT, CIT, Payroll) and across years. Data from third parties is then added (customs, procurement, transaction network).
(2) Choose indicators: discrepancies	Discrepancies are situations in which a self-reported tax liability can be considered as misreported or incomplete, by cross checking several data sources together.
(3) Choose indicators: anomalies	Anomalies correspond to abnormal reporting behavior, compared to peers. Anomalies suggest that firms should be monitored, but do not indicate tax evasion behavior with certainty.
(4) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons are done within clusters
(5) Assign values to indicators	The magnitude of the inconsistency is used to assign a value, ranging from one to ten (using deciles). For anomalies firms within the top decile of a particular indicator receive a value of one.
(6) Assign weights to indicators	Weights are assigned to each indicator reflecting beliefs about their relative importance.
(7) Aggregate indicators and years	The weighted risk indicators are first aggregated across indicators in each year. Then the yearly scores are summed up to form a total risk score covering the past four years of tax declarations. More recent years are slightly over-weighted.

4.3 Study Design

Starting in fiscal year 2018, DGID started using a mix of discretionary selection, risk-score selection and fully random selection. Figure 3 illustrates the timeline of the design and case selection. The selection of the audit program proceeded in three main steps. First, inspectors use their discretion to select cases: tax centers select cases for full audits, and each individual inspector selects cases for desk audits, which they then submitted to the central planning committee. The exact number of cases varies by tax center as displayed in Table 3.

Second, we assign risk scores to each taxpayer and rank taxpayers based on these scores within each tax center. The highest risk scores are assigned to full audits, as to match each tax center's list: that is for each case selected by inspectors for a full audit we assign a risk-score selected case, within the same center and within the same cluster (depending on centers: sector specific or geographical area specific). If a case has a high risk score but has already been selected by DGID, we assign the next highest score not yet assigned, as to obtain the same number of inspector and risk-score selected cases. The procedure for desk audits is similar: it excludes all cases already selected for full audits and keeps on moving down the risk-score list to match each discretionary selected case. One difference for desk audits, is that an extra 20% of cases selected at random are added to the audit list. Thus to summarize, the final list of full audits is composed of 50% of inspector selected cases and 50% of risk-score selected cases, and that of desk audits is composed of 40% of inspector selected cases, 40% of risk-score selected cases (those with lower scores than full audits) and 20% of cases selected randomly.

Third, these lists with the discretionary and risk-score selected taxpayers are sent to the planning committee for review. Almost all cases were approved, with the exception of firms that had already received a full audit last year, and a few state-owned companies. These approved list, are then sent to the tax centers and individual inspectors, with a sequence of alternating cases by selection method[9] An internal memo from the Director General's office informs them of the new environment for audit programming which is prepared in collaboration with researches and urges them to follow guidelines in completing cases at a prescribed order and to rigorously report audit results.[10] The files sent to inspectors mentions the names of the firm, their tax ID, if it was selected through the "tax inspector proposal" or the "new method" designed by external researchers (which contains algorithm and randomly selected cases). This is accompanied by a note specifying the risk indicators used and the logic of risk-based selection and a workshop organized by the intelligence unit of DGID.

Finally, for desk audits, we cross-randomize two types of information treatments to the three selection methods. Each audit case thus belongs to i) *Control*: no additional information on the case, ii) *Treatment 1*: information on top three risk indicators detected by the risk-score (if any) and iii) *Treatment 2*: information on top three risk indicators detected by the risk-score plus detailed taxpayer data including from third parties organized in a spreadsheet. Treatment 1 thus pinpoints the specific risks detected by the risk-score, potentially easing each inspector's work. Treatment 2 adds the data used to create these indicator flags: this could be important as not all tax data is digitized and third-party sources can be costly to access, given the lack of automatized exchange of information between DGID and

[9] We randomly ordered each tax inspector's list of cases. This ensures that the order of audits is uncorrelated with audit quality. However, we were unable to ensure discipline in following the designated sequence for the workload. For instance, inspectors could choose to prioritize cases they select themselves and which they believe could leave to higher yield. Nonetheless, the Director General signed a guideline urging staff to follow the sequence set in their assignments.

[10] This complements presentations by the intelligence unit of DGID and the research team to each center.

other government agencies. Thus, by providing the data in a spreadsheet, we ease the efforts that they would have used to work on a case, compared to control cases or treatment 1 cases.

4.4 Comparing Firms Across Audit Selection Methods

The risk-score algorithm selects firms that have a higher risk score. As explained above, the risk score is a weighted aggregation of several indicators constructed for each firm, in which the firm is compared to itself (inconsistency indicators) and to its peers (anomaly indicators). We also weighted the risk score by the log of the mean turnover (over time) of each firm, to increase the risk score of firms that are larger and also present large inconsistencies or anomalies indicators. Firms with largest risk score within each tax center were included in the audits program of 2019.

The risk-score algorithm results in a selection of firms that is markedly different from the discretionary method used by the tax inspectors. Table 7 summarizes in pre-audit firm characteristics across different selection methods: randomly selected firms, risk-score selected firms and tax authority selected firms. There were only 42 firms selected both by the tax authority and by the algorithm selection. In the comparisons of table 7, we disregard these firms, and we include a dummy to control for these cases in the regressions tables.

The table shows that the randomly selected firms present similar averages as the population of firms with respect to their declared turnover, profits, profit rates, and tax liability over the period 2015-2018. The mean differences are large with respect to the values in tax declarations of the year 2018, but the differences are not statistically distinguishable from 0 at the 95% confidence level. This suggests that even though the number of randomly selected firms was low (around 10% of the audits program), it is a somewhat representative group of the firms registered at the tax centers under analysis.

Firms selected by the algorithm and by DGID, on the other hand, are very different from the population's average (or the randomly selected firms'). Selected firms in both cases are larger in terms of their declared turnover, profits and tax liability. Moreover, as can be seen in table 7, DGID selected firms with substantially larger declared turnover than the risk score algorithm (53% larger on average), larger profits (difference is not significant), and larger profit rates. The firms in the two selection methods present similar amount of tax liability. The risk score is substantially larger for the algorithm selected firms. This is unsurprising, since the risk-score selection explicitly picks the firms with largest values of this variable. The key difference that stands out is indeed the mean declared turnover by firms. Inspectors have a clear preference for firms with large declared sales, even though this typically means selecting firms that are already paying high levels of taxes. The algorithm, on the other hand, adds risk score to firms that present abnormally low levels of tax or turnover declarations. Even though the algorithm in the end weighs the risk score by the declared turnover, the final selection results in firms with average self-declared size which are much smaller under the algorithm selection than under DGID's selection.

5 Results

This section presents our main empirical results. Note that the current analysis relies on audit results data reported by tax inspectors, and may change once we can use the data digitized from paper records, which we consider is of higher quality.

We focus on three outcomes of interest to understand the impact of the new selection method: the probability of the audit being opened, the audit yield (measured as the initial notice sent to the taxpayer) and the detected evasion rate (measure as the audit return divided by the firm's historical mean turnover). In ongoing work, we examine impacts on other outcomes, as we mention briefly below.

5.1 Empirical Specification

To examine the effect of the selection method on audit implementation and results, the simplest model we estimate is:

$$\begin{aligned} y_{io} =& \beta_0 + \beta_1 Algorithm_i + \beta_2 FullAudit_i + \beta_4 Random_i + \beta_3 Overlap_i \\ &+ \sum_{o=1}^{O} \gamma_o 1(tax\_office = o) + \varepsilon_i \end{aligned} \tag{1}$$

where y_i is the outcome of an audit for firm i, *algorithm* indicates cases selected by the algorithm, *FullAudit* indicates full audits, *Random* indicates randomly selected cases (which applies only for desk audits), *Overlap* indicates cases selected by both the algorithm and by inspectors and the γ_o series capture tax office fixed effects. We estimate this model using OLS.

In most specifications we also control for *Replacement* cases, i.e. additional marginal cases that were included in the algorithm list as potential replacement for cases that would be taken out due to political concerns. The inclusion of these safety cases meant that the algorithm-selection list was slightly longer than the inspector list. Inspectors were instructed to use these cases only as replacements, but in practice the replacement cases were often considered as part of the main list. Controlling for these cases or dropping them from the analysis does not substantially alter our results.

In additional specifications focused on desk audits, we add inspector fixed effects and dummies to capture whether inspectors were provided with the risk flags or with the risk flags and micro data for the case.

5.2 Effect of selection method on audit implementation

We start by analyzing whether audits were implemented as programmed. When deciding whether or not to open an audit case, and when to do so, inspectors know whether the case was selected by DGID or by the algorithm. If DGID-selected cases are easier to conduct or have a higher expected return, inspectors may be reluctant to open algorithm-selected cases. This is especially true in offices where inspectors are given the freedom to open cases which were not in the set audit program for the year.

Table 5 examines this hypothesis. Each column depicts results of linear probability models in which the outcome is a dummy variable which takes value 1 if the case has been opened, and 0 otherwise. A case is considered opened if an audit announcement or a notification has been sent to the taxpayer, regardless of whether the case has been concluded, was closed without notification, or is still ongoing. On average, algorithm-selected audits are 10-14% less likely to be opened. This is consistent with the fact that algorithm-selected firms are smaller, and the fact that larger firms are significantly more likely to be audited conditional on selection. The latter fact is true for both algorithm and DGID-selected cases, but more pronounced for DGID-selected cases (Figure 4). Turnover is the strongest predictor for audit implementation conditional on selection.

Both firm size and the selection method have a larger impact on the implementation of full audits compared to desk audits, possibly because inspectors invest more time in full audits and are hence more mindful of their return (Table 5, columns 5 and 6). Whether or not inspectors are provided with information and/or risk flags on the audit case does not have a significant effect on whether the case is opened (columns 3 and 4).

The results are similar across tax centers, with the notable exception of the smaller one of the two medium taxpayer offices, where algorithm-selected cases are slightly more likely to be conducted, at least for full audits (??).

The following tables show the relationship between the selection method and the outcomes (probability of audit being started, log of initial notification and evasion rate). The tables show that the execution of the audits was significantly weaker for algorithm selected cases than for DGID selected cases. Column 1 of table 5 shows that on average, algorithm cases on the audits program were 14.7% less likely to be started than DGID cases. Random cases, which are lumped together with algorithm cases in this regression, are not significantly different than algorithm cases. The inspectors indeed knew which cases were not chosen by them, but they could not distinguish between an algorithm selected case and a randomly selected case.

One reason why inspectors preferred their own cases is related to the firm's size: the algorithm selected firms which were on average smaller than the firms selected by DGID, as measured by the firm's mean declared turnover over the previous four years. Inspectors have a marked preference for larger firms: they select larger firms than the average of their tax center, they start more often audits of larger firms among those selected, and they finish more often the audits of larger firms among those that they start.

However, even controlling for firm size, it is clear that the probability of an algorithm-selected firm having an audit started is lower, on average, for almost all levels of firm size.

5.3 Audit Yield

As inspectors were more motivated to conduct inspector-selected audit cases, we might expect these cases to have a higher return. The comparison of the audit return by selection method is of course complicated by the incomplete implementation of the audit program, which leads us to work with the selected sub-sample of those audit cases actually implemented. The analysis is further complicated by the incomplete nature of the audit results data.

Table

10 shows slightly different results for the audit yield as a share of turnover, a proxy of the firm's evasion rate. Mechanically, turnover is negatively correlated with the outcome. The algorithm selection dummy is not positive, suggesting algorithm-selected cases exhibited a higher audit return, although this result is not statistically significant. The effect becomes

significant, however, in the tax offices in charge of liberal professional, among whom we would indeed expect evasion to be high and the predictive power of the algorithm to be strong.

6 Conclusion

In this paper, we have studied whether a data-driven algorithm can help improve the targeting of enforcement, focusing on the context of tax audits. Collaborating with the Senegalese tax administration DGID in an intervention at scale, we compare the implementation and return of audit cases which were selected by a risk-scoring algorithm to cases selected by tax inspectors based on a traditional discretionary procedure. We also test whether providing inspectors with easily analyzable information and with risk flags about the selected cases improves audit outcomes. Our analysis relies on partial outcome data, so that the results are still preliminary. We find that algorithm-selected audits are slightly less likely to have been implementation, and that inspector-selected audits focus on larger firms and detect the same evasion rate, thus yielding a higher return in absolute value. We do not find any evidence that the provision of information on the case or of risk flags affects audit outcomes. We aim to soon update our empirical analyses with manually digitized data on audit outcomes, which we consider more complete and of higher quality.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. *Handbook of economic growth*, 1:385–472.

Alm, J., Blackwell, C., and McKee, M. (2004). Audit selection and firm compliance with a broad-based sales tax. *National Tax Journal*, pages 209–227.

Bertrand, M., Burgess, R., Chawla, A., and Xu, G. (2018). The glittering prizes: Career incentives and bureaucrat performance. Technical report, mimeo.

Besley, T. and Persson, T. (2009). The origins of state capacity: Property rights, taxation, and politics. *American Economic Review*, 99(4):1218–44.

Besley, T. and Persson, T. (2010). State capacity, conflict, and development. *Econometrica*, 78(1):1–34.

Besley, T. and Persson, T. (2013). Taxation and development. In *Handbook of public economics*, volume 5, pages 51–110. Elsevier.

Besley, T. and Persson, T. (2014). Why do developing countries tax so little? *Journal of Economic Perspectives*, 28(4):99–120.

Best, M. C., Brockmeyer, A., Kleven, H. J., Spinnewijn, J., and Waseem, M. (2015). Production versus revenue efficiency with limited tax capacity: Theory and evidence from pakistan. *Journal of Political Economy*, 123(6).

Carrillo, P., Pomeranz, D., and Singhal, M. (2017). Dodging the taxman: Firm misreporting and limits to tax enforcement. *American Economic Journal: Applied Economics*, 9(2):144–64.

Duflo, E., Greenstone, M., Pande, R., and Ryan, N. (2018). The value of regulatory discretion: Estimates from environmental inspections in india. *Econometrica*, 86(6):2123–2160.

Finan, F., Olken, B. A., and Pande, R. (2017). The personnel economics of the developing state. In *Handbook of Economic Field Experiments*, volume 2, pages 467–514. Elsevier.

Gerardino, M. P., Litschig, S., and Pomeranz, D. (2017). Can audits backfire? evidence from public procurement in chile. *NBER Working Papers*, (23978).

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The quarterly journal of economics*, 124(4):1403–1448.

Jensen, A. (2016). Employment structure and the rise of the modern tax system. *Job market paper*, 37.

Khwaja, M. S., Awasthi, R., and Loeprick, J. (2011). *Risk-based tax audits: approaches and country experiences*. The World Bank.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., and Saez, E. (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica*, 79(3):651–692.

Kleven, H. J., Kreiner, C. T., and Saez, E. (2016). Why can modern governments tax so much? an agency model of firms as fiscal intermediaries. *Economica*, 83(330):219–246.

Monitor, I. F. (2017). Achieving more with less.

Murray, M. N. (1995). Sales tax compliance and audit selection. *National Tax Journal*, pages 515–530.

Naritomi, J. (2019). Consumers as tax auditors. *American Economic Review*, 109(9):3031–72.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review*, 105(8).

Restuccia, D. and Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics*, 11(4):707–720.

Slemrod, J., Blumenthal, M., and Christian, C. (2001). Taxpayer response to an increased probability of audit: evidence from a controlled experiment in minnesota. *Journal of public economics*, 79(3):455–483.

Slemrod, J., Collins, B., Hoopes, J., Reck, D., and Sebastiani, M. (2015). Does credit-card information reporting improve small-business tax compliance? Technical report, National Bureau of Economic Research.

Troiano, U. (2017). Intergovernmental cooperation and tax enforcement. Technical report, National Bureau of Economic Research.

Xu, G. (2019). The colonial origins of fiscal capacity: Evidence from patronage governors. *Journal of Comparative Economics*.

7 Figures and Tables

8 Figures

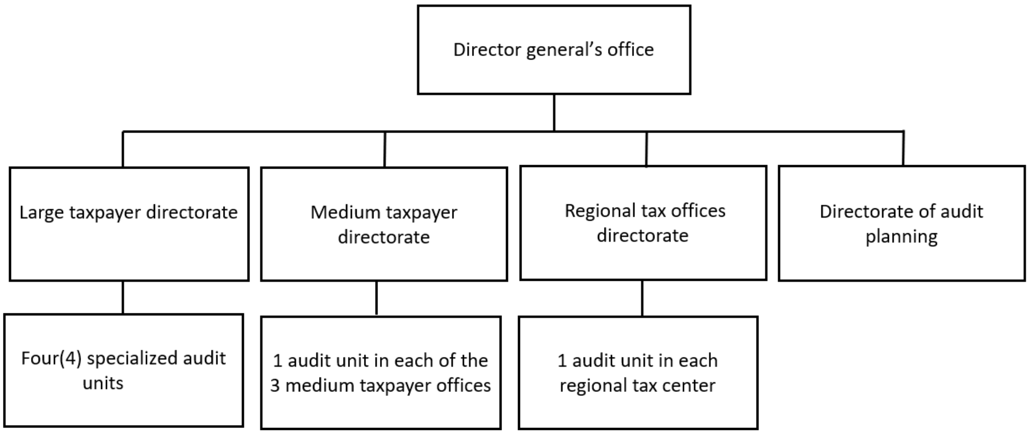

Director general's office
Large taxpayer directorate
Medium taxpayer directorate
Regional tax offices directorate
Directorate of audit planning
Four(4) specialized audit units
1 audit unit in each of the 3 medium taxpayer offices
1 audit unit in each regional tax center

Figure 1: DGID's organizational chart

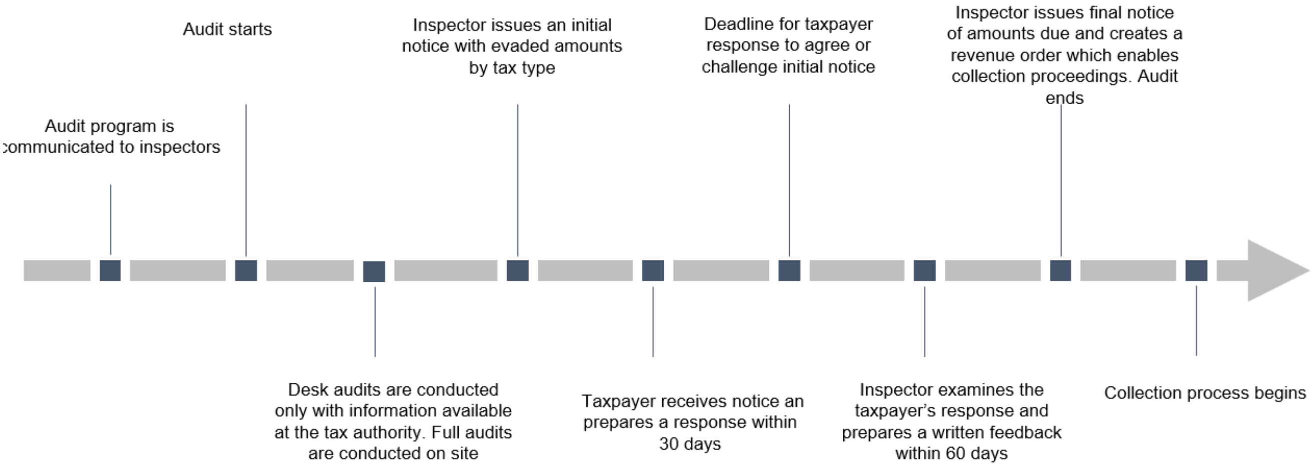

Audit program is communicated to inspectors
Audit starts
Desk audits are conducted only with information available at the tax authority. Full audits are conducted on site
Inspector issues an initial notice with evaded amounts by tax type
Taxpayer receives notice an prepares a response within 30 days
Deadline for taxpayer response to agree or challenge initial notice
Inspector examines the taxpayer's response and prepares a written feedback within 60 days
Inspector issues final notice of amounts due and creates a revenue order which enables collection proceedings. Audit ends
Collection process begins

Figure 2: Audit process

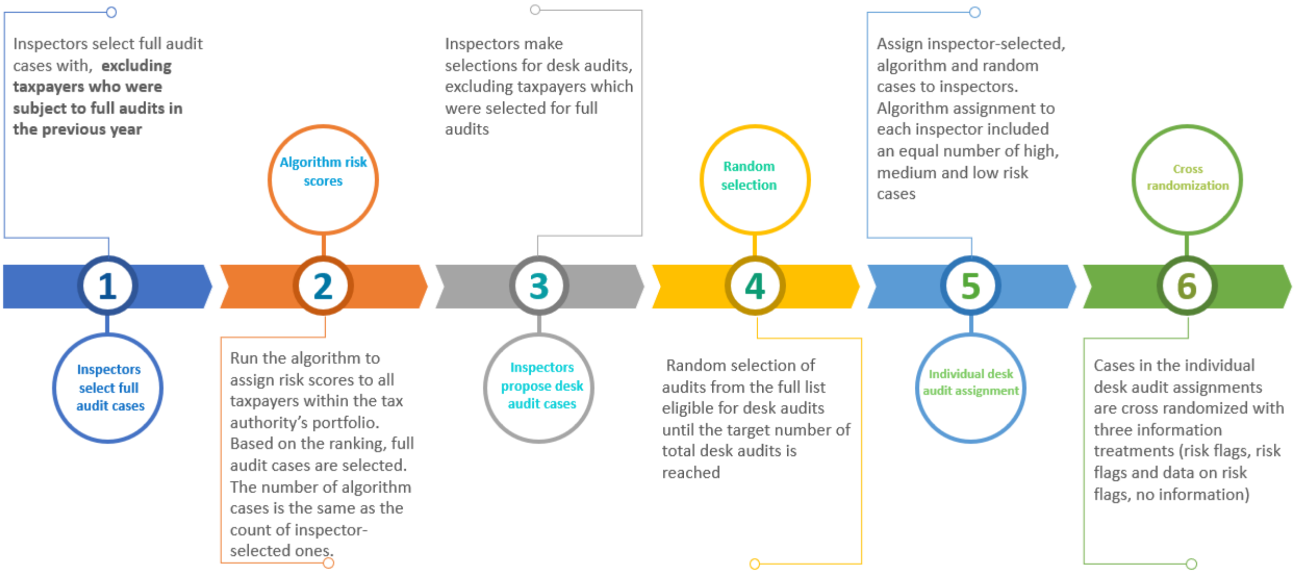

Inspectors select full audit cases with, **excluding taxpayers who were subject to full audits in the previous year**
1
Inspectors select full audit cases
Algorithm risk scores
2
Run the algorithm to assign risk scores to all taxpayers within the tax authority's portfolio. Based on the ranking, full audit cases are selected. The number of algorithm cases is the same as the count of inspector-selected ones.
Inspectors make selections for desk audits, excluding taxpayers which were selected for full audits
3
Inspectors propose desk audit cases
Random selection
4
Random selection of audits from the full list eligible for desk audits until the target number of total desk audits is reached
Assign inspector-selected, algorithm and random cases to inspectors. Algorithm assignment to each inspector included an equal number of high, medium and low risk cases
5
Individual desk audit assignment
Cross randomization
6
Cases in the individual desk audit assignments are cross randomized with three information treatments (risk flags, risk flags and data on risk flags, no information)

Figure 3: Program design and audit selection timeline

9 Program execution

The following sections provide an analysis of the 2019 audit reports, executed in the scope of an experiment in partnership with the Senegalese Internal Revenue Services (DGID in the French acronym, henceforth designated IRS). The experiment consisted in altering the selection method of the audits program of 2019 in some fiscal centers. Part of the audits program was chosen according to the IRS' discretionary method, and part was chosen according to an algorithm, following explicit rules. The tax authority was then asked to carry out the audits on the selected firms. At the end of the year, only part of the initially planned audits had been carried out. The purpose of the analysis is to establish whether the use of the algorithm improved the ability of the tax authority to select firms for audit, especially in terms of verified tax evasion.

The audits program of 2019 consisted of 1298 firms in seven different tax centers: the two centers for middle-sized enterprises (called CME 1 and CME 2 in the French acronym), the center for liberal professionals (CPR) and four location-specific centers for small and medium enterprises, all of them in the region of Dakar, Senegal's capital (the four centers were Dakar Plateau, Grand Dakar, Ngor Almadies and Pikine Guediawaye). Part of the 1298 firms were not initially in the list of selected firms, prepared in the beginning of 2019, but were added at the IRS' discretion during the course of the year. We added them as firms selected by the IRS in our analysis.

Table 2 summarizes the execution of the 2019 progam. Out of the 1298 selected firms, 1068 were chosen to be subject to "short audits" (also called CP in the Senegalese IRS' jargon), and the remaining 230 were supposed to be subject to "full audits" (VG in the IRS' jargon). The execution rate was around 50%, meaning that for half the firms in the list there is no indication that the inspectors audited them. For the remaining half, only 37% of them ended in a request for adjustment and eventual payment of a fine.

Table 2: Count of execution of 2019 program

Table 3: Count of execution of 2019 program

	Cases		Started		Adjustment	
	Short	Full	Short	Full	Short	Full
CME 1	147	55	46	13	27	1
CME 2	251	44	215	42	86	26
CPR	181	41	71	31	15	18
Dakar Plateau	206	37	135	2	7	0
Grand Dakar	51	17	38	7	23	3
Ngor Almadies	145	23	42	4	2	0
Pikine Guediawaye	84	15	28	3	7	1
All	1065	232	575	102	167	49

Obs: Count of total cases proposed, started and with tax adjustment in the 2019 audit program in Senegal. It includes cases selected by algorithm and by the tax inspectors.

Obs: Count of total cases proposed, started and with tax adjustment in the 2019 audit program in Senegal. It includes cases selected by algorithm and by the tax inspectors.

Table 4: Tax audit selection methods in selected countries

ntry	Discretionary selection	Risk analysis	Random selection
ya	Yes ; For all except large taxpayers	Yes ; Only for large taxpayers	No
egal	Yes	Yes, Introduced in FY 2018	Introduced in FY 2018
babwe	Yes; Inspectors rated on selection.	Yes; based on turnover variances	No
otho	No	No	Yes ; Randomly by ma
zania	Abandonned in 2007	Yes	
ted Kingdom	Yes; For 55% of audit cases	Yes; Risk scoring	Yes ; Simple random s
tzerland	Yes for all cases	No	Yes, periodically for so
ted States	No	Yes	
nce	Yes; For intelligence gathering	Yes; statistical techniques, data-mining	No
garia	Yes ; According to set criteria	Yes; Central risk analysis	No
key	No	Yes; Analysis by tax type	Yes ; to collect unbias

ources; Khwaja et al. 2011 and Authors' survey of select country tax officials.

9.1 Firms' characteristics

Table 5: Number of firms by data source

		2014	2015	2016	2017	2018	2019
Self reported	VAT	0	6138	6359	6486	5883	5842
	CIT	0	3823	3970	4245	4159	0
	CGU	0	16	34	63	76	62
	WIT	0	4503	4574	5101	5329	5344
	TAF	0	19	18	19	18	16
Third party	Imports	0	1500	1556	1483	1450	0
	Exports	0	446	463	441	429	0
	Treasury	0	547	547	428	444	0
	VAT annexes	0	0	0	0	0	0
Audits data	Fiches de suivi	0	0	0	0	0	1286
	Saisie	0	0	0	0	0	0

Note: Number of firms for which data was available, according to each data source. There are three main sources of data: self-reported tax declarations (Value Added Tax, Corporate Income Tax, simplified regime CGU, Withtheld Income Tax, financial services tax TAF), third party data (exports, imports, treasury payments and VAT annexes concerning inter-firm transactions) and the data produced by the tax inspectors regarding the audit program of 2019. The data includes the following tax centers in Senegal: medium taxpayers 1, medium taxpayers 2, liberal professionals, Dakar Plateau, Grand Dakar, Pikine Guediawaye, Ngor Almadies.

Table 6: Number of firms by data source

	Mean population	Mean random selection	Difference	p-value	Mean IRS selection	Mean algorithm selection	Difference	p-value
Turnover (mean 2015-2018)	162	172	-10	.87	463	290	173	0
Mean profit (mean 2015-2018)	1	3	-1	.85	6	-2	7	.23
Profit rate (2015-2018) Mean Payroll (2015-2018)	9	11	-3	.48	24	15	9	.01
Tax liability (total 2015-2018)	33	30	3	.76	94	84	10	.48
Risk score	0	156	-156	0	175	1618	-1443	0
Turnover 2018	233	326	-93	.4	597	400	197	.01
Profit 2018	15	8	7	.87	34	70	-36	.46
Number of employees 2018	414	12	403	.3	229	215	14	.93
N	11386	154	.	.	600	574	.	.

Note: Number of firms for which data was available, according to each data source. There are three main sources of data: self-reported tax declarations (Value Added Tax, Corporate Income Tax, simplified regime CGU, Withtheld Income Tax, financial services tax TAF), third party data (exports, imports, treasury payments and VAT annexes concerning inter-firm transactions) and the data produced by the tax inspectors regarding the audit program of 2019. The data includes the following tax centers in Senegal: medium taxpayers 1, medium taxpayers 2, liberal professionals, Dakar Plateau, Grand Dakar, Pikine Guediawaye, Ngor Almadies.

9.2 Outcomes

To analyze the data, we propose six outcomes: the probability that the audit started, the probability that there was an adjustment (conditional on audits having started), the amount of the first notification (the initial quantity of suspected evasion communicated to the taxpayer), the confirmed amount of evasion, the evasion rate as a percentage of the total tax liability, and the evasion rate as a percentage of mean turnover. A first comparison of the outcomes across short audits and full audits, and across selection methods, can be observe in table **??** below.

Table 7: Mean characteristics firms - All firms

	Mean Random	Mean IRS selection	Mean algorithm selection	Difference	p-value
1 probability being started	.58	.63	.49	.14	0
2 audit ending in adjustment	.31	.57	.31	.27	0
3 log (initial notice)	17.17	17.88	17.33	.55	.01
4 log (final notice)	16.59	17.22	16.75	.47	.02
5 evasion as % liability	.71	.68	.68	.01	.9
6 evasion as % of mean turnover	.4	.3	.4	-.1	.01
7 days spent on case	4.55	40.77	19.08	21.69	0
8 log turnover 2019	11.02	14.36	14.12	.24	.85

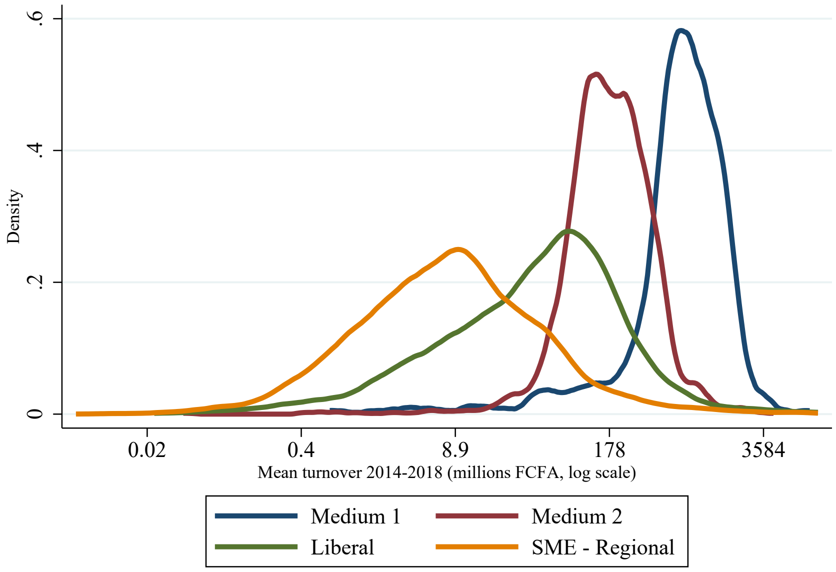
Note: Mean characteristics of firms in selection and in the population. Total tax liability includes only self declared tax liability in VAT, CIT, PAYE and CGU for firms. The data includes the following tax centers in Senegal: medium taxpayers 1, medium taxpayers 2, liberal professionals, Dakar Plateau, Grand Dakar, Pikine Guediawaye, Ngor Almadies. Values of turnover, tax liability and profits are expressed in Millions FCFA. Profit rate is in percentage of turnover, computed as the mean profit divided by the mean turnover. Number of employees refers to the number of employees in the PAYE declarations.

The definition of the outcomes is as follows:

- $i(Auditstarted)$: indicator function that takes value 1 if the audit contained any indication that the inspector worked on it. This variable takes value 1 whenever the audit report of the firm contains the indication of some evasion quantity, some qualitative variable, or even an indication of the date in which the audit was started. For many cases, the audit is started but not finished.
- $i(Adjustment > 0)$: indicator function containing some quantity of uncovered evasion. It can be the final amount the firm is asked to pay or the initially notified amount (which happens more often).
- $log(Notification)$: log of the value of the notified amount of evaded taxes. That is the amount of evasion that is assessed by the inspectors after the inspection. This amount is then negotiated with the firm, which provides some explanation about the problems, and is typically reduced in the confirmation stage.
- $log(Evasion)$: log of the assessed evasion of the firm. In this stage, we use the value of the *confirmed amount* of evaded taxes or the *final requested payment*. Whenever the two values are not the same (which happens very rarely) we take the max between them. We complement missing information with the value of notification (the outcome before) adjusted by the mean deduction from notification and confirmation at the tax office level and for each particular audit type (full or short audits). For example, in the Liberal Professions office, we observe that on average the confirmation is 57% the value of the initial notification (when both quantities are filled in) for full audits, so when we only have the value of notification (for full audits in that particular office) we complement the evasion variable by multiplying it by 57%.

9.3 Description of the firms and outcomes by firm size

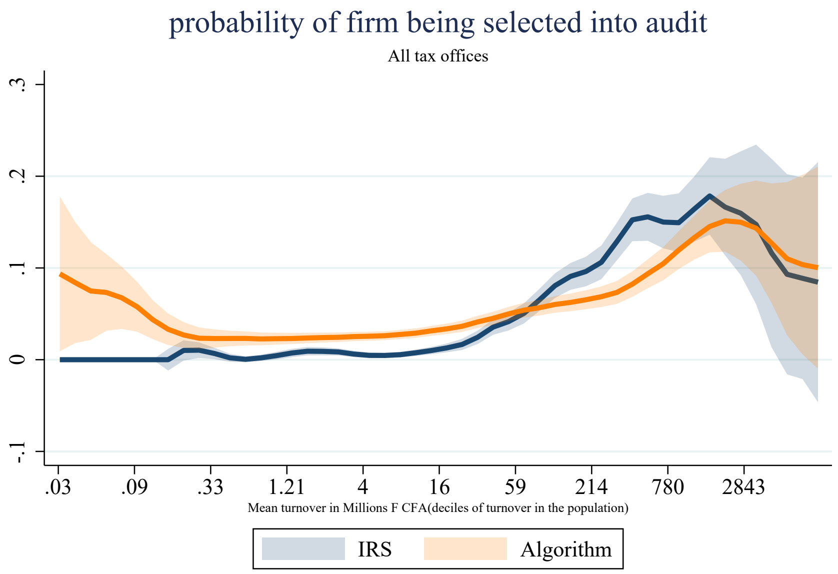
In the 2019 wave of the experiment, we proposed firms to be audited in tax centers in the Dakar area. The tax centers included small to medium enterprises. Based on their self reported yearly turnovers, we can plot the distribution of firm size in each of the tax centers below.



Note: graph only considered firms with positive mean turnover.

Figure 4

In every tax office, firms with larger declared turnover have a higher probability of being audited, in particular for IRS selected cases. The algorithm also gives explicitly more weight to firms with more declared turnover. Even though the algorithm explicitly gives more weight to firms with larger turnover, its selection is less concentrated at large firms than the inspector selection. The following figure shows how the two selections differ in terms of (self-declared) firms size. Firms with mean declared turnover lower than 16 Million FCFA (roughly 25 thousand euros) per year have virtually no chance of being selected for audit by the tax authority, while the algorithm assigns them positive probability of audit. In particular, firms with extremely low declarations had almost 10% chances of being selected by the algorithm, while no chance of being selected by the tax inspectors.



Note: Comparison between firms selected by IRS (598 firms) and algorithm (582 firms).

Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

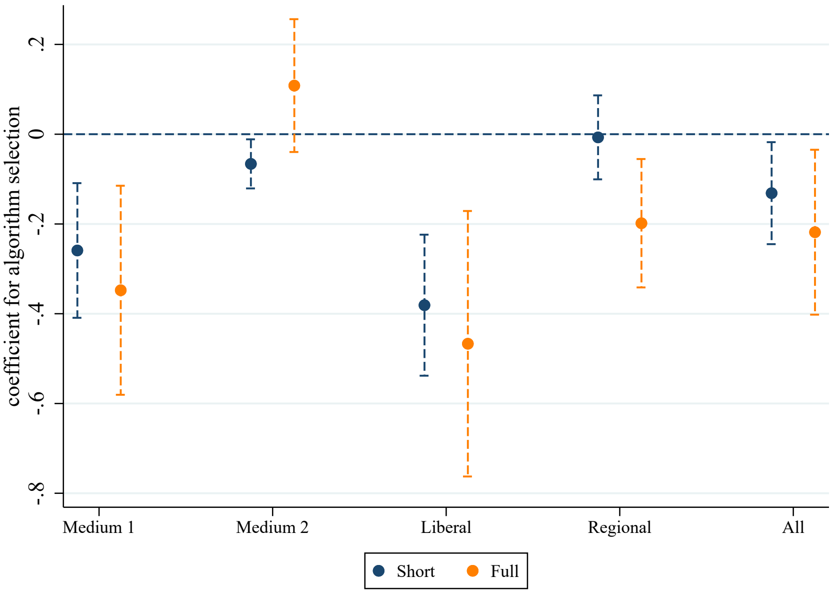
9.4 Impact of selection on outcomes

Table 8: Effect of algorithm selection on probability of audit being started

	(1)	(2)	(3)	(4)	(5)	(6)
	All audits	All audits	All audits	All audits	Short audits	Full audits
Algorithm selection	-0.178**	-0.0704	-0.0708	-0.187**	-0.161**	-0.250**
	(0.0614)	(0.0550)	(0.0551)	(0.0695)	(0.0512)	(0.0954)
Random audits	0.0671*	0.0171	0.0175	0.0835*	0.0591	
	(0.0327)	(0.0567)	(0.0557)	(0.0359)	(0.0433)	
log Mean turnover 2014-2018	0.00188	0.00539	0.00534	0.00245	0.00161	0.00481
	(0.00230)	(0.00601)	(0.00608)	(0.00432)	(0.00278)	(0.00556)
log Turnover 2018	-0.000111	0.000204	-0.00000511	0.000591	-0.0000486	0.00612
	(0.00178)	(0.00160)	(0.00162)	(0.00170)	(0.00138)	(0.00463)
Full audit	-0.0852	-0.0787	-0.0782			
	(0.138)	(0.137)	(0.136)			
Overlap (selected by algorithm and IRS)		0.0308	0.0297	0.102		
		(0.140)	(0.143)	(0.0990)		
Replacement cases		-0.236**	-0.226**			
		(0.0811)	(0.0720)			
Risk score		-0.0000203	-0.0000207			
		(0.0000186)	(0.0000191)			
Information treatment		0.0312				
		(0.0240)				
Info. treatment × Algorithm selection		-0.00615				
		(0.0381)				
Info. treatment (only risk indicators)			0.000911			
			(0.0490)			
Info. treatment (only risk indicators) × Algorithm			0.0126			
			(0.0640)			
Info. treatment (risk indicators plus data)			0.0688			
			(0.0504)			
Info. treatment (risk indicators plus data)× Algorithm			-0.0306			
			(0.0456)			
Tax Center fixed effects	Yes	Yes	Yes	No	Yes	Yes
Activity group fixed effects	No	Yes	Yes	No	No	No
Inspector fixed effects	No	No	No	Yes	No	No
N	943	872	872	907	753	190
R2	0.292	0.337	0.338	0.506	0.320	0.489

Note: OLS regression of probability of audit being started on the selection method. Different specifications controlling for the type of audit, the firm's mean turnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, medium enterprises 2, liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parentheses, and were computed clustered at the tax center level.

Figure 5: Effect of algorithm selection on probability of audit being started, by tax center



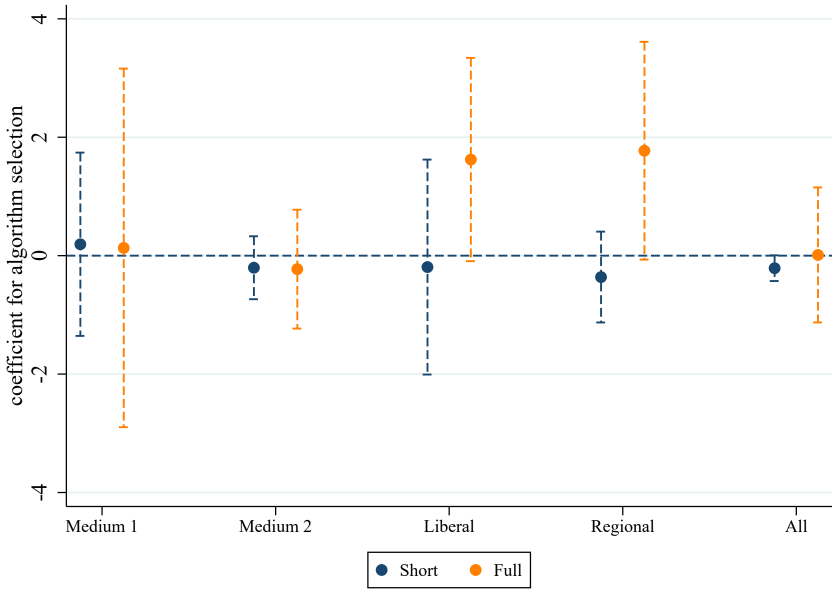
Obs: Coefficients of the regression of the outcome on the algorithm selection, controlling for mean firm turnover, by tax office and type of audit. The last two coefficients (All) represent the coefficients of same regression as the last two columns of the corresponding regression table.

Table 9: Effect of algorithm selection on log (initial notice)

	(1)	(2)	(3)	(4)	(5)	(6)
	All audits	All audits	All audits	All audits	Short audits	Full audits
Algorithm selection	-0.219	-0.567	-0.570	-0.195	-0.295	-0.0234
	(0.285)	(0.525)	(0.559)	(0.305)	(0.258)	(0.611)
Random audits	0.321	0.512	0.484	0.506	0.362	
	(0.333)	(0.431)	(0.422)	(0.482)	(0.355)	
log Mean turnover 2014-2018	0.00326	0.0951	0.0932	-0.0233	0.0302	-0.0471***
	(0.0408)	(0.0975)	(0.0985)	(0.0417)	(0.0677)	(0.00942)
log Turnover 2018	-0.00522	-0.0231	-0.0228	0.0187	-0.0138	0.0542**
	(0.0266)	(0.0262)	(0.0256)	(0.0256)	(0.0319)	(0.0188)
Full audit	1.067*	0.730	0.725			
	(0.508)	(0.530)	(0.544)			
Overlap (selected by algorithm and IRS)		0.239	0.268	0.468		
		(0.301)	(0.276)	(0.312)		
Replacement cases		0.0452	-0.0568			
		(0.337)	(0.251)			
Risk score		0.000128	0.000131			
		(0.0000889)	(0.0000982)			
Information treatment		-0.560				
		(0.294)				
Info. treatment × Algorithm selection		0.628				
		(0.382)				
Info. treatment (only risk indicators)			-0.616			
			(0.327)			
Info. treatment (only risk indicators) × Algorithm			0.806			
			(0.500)			
Info. treatment (risk indicators plus data)			-0.479			
			(0.305)			
Info. treatment (risk indicators plus data)× Algorithm			0.429			
			(0.378)			
Tax Center fixed effects	Yes	Yes	Yes	No	Yes	Yes
Activity group fixed effects	No	Yes	Yes	No	No	No
Inspector fixed effects	No	No	No	Yes	No	No
N	221	187	187	221	161	60
R2	0.230	0.349	0.351	0.462	0.235	0.140

Note: OLS regression of log (initial notice) on the selection method. Different specifications controlling for the type of audit, the firm's mean turnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, medium enterprises 2, liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parentheses, and were computed clustered at the center level.

Figure 6: Effect of algorithm selection on log(initial notice), by tax center

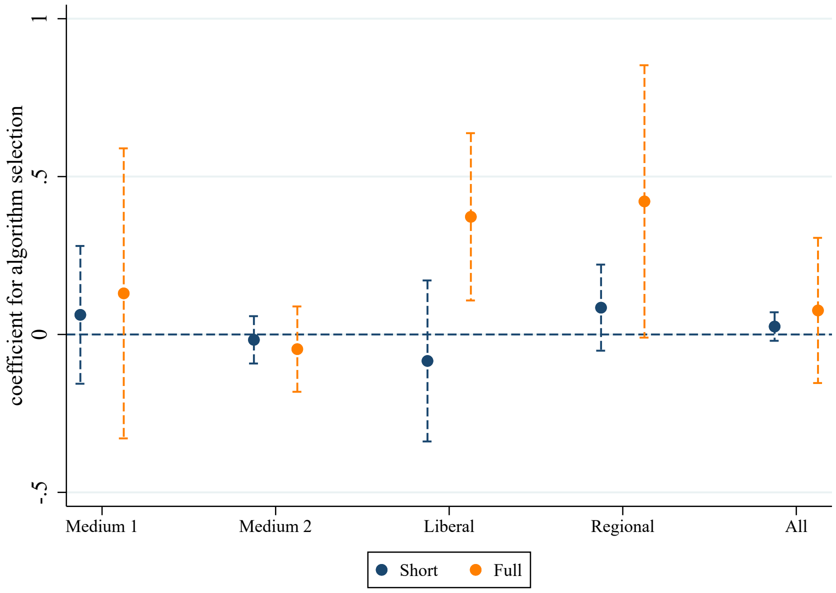


Obs: Coefficients of the regression of the outcome on the algorithm selection, controlling for mean firm turnover, by tax office and type of audit. The last two coefficients (All) represent the coefficients of same regression as the last two columns of the corresponding regression table.

Table 10: Effect of algorithm selection on evasion as % of mean turnover

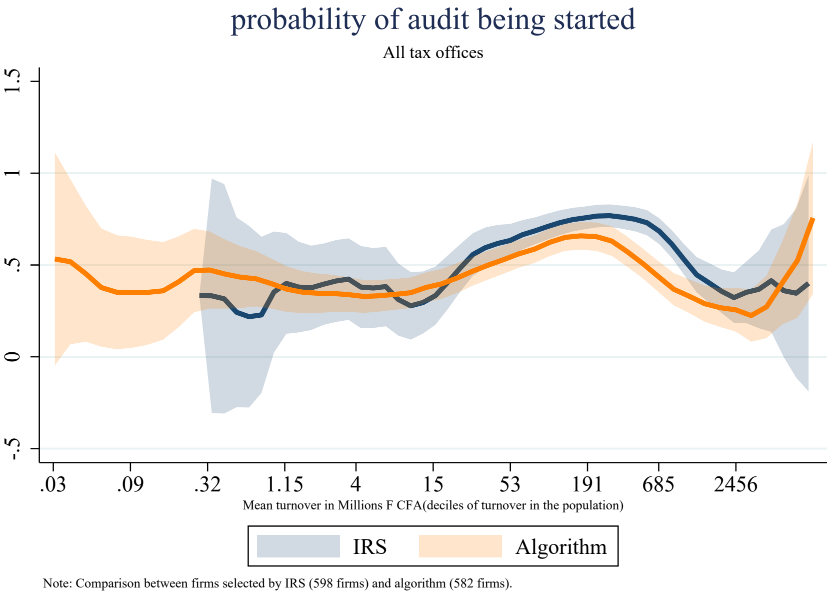
	(1)	(2)	(3)	(4)	(5)	(6)
	All audits	All audits	All audits	All audits	Short audits	Full audits
Algorithm selection	0.0484	-0.00390	-0.00210	0.0536	0.0336	0.0714
	(0.0644)	(0.0847)	(0.0908)	(0.0741)	(0.0547)	(0.138)
Random audits	-0.0391	-0.0161	-0.0248	-0.00630	-0.0226	
	(0.0691)	(0.0719)	(0.0732)	(0.103)	(0.0604)	
log Mean turnover 2014-2018	-0.0322***	-0.0507*	-0.0512*	-0.0326**	-0.0438***	-0.0277***
	(0.00857)	(0.0229)	(0.0223)	(0.00931)	(0.0116)	(0.00106)
log Turnover 2018	-0.0108	-0.00991	-0.0101	-0.00863	-0.0104	-0.00674*
	(0.00653)	(0.00682)	(0.00649)	(0.00614)	(0.00744)	(0.00249)
Full audit	0.0780	0.0321	0.0324			
	(0.0487)	(0.0680)	(0.0704)			
Overlap (selected by algorithm and IRS)		-0.0212	-0.0208	0.0192		
		(0.0264)	(0.0226)	(0.0354)		
Replacement cases		-0.00869	-0.0261			
		(0.0768)	(0.0609)			
Risk score		0.0000158	0.0000158			
		(0.0000141)	(0.0000159)			
Information treatment		-0.110**				
		(0.0386)				
Info. treatment × Algorithm selection		0.0877				
		(0.0668)				
Info. treatment (only risk indicators)			-0.140**			
			(0.0394)			
Info. treatment (only risk indicators) × Algorithm			0.134			
			(0.0733)			
Info. treatment (risk indicators plus data)			-0.0679			
			(0.0484)			
Info. treatment (risk indicators plus data)× Algorithm			0.0286			
			(0.0780)			
Tax Center fixed effects	Yes	Yes	Yes	No	Yes	Yes
Activity group fixed effects	No	Yes	Yes	No	No	No
Inspector fixed effects	No	No	No	Yes	No	No
N	221	187	187	221	161	60
R2	0.384	0.416	0.421	0.514	0.352	0.528

Note: OLS regression of evasion as % of mean turnover on the selection method. Different specifications controlling for the type of audit, the firm's mean turnover (with the information available over years 2015-2018), and dummies for the 6 tax centers (medium enterprises 1, medium enterprises 2, liberal professions, Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies). Standard errors are shown parentheses, and were computed clustered at the tax center level.



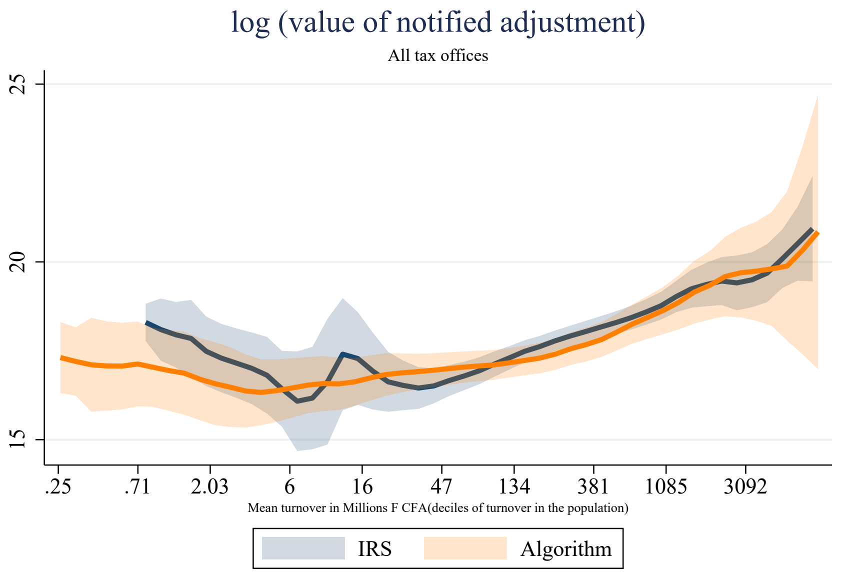
Obs: Coefficients of the regression of the outcome on the algorithm selection, controlling for mean firm turnover, by tax office and type of audit. The last two coefficients (All) represent the coefficients of same regression as the last two columns of the corresponding regression table.

Figure 7: Probability of audit being started



Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

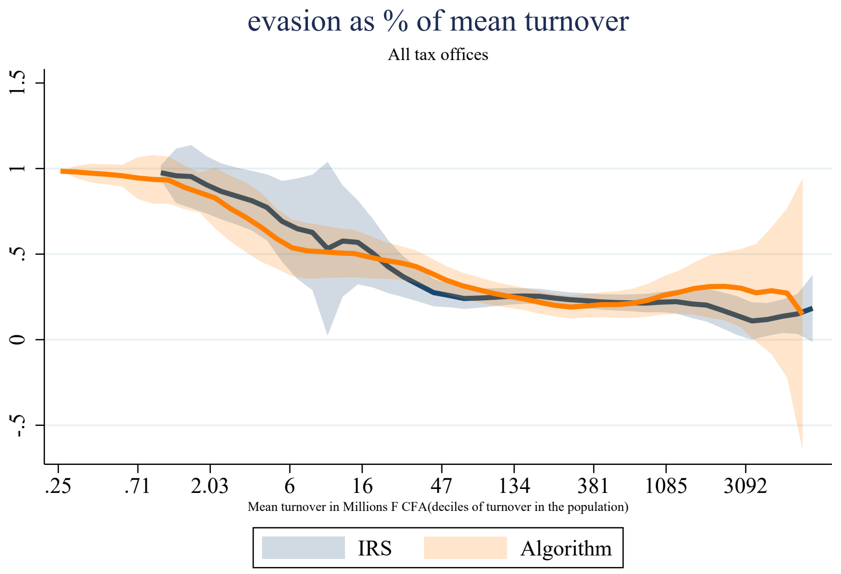
Figure 8: log(Initial notice)



Note: Comparison between firms selected by IRS (182 firms) and algorithm (85 firms).

Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

Figure 9: Evasion as a % of mean turnover



Note: Comparison between firms selected by IRS (155 firms) and algorithm (85 firms).

Obs: Non parametric regression of outcome on mean turnover (no controls), using Epanechnikov kernel, bandwidth computed according to the rule-of-tumb method.

9.5 Evaluation of risk score

Appendix C Risk Scoring of Tax Evasion

C.1 Motivation

A key feature of this project is to assist the Senegalese tax administration (DGID) to design a tool which assesses firms' tax evasion risk. Starting in 2017, the team held consultations with DGID leadership and former tax inspectors to map the compliance risks of Senegalese firms and to exploit all available data sources to assess this risk. Moreover, we discussed with experts in the field of taxation and risk management, who worked on tax evasion risk assessment in middle-income countries. With these inputs, we designed a risk-scoring tool, following best international practice, as implemented by the World Bank and its partner institutions.

Although the use of advanced machine-learning tools for prediction has exploded in economic analysis, it was decided together with DGID that the risk-score would be guided by simple variables which logically should predict evasion risk. The simplicity of the design is motivated by several factors, ranked by order of importance. First, the tool needed to be transparent, such that underlying compliance risks could be understood by tax inspectors, and explained to taxpayers when required. Second, the available data on historical audit results was sparse and not digitized, which limited the scope of our model calibration and model selection exercises (further details below). Finally, all cases concluded by 2017 were selected in a discretionary manner.

Thus, one should consider the risk-scoring tool as a transparent best-practice risk assessment, given the administrative capacity, rather than a fined-tool fully optimized algorithm. We note that the constraints faced by DGID are likely to bind in many low income countries, and especially in other West African countries, which often look at Senegal for administrative innovations.

Table XX summarizes the seven key steps in the design of the risk-score. Step (1) corresponds to the construction of a database covering all tax declarations across years and merged with third-party reported sources. Steps (2) and (3) determine specific risk indicators, based on discrepancies across sources or behavioral outliers, examples of which are discussed below. Step (4) defines the peer-group comparison: these clusters regroup firms by economic activity and either size or geographical zones, depending on the structure of each tax center. Step (5) assigns a numerical value to each risk indicator, depending on the size of the deviation (higher scores when larger discrepancies), while step (6) assigns weights to each indicator reflecting beliefs about their relative importance. Finally, step (7) aggregate the weighted indicators in each of the past four fiscal year, and then sums up the yearly scores to form a total risk score.

Table C1: Steps of risk-score design

Step	**Description**
(1) Prepare merged dataset	The tax declarations of each taxpayer are merged across type of taxes (VAT, CIT, Payroll) and across years. Data from third parties is then added (customs, procurement, transaction network).
(2) Choose indicators: discrepancies	Discrepancies are situations in which a self-reported tax liability can be considered as misreported or incomplete, by cross checking several data sources together.
(3) Choose indicators: anomalies	Anomalies correspond to abnormal reporting behavior, compared to peers. Anomalies suggest that firms should be monitored, but do not indicate tax evasion behavior with certainty.
(4) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons are done within clusters
(5) Assign values to indicators	The magnitude of the inconsistency is used to assign a value, ranging from one to ten (using deciles). For anomalies firms within the top decile of a particular indicator receive a value of one.
(6) Assign weights to indicators	Weights are assigned to each indicator reflecting beliefs about their relative importance.
(7) Aggregate indicators and years	The weighted risk indicators are first aggregated across indicators in each year. Then the yearly scores are summed up to form a total risk score covering the past four years of tax declarations. More recent years are slightly over-weighted.

C.2 Choosing indicators and weights

As explained above, the algorithm computes some ratios from the data of firms (declarations and third party data) and then calculates the value of the indicator based on the distribution of this ratio within a cluster of comparable firms. We tried several combinations of indicators before stabilizing the algorithm in a reduced set of them. The goal was to have a set of indicators that was sensible and correlated with evasion, but at the same time simple and understandable for the tax inspectors.

Table C1 summarizes the steps that we took to conceptualize the algorithm. We tried out several possible indicators that could suggest under-declaration of tax liability. We discarded most based on some analysis of data availability or statistical relevance. In the end, we discarded indicators that required information that was available for a reduced set of firms and indicators that did not seem to have any correlation with evasion, as per past evasion data. We tested these indicators on data from historical audits data. We performed out of sample regressions with LASSO and OLS and computed the out of sample mean squared prediction errors to compare different models. This allowed us to assert that the ranking normalization performed well with respect to alternatives (meaning that it presented a lower prediction error).

We refer to the appendix for an analysis of these indicators using historical audits data. From this analysis we decided to restrict the algorithm to a small list of indicators. Three of them are inconsistencies, plus a flag for inconsistent filing of taxes. On top of that, we have seven anomalies, of which two refer to value added tax, two refer to corporate income tax, one refers to third party data comparisons, one to share of imports from low tax countries and one refers to the financial services tax (only applicable to a reduces set of firms). The final list of indicators that is used in the algorithm, and the respective weights (ω and ξ in equation **??**) is summarized in the following table.

Some details for the calculation of the indicators are worth mentioning. In some cases of anomalies, the top decile within a cluster comprises more than 10% of cases. As long as the value is not zero, we include all these firms. Whenever there is not enough non-zero values that can fill un 10% of the firms, we only flag the non-zero values. We also top code (999 999 999) all values for which the denominator of te underlying ratio of the indicator is zero or missing. Therefore they belong by definition to the top decile. We also top code all values of negative tax liability, to make sure they also get flagged. The idea of the indicators is always that the larger the ratio, the less taxes the firm is paying.

We designed the risk-scoring scheme using best practices, drawing on policy documents from the World Bank (tax administration projects in Pakistan and Turkey), SKAT in Denmark, and the IMF's recommendations to DGID. We provide a high-level description of this process to preserve confidentiality around audit selection processes. We compute risk scores using information sets/tax returns submitted to DGID on corporate income taxes, VAT, personal income tax withholding remittance, as well external data from customs (imports/exports) and public procurement contracts, for the period 2013-2016 [11]. The score relies on two types of risk indicators: discrepancies and anomalies. Discrepancy indicators flag taxpayers whose self-reported information according to their tax returns differs from information in datasets obtained from customs or the government budget department in charge of paying state procurement. For instance, a discrepancy indicator is logged when taxpayers' reported turnover over multiple years is lower than its aggregate costs, that its imports plus its wage bill over the same period. Anomaly indicators use industry/sector benchmarking to flag firms with unusual behavior relative to their peers. An example would be a firm in petroleum retail with low profit rate compared to its peers, which might be associated with evasion. Discrepancies and anomalies are aggregated to produce a risk-score for each taxpayer.

[11] We also attempted to apply predictive analytics from the machine learning literature on these datasets and on previous audit results was conducted to check whether risk indicators could predict DGID audit returns. This exercise was inconclusive because of the selected nature of the sample for whom audit returns are available, the small number of observations and noise in the data.