How to Target Enforcement at Scale? Evidence from Tax Audits in Senegal

Pierre Bachas, Anne Brockmeyer, Alipio Ferreira, Bassirou Sarr*

November 3, 2023

Abstract

Developing economies are characterized by limited compliance with government regulation, such as taxation. Resources for enforcement are scarce and audit cases are often selected in a discretionary manner. We study how bureaucrats respond to the introduction of automatized audit selection. We design a data-driven algorithm to automatically select audits, based on transparent tax evasion indicators. To estimate the effects of automation on inspectors, we randomly place automatically selected audits in the yearly program of tax inspectors in Senegal. The experiment is at-scale, including all tax inspectors in the central and regional tax centers. We find that inspector-selected (i.e. discretionary) audits are more likely to be conducted, to uncover tax evasion, and detect larger evasion amounts. There is little evidence that inspectors actively boycotted automatically-selected cases; rather the lower execution rate seems linked to less attractive characteristics of the algorithm selected firms.

*Pierre Bachas: ESSEC-Business School and World Bank Research, bachas@essec.edu; Anne Brockmeyer: World Bank, abrockmeyer@worldbank.org; Alipio Ferreira: Southern Methodist University, alipioferreira@smu.edu; Bassirou Sarr: Ecole des Hautes Etudes en Sciences Sociales (EHESS). We thank Denis Cogneau, Laurent Corthay, Lucie Gadenne, Janet Jiang, Nicola Limodio, Jan Loeprick, Markus Kitzmuller, Dan Rogger, Eduardo Souza-Rodrigues for helpful comments and discussions, as well as participants at seminars and conferences at the IFS, IIPF, CES Ifo Public Economics Week, TARC Exeter, World Bank Tax Conference, CMI and TaxCapDev Conference, Oxford University, PUC-RJ, and University of New Mexico. We also thank the collaboration with the Senegal Tax Administration (DGID), in particular, Bassirou Samba Niasse, Amadou Abdoulaye Badiane, Oumar Diop Diagne, Hady Dieye, Mor Fall, Serigne Mabaye Fall, and Mathiam Thioub. We thank Samba Mbaye, Assane Sylla, and Medoune Sall from the CRDES, who implemented the taxpayer survey, and Oumy Thiandoum for excellent research assistance. Finally, we gratefully acknowledge the administrative support of the Paris School of Economics and CEPREMAP, and the financial support from the World Bank, UKAID via the Economic Development and Institutions Initiative (EDI) and the Centre for Tax Analysis in Developing Countries (TaxDev), and from the UKRI through Brockmeyer's UKRI Future Leaders Fellowship (grant reference MR/V025058/1).The findings, interpretations, and conclusions do not necessarily represent the views of the World Bank, its affiliated organizations, its Executive Directors or the governments they represent, nor the Senegal Tax Administration.

1 Introduction

The choice of where to allocate inspectors is a critical decision for tax enforcement agencies. Automating this decision can save time, reduce errors, and curb corruption. The benefit of using automation to target enforcement can be particularly large in low-income countries, where bureaucrats hold significant discretion over inspection selection. In contrast, tax audit selection tends to be automated in developed nations, where tax compliance is also higher (Khwaja, Awasthi, and Loeprick 2011). These stylized facts suggest that moving away from discretion is correlated with more efficient tax administrations. However, there is little empirical evidence about the impact of reducing discretion in favor of an automatized, data-driven selection method. We fill this gap by conducting an at-scale field experiment in Senegal, which introduces a novel automated audit selection method, partly replacing a purely discretionary selection method.

Automated selection can mitigate three inefficiencies generated by discretion. First, automation might help inspectors conduct audits by giving them specific reasons to investigate the firm, thus increasing audit execution rates. Second, automation might improve targeting by directing inspectors to non-compliant firms and reducing time wasted in uninteresting cases. Finally, automation lowers information asymmetries between inspectors and their hierarchy, which facilitates monitoring and reduces shirking behavior.[1] However, reducing inspectors' discretion may increase costs for inspectors by adding a cognitive burden to their work and reducing their choice flexibility. Therefore, the success of the transition towards automation relies on the quality of the selection tool and the inspectors carrying out the assigned audits.

This paper analyzes an experimental introduction of an automated selection for tax inspections at the inspector level. We estimate the impact of automation by comparing the performance of automated and discretionary cases in the probability of carrying out the audit and in the audit's outcome. In partnership with the Senegalese tax authority, we designed a risk-score algorithm to detect suspicious taxpayers. This algorithm selected half of the audit cases assigned to tax inspectors in Senegal. The other half of cases in each inspector's list were selected at discretion. This design allows us to observe how inspectors behave when assigned both discretionary and automated cases, the most frequent scenario in high-income countries. Moreover, the introduction of automation happened at scale, such that nearly all inspectors in the main tax offices took part in the experiment.

In 2018 the Senegalese tax administration mandated for the first time that an algorithm selects half of the cases assigned to each tax inspector in the main tax offices. The selection algorithm, constructed by the research team in collaboration with the authorities, leveraged vast amounts of administrative data to create risk indicators that flag potential under-payment of the Corporate Income Tax (CIT) and

[1] These hypotheses were registered at the AEA RCT registry, which can be accessed here: https://www.socialscienceregistry.org/trials/2488.

the Value-Added Tax (VAT), following international best practices. We then performed the selection for each tax inspector and collected data on audit execution throughout the year for both automatic and discretionaryselected cases.

Senegal has a relatively low fiscal capacity, only collecting 16% of its GDP in taxes despite it nominal tax structure being similar to international levels. It has an 18% general VAT rate, a 30% CIT rate, plus personal income taxes, payroll taxes, import duties, and other levies. However, the country's tax revenues suffer from a large informal sector (80% of the labor force) and high tax evasion in the formal sector. Therefore, policies to increase tax compliance are part of an effort to increase Senegal's state capacity. Increasing the tax enforcement agency's efficiency with data and automated processes is one key step in this direction.

We assembled a large quantity of data on the tax-registered firms in Senegal to create the indicators used in the automated selection. First, we merged firms' tax declarations (CIT, VAT, payroll) with third-party data on transactions (customs, procurement, domestic sales, and purchases). Then, following extensive consultations, we constructed evasion risk indicators following international best practices and adapted them to the Senegalese context. The risk score selection algorithm relies on discrepancy indicators, which flag taxpayers with incoherent quantities across datasets, and anomaly indicators, which flag taxpayers with outlying behavior relative to similar firms (e.g. low profit margin). We then aggregate these indicators into a single-dimensional risk score which is used to rank firms and select the most suspicious ones into the audit list.

The intervention took place in January 2018, 2019, and 2020, when inspectors designed the list of audits for the following year. Inspectors typically look at the tax declarations and discuss cases with their superiors, but there is no fixed rule about which firms should be audited, meaning that selection is discretionary. Then, their superiors officially approve the selected list, and inspectors keep a record of their actions regarding each case during the year. The experiment reduced by half (or less) the number of discretionarily selected within each inspector's list. The rest of the cases were selected automatically following the risk-score algorithm, with a small share being randomly selected.

The analysis results indicate that introducing risk-score selection reduces inspectors' probability of starting an audit case. Therefore the average evasion found in the automated list was lower than in the discretionary list. However, conditional on starting the audit, the average evasion amounts were similar across the two lists. Randomly selected cases were also less likely to be started since inspectors perceived them as automated cases and had less evasion than the discretionary cases, even after conditioning on started cases. Finally, overlapping cases, which were selected discretionarily and favored by the risk-score algorithm, showed the highest execution rates and audit yields. Therefore, the most prominent effect of the introduction of automated selection was to reduce audit execution relative to discretionary cases, without any clear compensation in terms of higher audit yields.

Why did automated selection reduce the probability that inspectors start cases? To explain this result, we propose three hypotheses. The first one is that the algorithm was ineffective at targeting relevant cases, and inspectors are better at it. The second explanation is that automated cases are harder to understand since the inspectors did not select them. Finally, the third explanation is that inspectors put less time and effort into algorithm cases because of their preferences or abilities.

The first explanation regards the quality of the algorithm. If the technology did a poor job of targeting non-compliant firms, inspectors would naturally be inclined to avoid carrying out those cases or to find less evasion in the ones they carried out. However, the algorithm risk score correlates positively with evasion within the inspectors' selected cases. Moreover, the risk score also correlates positively with the inspectors' probability of starting the cases in the discretionary list. These correlations suggest that the algorithm's negative result comes from inspectors receiving a new, different list of cases. Although the algorithm's quality can always improve, it does not seem to be a first-order reason for the inspectors' reluctance to do the automatically selected cases.

Indeed, the reason may be that cases selected by a machine may have looked bewildering to some inspectors, who would not have chosen those cases. To test this, we administered a randomized information treatment for each inspector, in which some cases were randomly accompanied with detailed information about the main reasons for suspicion. This random information treatment aimed to test whether providing the rationale for selection would help inspectors carry out the cases. However, we estimate a precise null effect for the information treatment, meaning that lack of justification for selection was again not a first-order driver of the main result.

The rest of the paper is organized as follows. After a review of the relevant literature, Section 2 provides the background for the study and discusses selection methods for tax audits in Senegal. Section 3 presents the data. Section 4 explains the experiment design and the empirical specification used to estimate effects. Section 5 presents the results, heterogeneity, and robustness checks. Section 6 discusses potential mechanisms, and Section 7 concludes.

1.1 Related Literature

Our research contributes to the literature about the use of technology to increase state capacity. Efficient administrations are vital to building state capacity (Besley and Persson 2013). by aligning correctly the incentives of bureaucrats with those of the state (Xu 2019, Bertrand et al. 2018; Finan, Olken, and Pande 2017) or improving enforcement. Other studies papers have studied how technology can help bureaucracies improve its activities by reducing resource waste (Banerjee et al. 2020), and targeting inspections in a more accurate way (Glaeser et al. 2016, Bullock 2019, Glaeser et al. 2021).

To design the algorithm, we made extensive use of third-party data. In countries that use data more

extensively in their enforcement targeting, the presence of third-party data seems to work as a strong deterrent against tax evasion (Kleven et al. (2011), Brockmeyer et al. (2019)). However, our paper shows that the introduction of a technology aimed at improving the efficiency of a bureaucracy had the unintended consequence of making bureaucrats more reluctant to carry out audits. Unintended consequences of enforcement interventions are also documented in Carrillo, Pomeranz, and Singhal (2017) and Gerardino, Litschig, and Pomeranz (2020).

The pros and cons of enforcement discretion is an open debate in the literature. This question is particularly relevant in developing countries, where the trade-offs of using discretion are clearer. Similar to our paper, Duflo et al. (2018) also propose an experimental approach to estimate the impact of selection methods on an enforcement system. They compare discretionary to a random audits for environmental standards of plants in Gujarat, India. They find that discretion outperforms randomly selected audits. Thus, despite the flaws theoretically associated with discretion, inspectors seem to be able to find infractions and punish them more effectively than under purely random selection. In contrast to that paper, we propose a risk-based algorithm to select audits, but reach similar conclusions: inspectors seem to do a good job at selecting audits and uncovering evasion.

The literature on the quality of bureaucracies tends to focus on human resources aspects, and their role in shaping outcomes such as regulatory compliance or quality of public services. Recent experimental evidence has shown that monetary incentives for tax inspectors improve the quality of inspections (Okunogbe and Pouliquen 2018) and increase revenues (Khan, Khwaja, and Olken 2015). Rasul and Rogger (2018) show evidence that management practices impact the quality of public services supplied by bureaucrats.

2 Background

Senegal has a low tax revenue by international standards, having raised 16.7% of GDP in Senegal between 2013 and 2019. Moreover, tax gap estimates indicate that at least 63% of potential income tax revenue (approximately 7% of GDP) and 23% of potential VAT revenue (2% of GDP) are evaded. Improvements in enforcement are expected to have high returns for the tax authority in this setting.

Most tax revenue in Senegal comes from firm taxation, particularly Value Added Tax (VAT) and income taxes (Corporate income tax, personal income tax and dividend withholding taxes), which accounted for 36% and 29% percentage of total tax revenue in 2019 Like in many other countries, most tax revenues are concentrated in large and medium corporations (Slemrod, Blumenthal, and Christian 2001; Slemrod and Velayudhan 2018). Firms also withhold income taxes on their employees' wages (Pay-as-You-Earn), which is often the only source of reporting on individual incomes, given the incompleteness of self-reported personal income taxes. Other significant revenue sources are customs duties (15% of total tax revenue) and excise taxes on petroleum. In this paper we concentrate on the

enforcement of firm taxation, with a focus on VAT, CIT, and PAYE.

The Senegalese Corporate Income Tax (CIT) is paid annually at a rate of 30% profits or 0.5% of turnover, whichever is larger. The Value Added Tax (VAT) is paid on a monthly basis at a standard rate of 18% of value added, or a reduced rate of 10%, typically applied for tourism-related activities. Financial sector firms pay a financial services tax instead of the VAT, also at a rate of 18%. 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) on turnover, which replaces the previously mentioned taxes and has rates varying from 1% to 8% depending on economic sectors and turnover. The Personal Income Tax applies to individuals, and is remitted by firms on behalf of their employees, provided they have a formal employment contract.

The Direction Générale des Impôts et des Domaines (DGID) is the administrative body in charge of domestic tax collection and enforcement, and reports to the Ministry of Finance. DGID is divided into several units: the Large Taxpayer Unit (LTU), which oversees firms whose turnover is greater than or equal to 3 billion CFA francs (approximately 5.3 million USD)[2]; two Medium Taxpayer Units (MTU), which oversees firms with a turnover between 100 million CFA francs and 3 billion CFA francs; a unit for regulated liberal professions, such as lawyers, notaries and medical practitioners; and 19 regional units for the remaining taxpayers, which are mostly individuals or small and medium enterprises (SMEs). Figure 1 displays DGID's organizational chart.

Tax enforcement uses mainly two types of audits: desk audits and full audits. Desk audits are carried out by individual inspectors from within the tax authority's premises, using the firm's tax returns and, potentially, third-party data. Taxpayers are not necessarily aware of these audits unless inspectors make information requests, for example, when data is missing or seems inconsistent. Full audits are carried out by a team of inspectors at the taxpayer's premises. Full audits are announced by letter to the taxpayer at least five days before the audit start date. Tax inspectors may collect information for several weeks, and may continue requesting information for up to 12 months.[3]

Figure 2 illustrates the steps in the audit process. After reviewing a case, inspectors list the detected irregularities and penalties and send them to the taxpayer in an "initial notice". They can also request additional information from the taxpayer. Upon receiving the initial notice, taxpayers have 30 days to respond to the inspector's findings.[4] The inspector examines the response and has 60 days to prepare and send a "confirmation notice" with the detected irregularities and penalties and the final amount to pay. The inspector then generates a revenue order for the tax collection unit, which requires the

[2]The LTU is subdivided into four sub-units, which are specialized by economic sectors. 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 general one with broad competence covering all other sectors.

[3]For firms with a turnover of less than 1 billion CFA francs (about 2 million USD), full audits can only last up to four months. These maximum limits are general rules. There may be extensions in cases with highly suspicious activity or when there is a delay in the transmittal of the requested information to auditors.

[4]If the taxpayer fails to respond, it means for legal purposes that they agree with the inspector's findings.

taxpayer to make a payment within ten business days. Taxpayers can appeal at the Ministry of Finance or a judicial court, and the appeal may suspend the payment process temporarily.

Before our intervention, the selection of cases for tax audits was fully discretionary, in the sense that it did not follow an explicit objective rule. Based on these numerical targets for each tax office, inspectors would draw up lists of proposed cases at the beginning of a fiscal year (normally in January). These lists are harmonized and validated by the manager of each unit, and a then sent to senior management for final adjustments.[5]

Although audit selection is discretionary, it follows some basic procedures. For desk audits, inspectors individually select their lists and send them to their superior, and are not usually required to motivate their choices. For full audits, selection is slightly more procedural, with a more detailed justification of the selection based on the firm's audit history and a summary of indicators such as total sales and profit margin. However, the justification form is not systematically used for all proposed cases and is not systematically stored. Therefore, the criteria used for case selection are usually not explicit and may vary across units and inspectors.

During the last decade, DGID has been investing resources into the digitization of its tax information, through a program called PROMAF. Thanks to these efforts, tax data availability has expanded dramatically in Senegal. Moreover, the large availability of data has enabled the tax authority to experiment with objective, automated selection rules for its audit program. The enforcement agency partnered with the research group to develop a selection tool and evaluate its impact as it is rolled out in the tax units.

3 Data and descriptives

Our study draws data from three administrative sources and two surveys. The administrative data include the tax declarations filed by taxpayers, third-party data on transactions, and audit outcomes.[6] All tax declarations and third-party data were used in the calculation of the risk scores. We complement the administrative data with a tax inspector survey and a taxpayer survey, which were designed by the research team. Only aggregated results from these surveys were shared with the tax administration.

Tax Declarations. Table 1, Panel A, provides an overview of the available tax declarations. Our primary sources of information are the declarations for the Corporate Income Tax (CIT), Value Added Tax (VAT), and the Pay-As-You-Earn (PAYE) income tax witholding reports covering the period of 2014-2019. The CIT data covers about 4,000 firms per year, and the VAT data around 8,000 firms.[7]

[5] Validation by each unit's management ensures that a firm is not selected by two different inspectors. In the selection lists shared with us, we can only observe the validated lists.

[6] We discuss details of the matching process between datasets in Appendix A.1.

[7] The number of VAT filers is higher than the number of CIT filers because self-employed individuals and unincorporated firms file VAT but not CIT.

The PAYE data allows us to calculate the number of employees and the aggregate wage bill for each firm. A small number of firms file tax under the simplified regime *Contribution Global Unique* (CGU), which is available only to firms with a yearly turnover below 50 Million FCFA (approximately 80 thousand USD). Less than 20 financial institutions pay the *Taxe sur les Activités Financières* (TAF), a VAT-substitute for the financial sector.

Third-Party Reports. Table 1, Panel B, describes the information about taxable transactions and activities that we obtain from third parties. Imports and exports are recorded by the customs authority, procurement from state institutions is recorded by the treasury, and firm-to-firm transactions are recorded in VAT annexes that firms file since 2018.[8] While these data are provided at the transaction level, we aggregate them at the firm-year level to merge with the tax declarations.

Audit Reports. We collect audit process and results data in two ways. First, we digitize all audit result reports for 2018-2020. The reports cover all process steps from audit announcement to notification, confirmation, and payment request. They contain the name(s) of the inspector(s) who conducted the audit, the taxes verified in the audit, infractions detected, evaded amounts, applicable penalties, and the dates of each step in the audit process. We use this information to compute our outcomes (e.g. audit yield). The data contain all audit cases, including those selected as part of the annual program and so-called *ad hoc* cases that were opened during the year outside of the planned program. In addition, we asked inspectors to report audit information in an excel sheet pre-filled with their individual list of audit cases. These excel files contain qualitative information about audit cases, such as a rating of the perceived difficult of the audit and indicators for specific difficulties encountered.

Tax Inspector Survey. In 2017, prior to our intervention, we conducted a detailed survey among all available tax inspectors involved in conducting audits, 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.

Taxpayer Survey. After the 2018 and 2019 audit programs were concluded, we surveyed approximately 750 firms in the Dakar region, most of which had been audited as part of the program. We conducted the taxpayer survey in two waves, from October to December 2020 and from March to May 2021. The survey allowed us to elicit taxpayers' views on the audit process, audit risk, and their general perception of the tax administration. We sample firms randomly from the 2018 and 2019 full and desk audits programs, hence targeting both inspector-selected and algorithm-selected.[9]

[8] All VAT filers above xxx of turnover have to file annexes that detail all transactions.

[9] In addition, we also interviewed a random sample of CIT filers which were not part of the audit program in 2018 and 2019, but we do not explore those results in this paper.

3.1 Description of discretionary selection

To shed light on patterns of discretionary selection based on observable firm characteristics, we estimated a selection model using information about the universe of taxpayers in Senegal. We created indicators of filing behavior for VAT, CIT, PAYE, and indicators for whether firms are exporters or importers. Moreover, using tax filings, we created indicators for whether the firm was in the top decile of total sales in its tax center and year, top decile for total material inputs (declared expenditure in VAT), and top decile for total payroll. Finally we also included an indicator for whether the firm declared a loss, and whether the firm was selected for audit. We aggregated these indicators for three years and used it to predict the probability of discretionary selection.

Given that selection is a very rare event in any given year, the vast majority of observations have an outcome of zero. Likewise, given the unsteady pattern of tax filings in Senegal, the predictive variables also contain many zeroes. For that reason, we estimate the model using a Poisson Pseudo-Maximum Likelihood estimator. The results are depicted in Table **??**. Besides the predictive variables described above, we also control for tax office and year fixed effects (for the years 2018, 2019, and 2020).

Column one shows the predictive power of each variable on the probability of discretionary selection as stated in the lists produced early every year. The pseudo-R2 is around 25%, and the variables that have the greatest positive impact are filing CIT and being in the top decile of turnover. VAT and PAYE filing have a weaker, but positive effect, and having been selected in the past has a strong negative impact on the probability of being selected. Exporting firms are weakly less likely to be selected. Column 2 expands the definition of discretionary selection by adding cases that were carried out despite not being planned (the ad hoc cases). The results are qualitatively similar for most cases, though exporting firms are more likely to be selected in this specification.

Finally, Column 3 shows the coefficients for algorithm selection, to show the contrast between the discretionary and risk-based methods. The most striking differences are that firms that fail to file CIT are more likely to be selected by the algorithm, as are firms with greater expenditures, and firms that declare losses. This is partly by design, as we explain in the next section.

4 Experimental Design

4.1 Intervention: Risk-Based Audit Selection

We collaborated with the tax administration to designed and introduce a risk-scoring algorithm for audit selection in Senegal. The objectives of the algorithm were a) to ensure that audit selection followed objective and quantifiable criteria, and b) to increase the audit yields, measured detected evasion amount plus associated penalties and fines. The algorithm generated a risk-score for each

firm and selected the riskier firms to form the audit lists within each tax unit.

To design the algorithm we were faced with two types of constraint. First, the algorithm needed to be intuitive, transparent, and easy to communicate to policymakers and tax inspectors. Second, the algorithm could not be trained on past audit data, using non-parametric statistical approaches from the machine learning literature. The reason is that data on past audits were scarce. Given these constraints, we designed an algorithm based on intuitive indicators discussed with and validated by the tax administration. The choice of indicators drew on technical assistance work done by the World Bank in Pakistan and Turkey, best practices shared by the tax administration in Denmark, the International Monetary Fund (IMF) recommendations for Senegal, and feedback from experts at the World Bank and the IMF. Our strategy is broadly applicable in other lower-income countries where similar constraints on algorithm design, relating to the communicability of the algorithm and the lack of historical audit results data, are likely to bind.

The algorithm's risk score aggregates two types of risk indicators at the firm level: inconsistencies and anomalies. Inconsistencies are *within-firm indicators* that flag taxpayers with inconsistent information across different datasets. For example, an inconsistency arises if the self-reported turnover is lower than the third-party reported turnover construct as the sum of exports, procurement payments, and purchases declared by other firms. In contrast, anomaly indicators are *across-firm indicators* that flag outlying behaviors potentially associated with tax evasion relative to the firm's peers. For example, one anomaly indicator flags firms with an extremely low profit margin relative to firms of the same economic sector and similar size. The inconsistency and anomaly indicators generate "points" which are aggregated using weights.[10] As mentioned above, it was not possible to estimate optimal weights given he lack of historical audits data, so we fixed weights based on expert judgment of the indicators' importance. We over-weighted the inconsistencies compared to anomalies to reflect the higher confidence that inconsistencies reflect non-compliance, while anomalies may also reflect temporary economic problems or poor management. The final algorithm includes four inconsistency indicators and six anomaly indicators to construct the risk score. Table A1 summarizes the steps to derive the risk score.

4.2 Experimental Design

We introduced algorithm selection in an experimental setting to evaluate the performance of algorithm audits compared to inspector-selected audits. The experiment introduced algorithm-selected audits in each list of audits, both for full and desk audits. The units participating in the experiment informed the total number of cases that each list should contain, and we agreed that they would select half of the cases (at discretion) and leave the other half to be selected by the algorithm. We ranked firms

[10]Each inconsistency and anomaly is captured in the form of a ratio, and numerical values are assigned based on the size of the ratio to take into account the severity of the irregularity.

by risk score within a tax office and selected the top cases on the list until the required number of cases was reached. In the experiment, each selection list (tax unit-specific for full audits and inspector-specific for desk audits) contained both inspector-selected and algorithm-selected cases. Our empirical analysis therefore consists of a "horse race" between the two case types.[11]

We randomized the order of case types displayed on the selection lists and asked inspectors to adhere to the proposed ordering. The ordering randomization tried to ensure that the two audit selection types were given a fair chance of execution and prevent the salience effects of cases at the top of the list from systematically favoring one type or the other.[12] For desk audits, in which the selection is individual by inspector, the algorithm-selected cases were randomly distributed across inspectors belonging to the same tax unit. The experimental introduction of algorithm cases in the audits lists enables the identification of the average effects of replacing discretionary selection with the algorithm.

We also took some steps to ensure compliance with the experiment, such as conducting workshops with inspectors in all tax units, producing reports in French with explanations about the algorithm and the indicators, and being available for any questions they might have. Managers also provided inspectors with a protocol emphasizing the importance of conducting all audit cases with the same rigor. In general, there was strong support from the hierarchy to implement the experiment, and although the experiment lasted from 2018 through 2020, DGID still asked us to produce algorithm selections for the years 2021 and 2022.

The experiment started with the Large and Medium Taxpayer Units in 2018 and included four regional tax offices in 2019 and 2020. In each year the algorithm's indicators and weights were slightly updated. Table xx shows the number of cases selected by year office and selection type, showing that the number of inspector-selected and algorithm-selected cases were approximately the same for most years and tax units. However, the relative number of algorithm and discretionary cases varied across years and units for several reasons explained in the Appendix.[13]

[11] This strategy is the best option in a context with few auditing units, in which it would not be feasible to randomize the use of the algorithm at the inspector or tax unit level. The combination of both selection methods within each unit conducting audits is also the most realistic policy, as fully replacing discretionary selection is risky from a revenue perspective and may face resistance among inspectors.

[12] Specifically, the order of cases was randomized across all selected cases for a unit/inspector in 2018. In 2019 and 2020, the first case on each list was inspector-selected, with subsequent cases alternating between the two selection methods. Cases were randomly allocated to slots on the list.

[13] One important reasons was the way we treated "overlapping cases", i.e. cases selected by inspectors but with high risk scores, were counted. In 2018, we flagged overlap cases so we can control for them specifically in the empirical analysis, but did not change the number of cases to be selected by the algorithm based on the overlap. In 2019 and 2020, we selected one additional algorithm case for each overlap case, to ensure that the total number of cases met the target. On the lists sent to inspectors, all overlap cases appear as selected by inspectors.

4.3 Additional Treatments for Desk Audits

The desk audit program has the advantages of being larger and having an individual inspector assignment. We took advantage of these features to introduce two different treatments in this list. First, in the years 2018 and 2019 we selected some cases at random (within the pool of firms of the respective tax unit) and included them in the audit lists. These random cases serve as a benchmark to assess the quality of the algorithm because inspectors were unaware that these cases were random and thought that they were algorithm cases.[14] This approach risked lowering inspectors' enthusiasm for algorithm cases if the random cases were consistently worse in terms of audit yields.

Second, the desk audit program is accompanied by an "information treatment" cross-randomized across all case types (inspector, algorithm, and random). The intervention gave case-specific information for two-thirds of desk audits: one-third of the cases received a summary report containing the three main risks flagged by the algorithm (e.g. abnormally low profit rate, turnover lower than third-party reported turnover); one-third received the same information plus excel spreadsheet with the firm's tax declarations and third-party data for the last four years (i.e. the data used by the algorithm); and the remaining third had no additional information. With this intervention, we aimed to test whether providing high-quality, readable information (one of the advantages of using data and algorithms) improves audit implementation and performance.

4.4 Empirical specification

We estimate the impact of changing audit selection methods on audit yields, especially the probability of a selected case being conducted. The introduction of the algorithm can impact audits through two main channels: i) by changing the set of selected firms, the algorithm affects the audit outcomes because of these firms' evasion behavior, and ii) by changing the set of cases assigned to inspectors, the algorithm affects how inspectors conduct the audits. For instance, inspectors' efforts or the composition of inspector teams might change (for full audits). Any difference in audit outcomes must be associated with one of these channels and cannot be due to changes in firm behavior because firms are not aware of how they are selected for audit and made their tax filing (and evasion) decisions before the audit selection took place.

As already alluded to before, our experimental design creates a natural and valid counterfactual to algorithm selection: the inspector-selected cases observed within each list. Because of the (roughly) equal number of inspector and algorithm-selected cases, the randomization or the cases' ordering in each list, and the randomized distribution of selected algorithm cases for inspectors within a tax unit, we can rule out selection bias issues. The systematic differences in audit outcomes between the two

[14] To avoid tax inspectors from ignoring the random cases, all cases that were not selected by tax inspectors themselves were labeled on the list provided to them as selected according to the "new selection method".

selection methods can therefore be interpreted as causal estimates of the impact of switching from one discretionary selection to the algorithm.

To evaluate how the introduction of the algorithm affects audit outcomes, we thus compute the difference in outcomes across the two types of audit selection, effectively running a "horserace" between them. To do this, we estimate a linear regression of audit outcomes on selection methods:

$$y_{i\ell} = \beta_0 + \beta_1 Algorithm_{i\ell} + \beta_2 Overlap_{i\ell} + \beta_3 Random_{i\ell} + \gamma_\ell + \varepsilon_{i\ell}, \quad (1)$$

where $y_{i\ell}$ is the outcome of an audit for case i selected in audit list ℓ, $\varepsilon_{i\ell}$ is a conditional mean zero error term. The $\gamma_{i\ell}$ term denotes the list of fixed effects. The audit lists are tax office-year specific for full audits, and inspector-year specific for desk audits. Therefore, this level of fixed effects allows us to control for variation across audit types, tax offices, and selection years.

$Algorithm$ is an indicator that the case was selected by the algorithm. The main coefficient of interest is β_1, which captures the difference in audit outcomes between algorithm-selected and inspector-selected cases, the omitted category. A fair comparison between the two selection methods requires that the number of cases selected with both methods is equal and that inspectors' performance incentives were the same for both case types.

The variable $Overlap$ indicates that the algorithm prioritized a case selected by the inspectors. These cases appeared in the lists as inspector-selected. The variable $Random$ is only relevant for desk audits, and indicates that the case was picked at random. These cases appeared in the list as algorithm-selected. These variables are useful to assess the heterogeneity of cases within each of the two main selection methods.

To perform inference, we estimate the standard errors of each case by clustering the error term at the level of the audit list. This procedure allows for the correlation of unobservable characteristics within the cases selected by a single inspector (for desk audits) or within a tax office (for full audits).

We are mainly interested in understanding how the introduction of the algorithm affected the probability of an audit being conducted, and the audit yields such as the probability of detecting positive evasion and the evasion amount. Because audit yield is only observed conditional on conducting the audit, only the β_1 for the probability of conducting the audit can be causally interpreted, whereas the remaining outcomes rely on selected samples. Our initial sample consists of the lists of planned desk and full audits in the years 2018-2020, and we disregard eventual additional audits that may have been conducted during these years despite not being initially programmed.

5 Results

We describe the results of the intervention in several steps. We first consider the impact of the selection mechanisms of audit execution, given that only a part of the programmed audits is ultimately implemented. Second, we consider the effect on audit outcomes, including whether evasion is detected, the amount of evasion detected, and whether the taxpayer accepts or disputes the audit outcomes. Third, we examine results on the duration and human resource cost of audits.

5.1 Main results

Audit Execution

Based on the inspectors' audit reports, we observe which cases on the list were started and which ones were not. We consider an audit as executed if the case was started, even if it did not lead to a fine. Overall, only 53% of cases selected for full audits and 33% of cases selected for desk audits are actually audited. The implementation rates vary across years but are not lower in later years. The incomplete implementation is hence not due to the fact that some audits are implemented after the year in which they are scheduled. We consider all audits implemented until the end of 2021, and merge this information with audit program lists for 2018. 2019 and 2020.

Table 4 summarizes the main results. Full audit cases selected by the algorithm are 18 percentage points less likely to be implemented than inspector-selected cases, corresponding to a 34% reduction compared to the mean implementation rate (column 1). "Overlap" case, which were selected by both the inspectors and the algorithm (but which appeared as "Inspector selected" on the lists) were 15 percentage points more likely to be implemented than cases selected only by inspectors.[15]

The results for the desk audit program are qualitatively similar, but the differences are much smaller (columns 2 and 3). Algorithm-selected desk audits were 4 percentage points less likely to be implemented, which corresponds to 13% of the mean. The point estimate is statistically significant only when we control for inspector fixed effects (column 3). Overlap cases are again more likely to be implemented, though the coefficient is not statistically significantly. Random cases are no less likely to be implemented than algorithm-selected cases, likely because the two case types appeared in the same way to inspectors (as selected by "new methods") and the expected return of a case would not ex ante be clear to inspectors.[16]

[15]On the other hand, replacement cases, which appeared on the bottom of the list sent to inspectors and which were explicitly marked as replacements for algorithm-selected cases that turned out to be void (e.g. taxpayers that had become inactive or were not reachable) were 47% less likely to be implemented than other algorithm-selected cases (coefficients not shown). These replacement cases were labeled as replacement cases (rather than algorithm x replacement) on the list sent to inspectors, but DGID audit lists did not previously include replacement cases. So it is reasonable to assume that inspectors perceived these cases as algorithm-selected cases that were also lower-ranked in the priority given their role as a replacement.

[16]Replacement cases are significantly less likely to be implemented but the magnitude of this effect is smaller than it is

The reduction in the probability of conducting audits is our main result. It means a case is less likely to be looked at when it is selected by the algorithm than when it is selected by the inspectors themselves. This result may be explained by the fact that inspectors inherently prefer the cases chosen at discretion, devoting more effort to these cases. On the other hand, inspectors may also be simply sorting cases based on observable characteristics regardless of their selection method, and algorithm cases are simply less interesting by that measure. We discuss these two hypotheses in section 6.

In what follows, we present evidence of how the cases selected by the algorithm differ in terms of several other outcomes, conditional on execution.

Audit yields

Table 4, columns 4-6, show the relationship between the selection method and the probability of evasion detection, conditional on implementation. Overall, 89% of full audits and 74% of desk audits detect some evasion. For full audits (column 4), algorithm-selected cases are 3.7 percentage points more likely to detect evasion, and overlap cases are 8.1 percentage points more likely. These effects are economically substantial, but only the point estimate on overlap cases is marginally statistically significant.

For desk audits (columns 5 and 6), algorithm-selected cases are 4.7 percentage points less likely to detect evasion, and overlap cases are 7.4 percentage points less likely to do so, although only the latter point estimate is marginally statistically significant. The point estimates on random and replacement cases are relatively precisely estimated zeros, suggesting that inspector selection does not perform better than random selection. Overall, these results indicate that conducted cases were similarly likely to detect evasion for full and desk audits. However, conditional on positive detection, we find that algorithm-selected full audits detect significantly smaller amounts of taxes evaded plus fines (columns 7-9), though the result is not significant for desk audits when controlling for inspector fixed effects (column 9).[17]

To summarize, we find that algorithm-selected audits are less likely to be implemented, but similarly likely to yield a detection of evasion when implemented. Conditional on detection, algorithm audits are associated with slightly smaller amounts of evasion, in the case of full audits only. In general, our results are starker for full audits and are much smaller for desk audits. This is consistent with the fact that full audits are more costly and typically involve higher stakes for both the administration and the taxpayer.

for full audits. This is because the share of replacement cases in the total number of cases is larger for desk audits than for full audits.

[17] Similarly, we do not detect statistically significant differences in the detected evasion rate, measured either as a share of liability or as a share of the firm's mean turnover across several years, by selection method. We show these results in the Appendix, in Tables A5 and A6).

5.2 Robustness tests

We now examine the robustness of our results across subsamples. This allows us to show that our results are not driven by shortcomings in the implementation of our intervention, and that there is little heterogeneity in the results across tax offices and time periods.

First, we tackle the issue that the number of inspector-selected and algorithm-selected cases in a list sometimes deviated from the intended 50-50 split. The most important deviation occurred in the Large Taxpayers Office, where in 2019 and 2020 the tax administration only allowed 30% of the full audit cases to be selected by the algorithm. Other deviations occurred when managers adjusted the number of cases selected at discretion slightly upwards or downwards. If both inspectors and the algorithm ranked cases by perceived risk and then selected cases following the rank order, the average riskiness is higher when a smaller number of cases is selected. We hence rerun our results for the following three subsamples: a) lists with an identical number of inspector-selected and algorithm-selected cases (as they appear to the researcher), b) lists with an identical number of inspector-selected and algorithm-selected cases (as they appear to the tax inspectors), and c) lists in which the number of algorithm cases is equal to or smaller than the number of inspector-selected cases. The latter subsample is less restrictive than the first two, but also less clean, as it may give the algorithm cases an advantage.[18]

Second, we recognize that there is wide heterogeneity across tax units and years. The Large Taxpayer Unit typically has the best inspectors and the lowest number of firms per inspector. Although these firms are more complex than smaller firms, inspectors are more familiar with them. Firms also have an expectation of being audited every four to five years. The association between audit selection method and outcomes may thus be different in the LTU than in other tax offices. We hence rerun our analysis excluding the Large Taxpayers Unit.

Third, the experiment's implementation differed in each of the three years. The year 2018 was the first one, and inspectors may have felt the cost of the "novelty" of the algorithm. In 2020, both firms and tax inspectors were affected by the pandemic. The 2019 implementation is the cleanest, also because the research team tightly monitored the execution of the experiment, including by asking inspectors to fill out special forms about the audit execution. We thus reestimate the results for the 2019 selection only.

Figure 3 depicts the coefficients on the algorithm selection indicator from Equation 1 (relative to inspector-selected cases) for the three main outcomes: the probability of starting the audit, the prob-

[18]The difference between sub-samples a) and b) arises from a change in the method we implemented over time. In 2018, we ran the algorithm to select a number of cases that was equal to the number of inspector-selected cases, allowing for overlap cases. The overlap cases appear to inspectors as inspector-selected cases. Thus, the list of "pure" algorithm cases on their list is lower than the number of inspector-selected cases. In 2019 and 2020, we corrected for this by adding one additional algorithm case for each overlap case.

ability of detecting positive evasion, and the amount of detected evasion (plus fines, in log points). Panel A is for full audits, and Panel B for desk audits. The two panels show the baseline coefficient (discussed above) with two types of standard error computation, and the various subsamples as discussed above and listed in the legend.

The figure shows that the point estimates are remarkably stable across subsamples, and the findings highlighted above are qualitatively and quantitatively robust for full audits. Algorithm cases are less likely to be conducted, no more likely to yield a detection of evasion (though point estimates are consistently positive), and conditional on detection, algorithm cases exhibit significantly smaller amounts of evasion. For desk audits, however, the point estimates are almost all very small and statistically indistinguishable from zero. The negative association between algorithm selection and audit execution is present only in the full sample, but disappears when we limit the analysis to lists with more comparable numbers of algorithm and inspector-selected observations. Our inability to detect statistically significant differences for desk audits is not because of a lack of power. In fact, the number of desk audits is higher and the variation of outcomes among these cases smaller than for full audits. Our experimental design also allows a within-inspector comparison.

5.3 Other differences in audit outcomes

Time and Human Resource Costs

Table 6 examines how the selection method affects the effort and human resources required for audits, and the audit length. These analyses are conditional on audit implementation. The most accurately captured measure of audit cost is the number of inspectors working on an audit, which is reported in the administrative data. Column 1 shows that algorithm-selected (full) audits require teams that are 10% smaller than the average team of 2.9 members.

The remaining columns show the association between audit selection and three different measures of audit length. Overall, algorithm audits appear shorter and require less inspector time, although most point estimates are statistically insignificant. In columns 2-3, the outcome is the self-reported number of days that inspectors work on a case. For full audits, we multiply the reported number of days by the number of inspectors working on a case. Full audits require 188 inspector-days while desk audits require only 9 days on average. Although the point estimates on the algorithm dummy are negative, we detect no statistically significant difference between algorithm and inspector cases, possibly because the number of days spent on the case is noisy. For full audits, it is reported for only 51 cases, and it is unclear whether inspectors were referring to their own time investment or the investment of the full team when providing the information.

In columns 4-5, we use as outcome the audit duration as measured by the difference between the date of notification and the audit start date. The measure captures the length of an audit but not

necessarily the time inspectors actively spend working on the case, as they might work on several cases simultaneously.[19] The average full audit takes 157 days until notification. Algorithm-selected audits are completed 20 days faster. Desk audits are only slightly shorter, taking on average 113 days, and algorithm-selected cases take 15 days longer. In both cases, the difference between the algorithm and inspector-selected cases is not statistically significant.

Column 6 shows the results for the audit duration as reported by firms subject to a full audit.[20] The mean here is very different from the number of days from start to finish (28 days vs 157 day), which is not surprising as the firm would use the number of days that auditors physically worked in its premises to estimate the audit duration. Algorithm-selected audits take significantly fewer days, completing in 19.5 days on average.

Taxpayers' perceptions

An alternative measure of inspectors' effort and performance comes from the taxpayers' point of view. We surveyed taxpayers in the Dakar area and asked them about their impressions of interactions with the tax authority. Table 8 shows the regression results for the taxpayers' responses conditional on the selection method that led to their being audited. Panel A uses the sample of all interviewed firms that were also in the audit selection, and Panel B conditions on firms that had a conducted audit.

Columns 1 and 2 show the results of the evaluation given by the taxpayer to the last interaction with the tax authority during a full audit. This variable is the means of three grades (from 0 to 10) given by taxpayers to the inspectors' technical knowledge, honesty, and efficiency. Firms selected by the algorithm reported lower grades, meaning that they had an overall poorer opinion of their last interaction with the tax authority. However, the results are only marginally significant for full audits and no longer significant when we condition on firms with conducted audits (Panel B). Firms that were "overlap" cases reported higher evaluation grades of inspectors.

Columns 3 and 4 report results for the taxpayers' agreement with the statement that "Inspectors manage to uncover evasion during a full audit", where they were able to give 1 to 5 answers that ranged from "I strongly disagree" to "I strongly agree." The mean of this variable is around 4.1, suggesting that taxpayers tend to strongly agree with the statement. However, algorithm firms again had negative coefficients, though only marginally significant for full audit cases in Panel A. Overlap cases reported stronger agreement with the statement.

Finally, Columns 5 and 6 have the coefficients for the regressions of the perceived incidence of corruption among full audits. Taxpayers were asked about their beliefs about the percentage of full audits

[19] As discussed above, a contest and lengthy negotiations of audit results can extend the duration between notification and confirmation.

[20] The sample includes firms that we can observe as having been subject to a full audit in our administrative data, although the question asks firms about the duration of the *last* full audit, so they may or may not be referring to the one that we have in our data.

that end in corruption in Senegal. Their staged mean was 14% among firms selected for full audits and 30% for firms selected for desk audits. The results do not suggest statistically significant differences across selection methods, but similarly to the other outcomes, algorithm cases show negative coefficients for full audits, and overlap cases report higher corruption perceptions. The coefficient on overlapping cases is strongly significant on column 6, Panel B.

The taxpayer survey results reinforce the pattern that inspectors seemed to be less invested in algorithm cases. Given that the questions asked about interactions in "full audit" interactions, it is not surprising that the coefficients among firms selected for desk audits are mostly small and not significant. For the same reason, it is difficult to explain the strong coefficient found for corruption among conducted desk audit cases, and it is possible that this result is a spurious correlation.

Dispute of Audit Outcomes

Another dimension of tax audits that is relevant for tax authorities is the extent to which taxpayers push back on audit findings. Once an audit has detected tax evasion, the taxpayer is provided with a notification and the opportunity to dispute the audit results. After dispute and negotiation with the tax administration, a confirmation of the audit results is issued. The confirmed amount of evasion is usually lower than the notified amount. The confirmation amounts match the notification amounts only in 22% of conducted full audits and 29% of conducted desk audits. Having differences between confirmation and notification may reflect that audits were poorly conducted or collusive behavior between taxpayers and inspectors to alleviate the penalties.

We compute the confirmation amount as a share of the notification amount for the audits in our sample, conditional on having positive detection. We plot the distribution of this share for inspector- and algorithm-selected cases, separately for full and desk audits. A large amount of cases have a substantially low confirmation amount relative to the initial notification. For full audits, the mean and median confirmation values are 40% and 18% of the notification value. For desk audits, these figures are 62% and 58%. We analyze whether the intensity of these revisions change for algorithm and inspector cases.

Table 5 shows the regression results of the probability that the notification matches the confirmation on selection methods, as well as the notification and confirmation values themselves. It shows that conducted full audit cases were substantially more likely to have identical confirmation and notification amounts when the algorithm selected them. However, the algorithm notification amounts also tended to be smaller, as mentioned previously in the main results. No differences are detected for desk audits, in line with previous results.

The same results can be seen graphically in Figure 4, which shows the distribution of the confirmation-notification ratio for inspector- and algorithm-select cases among full audits (the figure excludes cases with no confirmation, which represent about a third of these cases). The figure shows that the density

at 100% is noticeably different between algorithm and inspector cases, with algorithm cases being more likely to have identical values. For desk audits, there are no significant differences between the algorithm and inspector-selected audits (Panel B).

6 Mechanisms

As shown in the main results, the algorithm introduction reduced the probability of audit implementation, especially for full audits and, to a lesser extent, for desk audits. Our additional results suggest that algorithm audits finished with lower fines used fewer human resources, took less time, and faced less dispute from taxpayers. It is unclear, however, why switching to the algorithm reduced the execution rate and whether the other differences in audit outcomes are due to different types of firms selected by the algorithm or simply different levels of effort put by inspectors into these cases.

We put forward there are three testable reasons why inspectors may reduce their execution rates for algorithm cases: i) targeting: the algorithm selects firms whose characteristics determine the worse outcomes; ii) justification: inspectors had difficulty understanding how to carry out algorithm cases for lack of information about the cases; and iii) effort: inspectors systematically put less effort into algorithm cases. We investigate each of these cases by exploiting the characteristics of selected firms, by analyzing the impact of the information treatment, and by exploiting a discontinuity in the algorithm selection.

6.1 Targeting: Characteristics of selection

One possible explanation for the under-performance of algorithm cases relative to discretionary cases is that the algorithm cases were deemed uninteresting to inspectors. If this is the case, then the differences in the probability of starting a cases could be attributed to the characteristics of the firms selected by the algorithm. The lower rate of execution could, therefore, reflect that these firms had particular features that reduced the inspectors' willingness to audit them. In this case, the execution gap between algorithm and inspector cases would not reflect any animosity towards the algorithm but simply the application of the same decision rule to a different set of firms.

The first test of the algorithm's targeting ability is to assess the quality of its selecting variable: the risk score. Is the risk score systematically favoring uninteresting firms to the inspectors, or low evasion firms? We test that hypothesis by assessing the correlation between the risk score and audit outcomes within the sample of inspector-selected cases. Restricting the sample to the inspector-selected isolates the quality of the risk score, since it is possible that inspectors were behaving differently toward algorithm cases. Table 11 shows that among inspector-selected cases, there is a clear positive correlation between the risk score and the audits' outcomes. However, the correlation is non-existent in the total sample, including discretionary and algorithm-selected cases (Table 11), reflecting the impact of the

negative behavioral impact of the algorithm on inspectors' ability to carry out the cases. The result suggests that the risk score itself is not negatively correlated with audit outcomes, absent other effects of the algorithm on inspectors' behavior.

A richer way to exclude the targeting hypothesis is to use the characteristics of selected firms to predict whether a case is conducted. If inspectors decide to open or not a case based on observable characteristics, such as firm size, profitability, or sector of economic activity, these characteristics may be driving the differences between algorithm cases and discretionary cases. To test this hypothesis, we estimate the inspectors' decision-making rule by estimating the conditional choice probabilities of conducting cases based on a rich set of controls. The controls are the distances from the firms to the tax office, dummies for the selection list (year x tax office for full audits, year x inspector for desk audits), turnover, profit rate, and payroll in each of the three years before the audit. To give more flexibility to the functional form, we include these variables as dummies of the firm decile in the variable distribution, computed at the tax office level. We use several different prediction models and train them on the sample of inspector-selected cases: linear probability model (OLS), logit, LASSO, Random Forest, and a Random Forest using the continuous variables as predictors (instead of the deciles). We then use the models trained on the inspector-selected cases to predict the probability of conducting algorithm-selected cases. The predicted probabilities correspond to how inspectors would behave if they used the same decision-making rule for conducting algorithm-selected cases as they did for inspector-selected cases.

We plot in figure 5 the predicted mean rate of audit execution estimated using the sample of inspector-selected cases, according to various models. Panel A shows the mean prediction for full audits, and Panel B for desk audits. The horizontal lines show the actual mean execution for each case.[21] Unsurprisingly, there is not much difference between predicted and executed rates among inspector-selected audits since the models were trained exclusively within that sample. However, as we extrapolate the predictions to the sample of algorithm-selected cases, the right-hand side of Panel A shows that there remains an important gap between predicted rates and executed rates. This gap means that observable characteristics cannot fully account for the negative effect of the algorithm selection on execution. For desk audits, where the difference was already much smaller, the models present a heterogeneous behavior, with simpler, more parametric models displaying a gap between prediction and realization, and Random Forest models suggesting the observable characteristics explain well the mean execution.

We also run the baseline regression (equation 1) to compute the effect of including the algorithm cases in the audit program conditioning on the predicted probability according to several model. The results are shown in Table 13 for full and desk audits, using the predicted probability as computed by different models. The predicted probabilities are always strongly correlated with the probability

[21] Notice that since the plots an unconditional mean execution, so these values should not be directly compared to estimate differences between algorithm and inspector selected execution. The appropriate comparisons are the ones at the list level, which we presented in the Results section of this paper.

of conducting the audit, and much more so for the random forest models than for the logit model. Moreover, the inclusion of these controls reduces the magnitude of the algorithm's negative effect on execution, suggesting that observables partly explain the unwillingness to conduct algorithm cases. From the baseline effect of -18.3 percentage points for full audits and -4.3 percentage points for desk audits (columns 1 and 3 of main results Table 4), controlling for the predicted probability of execution estimated with the random forest model, the effect of the algorithm becomes -13.3 percentage points and -2.5 percentage points (columns 5 and 6 of table 13). The coefficient on the algorithm is not statistically different from zero for desk audits, which is consistent with the view that the underlying characteristics of the cases explain all the execution gaps between discretionary and algorithm cases for desk audits. Nevertheless, among full audits, inspectors still seem to have penalized algorithm cases for other reasons.

6.2 Information

The second reason why the algorithm might have performed worse is because inspectors had trouble carrying out the cases. Such difficulty is warranted since the selection was automatized, and inspectors did not take part in it. Consequently, inspectors could be surprised by the selection and struggle to understand its rational. To test this hypothesis, we exploit the effect of the randomized information treatment on the audit performance. This treatment provided detailed information for inspectors on selected cases about the main indicators used for the selection, and (for a subset) a user-friendly spreadsheet containing the full information available about the selected firm's tax declarations and third party data. The goal of this information treatment was to help inspectors carry out the audits, which might be particularly important for automatically selected cases.

To estimate the effect of the randomized information treatment, we adapt the specification of equation 1 by including interaction terms between the selection methods and an indicator function for the information treatment. This specification allows us to retrieve the treatment effect of the information treatment and the heterogeneous effects it may have, depending on the selection method. The estimating equation is as follows:

$$\begin{aligned} y_{i\ell t} =& \alpha_0 + \alpha_1 Algorithm_{i\ell t} + \alpha_2 Overlap_{i\ell t} + \alpha_3 Random_{i\ell t} \\ &+ \alpha_4 Information_{i\ell t} + \alpha_5 Algorithm_{i\ell t} \times Information_{i\ell t} \quad + \gamma_\ell + \mu_t + \epsilon_{i\ell t} \end{aligned} \tag{2}$$

Since the information treatment was randomized for the selected cases, regardless of the selection method, average treatment effect is identified by comparing the average outcome among treated cases with the average outcome of the control cases. This comparison between means is essentially what the estimation of equation 2 performs. We run balancing tests (Table A4) to ensure that the two groups

are indeed similar along relevant characteristics.[22] Roughly half desk audit cases received some form of the information treatment.

Table A17 shows the results for the main audit outcomes. The randomized information treatment had no effect on the probability of starting a case or subsequent outcomes. Contrary to our prior, the information treatment did not help inspectors in the algorithm cases. The negative effect of algorithm selection on the main audit outcomes remains significant, whereas the interaction with information treatment has a small and statistically insignificant coefficient. Columns 4, 5, and 6 show the information treatment separated by the cases with "only indicators" and the cases with "indicators and data". There was no effect in either version of the intervention. The information treatment did not have any effect, and therefore, lack of information was not the driving force behind the negative effect on audit outcomes in the algorithm list.

6.3 Inspector ability

Our third potential mechanism is that switching selection methods requires the buy-in and ability of inspectors. On the one hand, the average effect of the change led to a reduction in audit execution, but it could be that the algorithm would have a better performance among inspectors with higher training, more experience, and stronger enthusiasm for using algorithm selection. To test this hypothesis, we analyze the heterogeneous performance of inspectors based on characteristics collected in the inspector survey conducted at the beginning of the study. In the survey, we collected information about inspectors' education levels (e.g., if they had a Master's or PhD degree), their experience at the tax authority (i.e., number of years and months at the institution), and how positively they saw the use of algorithms to automate audit selection. We then analyze how the algorithm performed differently from discretionary cases in a specification that includes interactions with these characteristics. This analysis can only be done for desk audits because these audit assignments are personalized for each inspector, whereas for full audits, there are no inspectors in charge of the case at the assignment stage.Remember that these audits are assigned at the tax unit level, and carried out by groups of inspectors. Therefore, the selection list does not contain the name of any inspector, and we only know the inspector characteristics of the cases that were carried out.

Table 10 shows the results for the three main outcomes. Notice that the sample size falls relative to the main results for desk audits (Table 4) because not all inspectors were interviewed in the survey. The results show that only inspectors' experience matters among the three analysed heterogeneity dimensions. Algorithm cases assigned to inspectors with more than the median number of months at the tax authority (the median was 99) are 15% more likely to be conducted, though these highly ex-

[22] Despite the randomization, small samples may present differences along relevant characteristics between control and treatment groups, biasing the estimate of the average treatment effects. Table A4 shows regressions of three outcomes on the information treatment: turnover, profit rate, and payroll. All variables were computed according to the 2018 tax declarations, to ensure that they were not affected by the intervention, which started in 2018.

perience inspectors in general have lower execution rates overall by 11%. Moreover, algorithm cases assigned to highly experienced inspectors were also 9.9% more likely to detect evasion, though this coefficient is only significant at the 10% significance level. Including inspector characteristics does not help explain the probability of detecting evasion or the amount of detected evasion. However, including inspector characteristics substantially increases the magnitude of the coefficient on algorithm cases from approximately -6% to -15%.[23] This result suggests that the treatment effect of switching selection methods is indeed heterogeneous across types of inspectors, and while inspectors seem to have displayed a lower probability of conducting algorithm cases, highly educated inspectors were less inclined to do so.

7 Conclusion

This paper assesses the impact of introducing automated selection in a tax enforcement agency. Collaborating with the Senegalese tax administration on an intervention at scale, we designed a selection algorithm based on risk indicators, and introduced automated selection in the audit lists during the years 2018, 2019, and 2020. We then compared the implementation and return of audit cases selected by the risk-scoring algorithm to cases selected by tax inspectors based on a traditional discretionary procedure. We find that algorithm audits were substantially less likely to be conducted relative to inspector-selected cases, especially among full audits. Among conducted audits, they were similarly likely to detect evasion, but algorithm-selected full audit detected smaller amounts of evasion. We also study the differences in several characteristics of the audits, showing that, among the conducted cases, algorithm cases used less human resources, were shorter, had less infractions, and had infractions in fewer years.

We test whether the worse performance of the algorithm is due to underlying characteristics of the cases, low levels of information about the selected cases, or inspector characteristics. Using a rich set of observable characteristics, we still cannot rule out that inspectors were less likely to carry out algorithm cases even for similar levels of observable characteristics. We tested the information hypothesis by providing inspectors with additional information in a random set of cases, but the additional information has not impact on audit implementation or other outcomes. Finally, we find that heterogeneity in inspector experience accounts partly for the lower performance of the algorithm, with highly experienced inspectors being less likely to "penalize" the algorithm.

Our study sheds light on the important issue of introducing "good practices" into a government organization, an particularly into an enforcement agency. Despite its initial apparent advantages, the automation of audit selection using evasion indicators faced strong execution hurdles and seems to have benefited from a reduced effort by inspectors in the field. Moreover, even among desk audits, in

[23]The effect of the algorithm on audit implementation among desk audits is -4.3% on Table 4. However, this result increases in magnitude to -6% in the sample of cases for which we have inspector characteristics data.

which inspectors executed algorithm and discretionary audits at similar rates, there was no discernible advantage of using the algorithm to select audits. Our study suggests that, at least in the short run, the reduction of discretion in audit selection fails to produce benefits.

References

Banerjee, Abhijit, Esther Duflo, Clement Imbert, Santhosh Mathew, and Rohini Pande (2020). "E-governance, accountability, and leakage in public programs: Experimental evidence from a financial management reform in india". In: *American Economic Journal: Applied Economics* 12.4, pp. 39–72.

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu (2018). *The Glittering Prizes: Career Incentives and Bureaucrat Performance*. Tech. rep. mimeo.

Besley, Timothy and Torsten Persson (2013). "Taxation and development". In: *Handbook of public economics*. Vol. 5. Elsevier, pp. 51–110.

Brockmeyer, Anne, Spencer Smith, Marco Hernandez, and Stewart Kettle (2019). "Casting a wider tax net: Experimental evidence from costa rica". In: *American Economic Journal: Economic Policy* 11.3, pp. 55–87.

Bullock, Justin B (2019). "Artificial intelligence, discretion, and bureaucracy". In: *The American Review of Public Administration* 49.7, pp. 751–761.

Carrillo, Paul, Dina Pomeranz, and Monica Singhal (2017). "Dodging the taxman: Firm misreporting and limits to tax enforcement". In: *American Economic Journal: Applied Economics* 9.2, pp. 144–64.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan (2018). "The value of regulatory discretion: Estimates from environmental inspections in India". In: *Econometrica* 86.6, pp. 2123–2160.

Finan, Frederico, Benjamin A Olken, and Rohini Pande (2017). "The personnel economics of the developing state". In: *Handbook of Economic Field Experiments*. Vol. 2. Elsevier, pp. 467–514.

Gerardino, Maria Paula, Stephan Litschig, and Dina Pomeranz (2020). "Distortion by Audit". In:

Glaeser, Edward L, Andrew Hillis, Hyunjin Kim, Scott Duke Kominers, and Michael Luca (2021). "Decision Authority and the Returns to Algorithms". In:

Glaeser, Edward L, Andrew Hillis, Scott Duke Kominers, and Michael Luca (2016). "Crowdsourcing city government: Using tournaments to improve inspection accuracy". In: *American Economic Review* 106.5, pp. 114–18.

Khan, Adnan Q, Asim I Khwaja, and Benjamin A Olken (2015). "Tax farming redux: Experimental evidence on performance pay for tax collectors". In: *The Quarterly Journal of Economics* 131.1, pp. 219–271.

Khwaja, Muwer Sultan, Rajul Awasthi, and Jan Loeprick (2011). *Risk-based tax audits: approaches and country experiences*. The World Bank.

Kleven, Henrik Jacobsen, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez (2011). "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark". In: *Econometrica* 79.3, pp. 651–692.

Okunogbe, Oyebola Motunrayo and Victor Pouliquen (2018). "Technology, taxation, and corruption: evidence from the introduction of electronic tax filing". In: *World Bank Policy Research Working Paper* 8452.

Rasul, Imran and Daniel Rogger (2018). "Management of bureaucrats and public service delivery: Evidence from the nigerian civil service". In: *The Economic Journal* 128.608, pp. 413–446.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian (2001). "Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota". In: *Journal of public economics* 79.3, pp. 455–483.

Slemrod, Joel and Tejaswi Velayudhan (2018). "Do firms remit at least 85from India". In: *Journal of Tax Administration* 4.1. ISSN: 2059-190X. URL: http://jota.website/index.php/JoTA/article/view/163.

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

Tables

Table 1: Number of Firms by Data Source

		2014	2015	2016	2017	2018	2019	2020
A Self reported	CIT	5136	5969	6218	6594	6720	7233	0
	VAT	11181	11901	12699	13352	13969	14213	13538
	PAYE	7061	7518	7870	8513	8782	9005	8621
	CGU	1581	1827	2026	2203	2650	2671	2801
	TAF	86	105	112	122	121	111	118
B Third party	Imports	8963	12427	13068	11859	13551	13677	10591
	Exports	1398	1724	1881	1824	1697	1659	1558
	Procurement	809	735	1380	1340	1903	1897	1684
	VAT annexes	6	9	21	805	3606	3209	0
C Audits data	Digitized	0	0	1	3294	2753	2946	3714
	Self-reported (Excel)	0	0	0	102	561	664	51

Notes: This table shows the number of taxpayers (firms) in the main datasets used in the algorithm and the analysis. This table is discussed in Section 3.

Table 2: Count of selected audits by year, tax office, and selection method

		Algorithm	Discretion	Random	Replacement	Total
LTU	2018	153	175	60	7	395
	2019	25	96	0	0	121
	2020	110	314	0	85	509
SME	2018	183	193	83	6	0
	2019	172	194	52	23	0
	2020	88	88	0	38	0
Liberal	2018	101	110	59	2	272
	2019	81	85	26	12	204
	2020	90	86	0	70	246
Regional	2018	0	0	0	0	0
	2019	206	215	84	41	0
	2020	268	278	0	237	0
All	2018	437	478	202	15	1132
	2019	484	590	162	76	1312
	2020	556	766	0	430	1752
Total		1477	1834	364	521	4196

Note: Number of selected cases (full audits and desk audits) by year and tax office. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion. Random cases and replacement cases are exclusive to desk audits. This table is discussed in Section 3.

Table 3: Firm characteristics of algorithm vs discretionary selection

	Tax declarations				Administrative			Taxpayer Survey	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log(turnover)	log(payroll)	profit rate	P(trade)	Duration trip	Age	N. employees	% sales in cash	Audits frequency
Algorithm	-1.323***	-2.131***	-0.0261**	0.0152	-3.605	1.475***	1.636	8.282**	-0.0981
	(0.276)	(0.312)	(0.0117)	(0.0190)	(15.70)	(0.393)	(9.400)	(3.744)	(0.164)
Inspectors x Overlap	-0.0915	-0.114	-0.0172	0.0678*	75.39*	2.405***	-28.05	6.329	0.0664
	(0.505)	(0.565)	(0.0205)	(0.0377)	(45.27)	(0.853)	(19.71)	(13.95)	(0.586)
Algorithm x Random	-0.247	0.439	0.00344	-0.0857***	-17.54	-1.149*	-9.011	3.438	0.136
	(0.433)	(0.508)	(0.0152)	(0.0302)	(17.01)	(0.607)	(7.805)	(5.932)	(0.308)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	3310	3310	2976	3585	2775	3153	671	680	617
R2	0.207	0.232	0.0906	0.258	0.0714	0.127	0.256	0.198	0.169
Mean outcome	17.20	12.05	-0.03	0.40	66.79	13.67	32.52	49.85	2.76

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. This table depicts the regression coefficients of characteristics on selected methods, using the sample of selected cases across the three years of the experiment, including full and desk audits. The table shows OLS results with fixed effects at the list level (year X tax office for full audits, year X inspector for desk audits). Standard errors are clustered at the list level. The characteristics of the firms stem from three sources. From the tax declarations, we use the log of the yearly declared turnover, the log of the yearly declared payroll, the profit rate, and the probability that the firm trades with foreign countries. All these variables refer to the firms' declarations one year before selection to prevent them from being affected by an audit. Moreover, we use administrative data on the firms to compute the duration of the trip from the tax authority's premises to the firm's address, which we computed using GoogleMaps. We also compute the firm's age based on its date of creation. Finally, we use the taxpayer survey to compute the (self-reported) number of full-time employees, the share of total sales done in cash, and the perceived yearly frequency of full audits. This table is discussed in Section 5.

Table 4: Results of main outcomes

	P(execution)			P(detection | execution)			log(evasion) | detection		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.183***	-0.0447	-0.0431**	0.0371	-0.0465	-0.0441	-0.579*	-0.178	-0.0818
	(0.0369)	(0.0289)	(0.0208)	(0.0309)	(0.0367)	(0.0303)	(0.283)	(0.168)	(0.152)
Inspectors x Overlap	0.158**	0.0393	0.0348	0.0809*	-0.0744*	-0.0756	0.0689	-0.213	-0.0264
	(0.0704)	(0.0407)	(0.0351)	(0.0441)	(0.0367)	(0.0464)	(0.312)	(0.428)	(0.350)
Algorithm x Random		0.00666	0.00639		0.00284	0.00534		-0.0921	-0.0836
		(0.0396)	(0.0310)		(0.0417)	(0.0331)		(0.131)	(0.172)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	959	3237	3237	507	1016	997	453	751	732
R2	0.262	0.243	0.324	0.151	0.324	0.430	0.214	0.271	0.402
Mean outcome	0.53	0.33	0.33	0.89	0.73	0.73	18.84	17.42	17.39

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included. This table is discussed in Section 5.

Table 5: Dispute of Audit Outcomes

	P(conf.=not.)		log(not.)		log(conf.)	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	0.121**	0.0101	-0.349*	-0.160	-0.191	-0.193
	(0.0433)	(0.0519)	(0.202)	(0.173)	(0.258)	(0.236)
Inspectors x Overlap	-0.0310	0.0291	0.665	-0.569	0.543	-0.0598
	(0.0651)	(0.1000)	(0.401)	(0.447)	(0.385)	(0.479)
Algorithm x Random		-0.00779		-0.113		0.0537
		(0.0954)		(0.261)		(0.249)
N	264	372	260	346	260	346
R2	0.171	0.235	0.325	0.547	0.189	0.509
Mean outcome	0.15	0.21	19.46	17.96	18.68	17.33

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). The sample includes all cases with **non-missing confirmation and notification**, and for the second and third outcome, the sample is conditional on **both notification and confirmation being non-zero**. Standard errors are clustered at the list level. This table is discussed in Section 5.

Table 6: Resources Allocated to Audit Execution

	Number of agents	N Days Working on Case (Self-reported)		N Days From Open to Close (Admin Data)		N Days Duration (Firm Survey)
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Full audits	Desk audits	Full audits	Desk audits	Full audits
Algorithm	-0.211**	-7.226	-3.500	-26.03*	2.570	-8.659**
	(0.0795)	(18.76)	(2.401)	(14.96)	(18.96)	(3.956)
Inspectors x Overlap	0.0187		-22.76	-10.39	25.96	-26.52
	(0.133)		(18.24)	(35.33)	(37.59)	(26.90)
Algorithm x Random			-2.038		2.934	
			(1.711)		(19.47)	
N	507	51	108	285	238	214
R2	0.235	0.263	0.568	0.246	0.355	0.255
Mean outcome	2.87	188.03	9.06	164.44	124.27	28.16

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of number of agents and audit duration (in days) on the selection method of the case. The sample includes all cases selected and started in the audit programs of 2018, 2019, and 2020, for which we had data on the outcome variables. We use three measures of audit duration. The number of agents is obtained from the audit records, and represents the number of enforcement agents involved in the case. The first measure of duration is the self-reported mean number of days worked on the case, and it is based on a spreadsheet that inspectors needed to fill out during the audit program of 2019. We multiplied the answer by the number of inspectors that worked in the cases, obtained from the audit reports. This measure is only available for 2019. The second measure is a proxy for the duration of the inspection work based on administrative data. It is computed as the difference between the date of notice of infraction and the date of the start of the audit (either verification announcement or information request or start date of audit). It can only be computed for cases that had a first notice. The third measure is the number of days the audit took as reported by respondents in a survey of firms subject to full audit. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year (list) levels, and inspector x year levels when inspector-fixed effects are included. This table is discussed in Section 5.

Table 7: Audit outcomes of algorithm vs discretionary audits using thoroughness of investigation

	N. years		N. infractions		Fine/Evasion	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.701***	-0.235***	-0.227	-0.215**	0.0111	0.0201
	(0.118)	(0.0663)	(0.138)	(0.0879)	(0.0156)	(0.0135)
Inspectors x Overlap	0.892**	-0.0215	0.303	-0.240	0.0529	-0.0460
	(0.350)	(0.102)	(0.210)	(0.173)	(0.0370)	(0.0341)
Algorithm x Random		0.0967		0.0192		0.0188
		(0.101)		(0.0930)		(0.0241)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	Yes	No	Yes	No	Yes
N	944	2731	507	997	294	392
R2	0.230	0.177	0.149	0.327	0.182	0.309
Mean outcome	1.67	0.70	2.17	1.39	0.40	0.41

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case, controlling for list fixed effects (year x tax office for full audits, year x inspector for desk audits). The outcomes are extracted from the audit reports. The three outcomes are i) the number of years from the taxpayers' declarations in which the inspector found an infraction (notice that the inspection can investigate tax declarations up to four years before the audit date according to Senegalese law), ii) the number of different infractions found by the inspector, and iii) the severity of the infraction as indicated by the ratio of fine to the evaded amount. Standard errors are clustered at the list leve. Sample is conditioned on cases that started an audit. This table is discussed in Section 5.

Table 8: Selection method and survey results

Panel A: All interviewed firms

	Evaluation		Detection		Corruption	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.397*	-0.166	-0.500*	-0.372	-6.536	1.075
	(0.195)	(0.263)	(0.268)	(0.276)	(4.230)	(4.587)
Inspectors x Overlap	0.550**	0.417	0.956***	-0.239	4.917	15.21
	(0.263)	(0.333)	(0.161)	(0.976)	(11.28)	(12.62)
Algorithm x Random		0.0241		0.0272		-6.233
		(0.385)		(0.352)		(5.329)
N	216	255	254	369	159	204
R2	0.138	0.176	0.131	0.158	0.140	0.303
Mean outcome	6.70	6.65	4.20	4.20	14.93	16.66

Panel B: Only audited firms

	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-0.256	-0.0271	-0.436	-0.0891	-6.861	-2.260
	(0.424)	(0.327)	(0.525)	(0.384)	(4.903)	(8.667)
Inspectors x Overlap	0.617***	0.721*	0.351	-0.467	5.285	41.02***
	(0.188)	(0.397)	(0.243)	(1.710)	(13.49)	(11.35)
Algorithm x Random		0.889*		-0.134		-0.0823
		(0.485)		(0.503)		(9.150)
N	108	126	116	179	81	99
R2	0.161	0.238	0.0687	0.194	0.0920	0.315
Mean outcome	6.83	6.68	4.15	4.13	11.04	14.20
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	No	No	No	No

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results with fixed effects at the year x tax office level for outcomes from the taxpayer survey. Notice that given the low number of observations in the taxpayer survey, it is not possible to run the regression with year x inspector fixed effects as is done with outcomes extracted from administrative data. The sample in Panel A shows the results for all selected firms interviewed in the survey, whereas Panel B restricts to firms that were audited according to the audit reports. Standard errors are clustered at the year x tax office level. The meaning of outcomes if the following: "Evaluation" is the average grade (0 to 10) given by the taxpayer to their latest interaction with inspectors regarding their technical preparation, honesty, and efficiency; "Detection" is the degree of agreement (from 1 to 5 where 1 means "I strongly disagree" and 5 means "I strongly agree") with the statement that "Inspectors manage to uncover evasion during a full audit"; and "Corruption" is the declared belief of the percentage of audits in Senegal that end in corruption. This table is discussed in Section 5.

Table 9: Inspectors' characteristics among conducted cases

Panel A: Selected audits

	Age		Masters/PhD		Enthusiasm for Alg.		High experience	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Full audits	Desk Audits	Full audits	Desk Audits	Full audits	Desk Audits	Full audits	Desk Audits
Algorithm		0.107**		-0.00557		0.00431		0.0132
		(0.0379)		(0.00739)		(0.0126)		(0.0165)
Inspectors x Overlap		0.525		-0.00792		-0.00551		0.0398
		(0.400)		(0.0529)		(0.0450)		(0.102)
N		1906		1272		1272		1272
R2		0.354		0.397		0.0566		0.114
Mean outcome		34.99		0.69		0.82		0.42

Panel B: Conducted audits

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Algorithm	-0.182	0.203	0.0205	-0.0172	-0.0326	0.0481**	0.0163	0.0162
	(0.309)	(0.362)	(0.0207)	(0.0177)	(0.0260)	(0.0221)	(0.0275)	(0.0286)
Inspectors x Overlap	0.306	-1.035	0.0244	-0.124**	-0.00754	0.0453	-0.0466	0.0375
	(0.582)	(0.726)	(0.0484)	(0.0465)	(0.0530)	(0.0545)	(0.0447)	(0.0510)
N	459	884	423	599	423	599	423	599
R2	0.360	0.400	0.321	0.134	0.335	0.135	0.531	0.345
Mean outcome	37.02	35.39	0.81	0.80	0.82	0.90	0.39	0.28

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the differences in inspector characteristics for selected and conducted cases. It is not possible to compute the characteristics of inspectors for full audits at the selection stage because at this stage they have not yet been assigned to individual inspectors. The age of the inspectors is extracted from administrative data, and the other characteristics from the taxpayer survey. The variable "Masters/PhD" indicates whether the inspector has a Masters or PhD degree. "Enthusiasm for Algorithm" indicates that the inspector answered "I agree" to the statement that "It would be useful to have automated selection for audit cases". "High experience" means that the inspector had more experience at the tax authority than the median (the median was 99 months). For Panel A, we show results at the selection stage for desk audits, using information about the inspector to which the case was assigned. In Panel B, we use the information about the (teams of) inspectors that conducted the audits, taking the mean across them when multiple inspectors were working on the same case. The regressions include fixed effects at the year X tax office level (unlike in the main results, where we controlled for inspector x year fixed effects). Standard errors are clustered at the tax office x year level. This table is discussed in Section 6.

Table 10: Heterogeneity by inspectors' characteristics

	P(start)		P(detection)		log(evasion)	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.517***	-0.149**	0.131	-0.00366	-0.429	-0.328
	(0.184)	(0.0637)	(0.107)	(0.0717)	(1.130)	(0.560)
Algorithm x Masters/PhD	0.179	0.0396	-0.0763	0.00727	-0.393	0.610
	(0.235)	(0.0550)	(0.0746)	(0.0728)	(0.984)	(0.466)
Masters/PhD		-0.0606	0.0740	0.0331	0.454	-0.164
		(0.0452)	(0.0568)	(0.0512)	(0.355)	(0.338)
Algorithm x Enthusiasm	0.138	0.0189	-0.0235	-0.0734	0.105	0.0881
	(0.228)	(0.0623)	(0.0861)	(0.0678)	(0.571)	(0.565)
Enthusiasm for Alg.		-0.0419	-0.0112	-0.00457	-0.974	0.143
		(0.0480)	(0.0509)	(0.0541)	(0.687)	(0.343)
Algorithm x Experience	0.129	0.134**	-0.103	0.129**	-0.192	0.303
	(0.149)	(0.0563)	(0.0765)	(0.0625)	(0.511)	(0.405)
Experience		-0.103**	0.0519**	-0.0705*	0.386	-0.451
		(0.0401)	(0.0240)	(0.0420)	(0.358)	(0.314)
N	893	1272	423	428	404	362
R2	0.245	0.153	0.0505	0.401	0.195	0.288
Mean outcome	0.54	0.33	0.95	0.84	18.97	17.67

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the OLS results of the probability of starting the audit on the selection method, controlling for three inspector characteristics. The assignment at the inspector level is only done for desk audits, and we use the characteristics of the originally assigned inspector in the regressions of columns 2, 4, and 6. For full audits, the assignment is done at the tax office level. In these cases, we take the average characteristics of inspectors working at the tax office in the year of the selection for column 1. Since these values are fixed within year x tax office, they are collinear with the year x tax office fixed effects and are excluded from the first specification. For columns 3 and 5, we use the characteristics of the inspectors who effectively worked on the full audit case. Information about inspectors comes from the inspector survey and covers most of the inspectors in the selection. The variable "Masters/PhD" indicates whether the inspector has a Masters or PhD degree. "Enthusiasm for Algorithm" indicates that the inspector answered "I agree" to the statement that "It would be useful to have automated selection for audit cases". "High experience" means that the inspector had more experience at the tax authority than the median (the median was 99 months). The regression has fixed effects at the year X tax office level (unlike in the main results, where we controlled for inspector x year fixed effects). Standard errors are still clustered at the list level (tax office x year level for full audits, inspector x year level for desk audits), like in the main tables. This table is discussed in Section 6.

Table 11: Risk score quality

	P(start)		P(detect)		log(evasion)	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.295***	-0.0735**	-0.00245	-0.0762*	-1.232***	-0.458*
	(0.0656)	(0.0305)	(0.0258)	(0.0399)	(0.386)	(0.236)
Inspectors x Overlap	0.0330	0.0395	0.0697	-0.0870*	-1.019**	-0.802*
	(0.0747)	(0.0458)	(0.0622)	(0.0521)	(0.443)	(0.424)
Algorithm x Random		0.0416		0.0445		-0.0385
		(0.0379)		(0.0369)		(0.236)
Risk score	0.0482*	-0.00132	0.00341	0.00375	0.448**	0.586***
	(0.0278)	(0.0191)	(0.0210)	(0.0219)	(0.167)	(0.170)
Alg. x Risk score	-0.0147	0.0215	0.0118	0.0145	-0.319*	-0.552***
	(0.0296)	(0.0200)	(0.0271)	(0.0206)	(0.185)	(0.168)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	Yes	No	Yes	No	Yes
N	944	2731	507	997	453	732
R2	0.269	0.318	0.153	0.431	0.232	0.420
Mean outcome	0.53	0.37	0.89	0.73	18.84	17.39

Obs: Only desk audit cases were used in the intervention. OLS results with fixed effects at the year X inspector level. Standard errors are clustered at the inspector level. This table is discussed in Section 6.

Table 12: Treatment effect of information intervention

	(1) P(start)	(2) P(detect)	(3) log(evasion)	(4) P(start)	(5) P(detect)	(6) log(evasion)
Algorithm	-0.0644**	-0.0695	-0.184	-0.0615**	-0.0695	-0.179
	(0.0273)	(0.0428)	(0.216)	(0.0274)	(0.0430)	(0.216)
Algorithm x Random	0.000589	0.00509	-0.0883	-0.00117	0.00510	-0.0885
	(0.0322)	(0.0329)	(0.172)	(0.0319)	(0.0329)	(0.173)
Information	0.0112	-0.0423	-0.0682			
	(0.0299)	(0.0375)	(0.213)			
Info. (indicators)				0.0257	-0.0431	0.0198
				(0.0395)	(0.0477)	(0.240)
Info. (indicators+data)				0.00596	-0.0413	-0.136
				(0.0370)	(0.0465)	(0.279)
Algorithm x Information	0.0198	0.0552	0.173			
	(0.0355)	(0.0445)	(0.293)			
Alg. x Info. (Indicators)				0.0273	0.0572	0.165
				(0.0489)	(0.0539)	(0.325)
Alg. x Info. (Indicators+data)				0.0135	0.0531	0.178
				(0.0450)	(0.0540)	(0.383)
N	2731	997	732	2731	997	732
R2	0.317	0.430	0.403	0.317	0.430	0.403
Mean outcome	0.37	0.73	17.39	0.37	0.73	17.39

Obs: This table shows the estimation of the information treatment intervention's effect on the probability audit execution, and the correlation of the information treatment with subsequent audit outcomes. Only desk audit cases were used in the intervention, so the sample does not include any full audit case. Columns 1, 2, and 3 show the results for a specification containing a dummy indicating whether the case was treated. Columns 4, 5, and 6 distinguish between two modalities of the treatment: providing only indicators of risk about the case to the inspectors and providing risk indicators plus a spreadsheet with data on the taxpayers' tax declarations and third-party data. OLS results with fixed effects at the year X inspector level. Standard errors are clustered at the inspector level. This table is discussed in Section 6.

Table 13: Algorithm effect on probability of starting case, controlling for observables

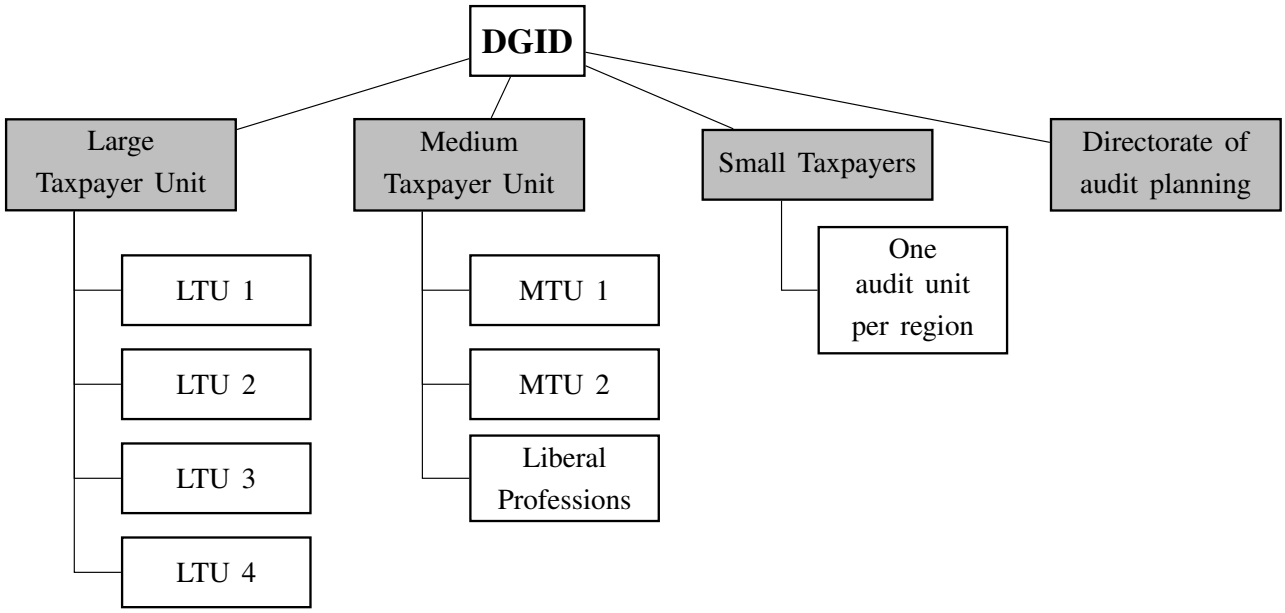
	(1) Full audits	(2) Desk audits	(3) Full audits	(4) Desk audits	(5) Full audits	(6) Desk audits	(7) Full audits	(8) Desk audits
Algorithm	-0.158***	-0.0911***	-0.155***	-0.0496**	-0.132***	-0.0378*	-0.133***	-0.0248
	(0.0358)	(0.0217)	(0.0354)	(0.0212)	(0.0406)	(0.0212)	(0.0344)	(0.0210)
Inspectors x Overlap	0.115*	0.0417	0.109*	0.0395	0.0278	0.0176	0.0324	0.0315
	(0.0564)	(0.0400)	(0.0550)	(0.0351)	(0.0237)	(0.0206)	(0.0260)	(0.0201)
Algorithm x Random		0.0143		0.0145		0.0596*		0.0493
		(0.0345)		(0.0310)		(0.0323)		(0.0332)
Prediction logit	0.484***	0.516***						
	(0.0571)	(0.0470)						
Prediction lasso			1.104***	1.017***				
			(0.127)	(0.112)				
Prediction RF					1.242***	1.025***		
					(0.0495)	(0.0356)		
Prediction RF cont.							1.170***	1.050***
							(0.0565)	(0.0375)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	925	2246	944	2731	944	2731	944	2731
R2	0.329	0.297	0.351	0.342	0.617	0.534	0.602	0.524
Mean outcome	0.54	0.39	0.53	0.37	0.53	0.37	0.53	0.37

Obs: This table shows the estimation of the information treatment intervention's effect on the probability audit execution, and the correlation of the information treatment with subsequent audit outcomes. Only desk audit cases were used in the intervention, so the sample does not include any full audit case. Columns 1, 2, and 3 show the results for a specification containing a dummy indicating whether the case was treated. Columns 4, 5, and 6 distinguish between two modalities of the treatment: providing only indicators of risk about the case to the inspectors and providing risk indicators plus a spreadsheet with data on the taxpayers' tax declarations and third-party data. OLS results with fixed effects at the year X inspector level. Standard errors are clustered at the inspector level. This table is discussed in Section 6.

Figures

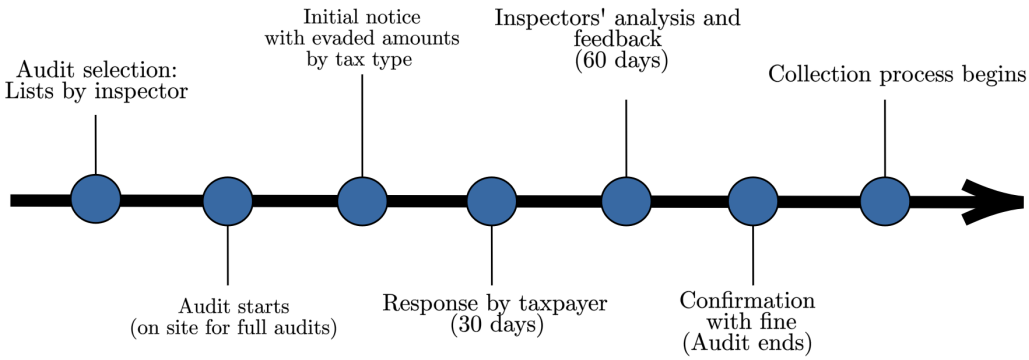
Audits in Senegal

Figure 1: Organizational chart of the Senegalese tax authority (DGID)



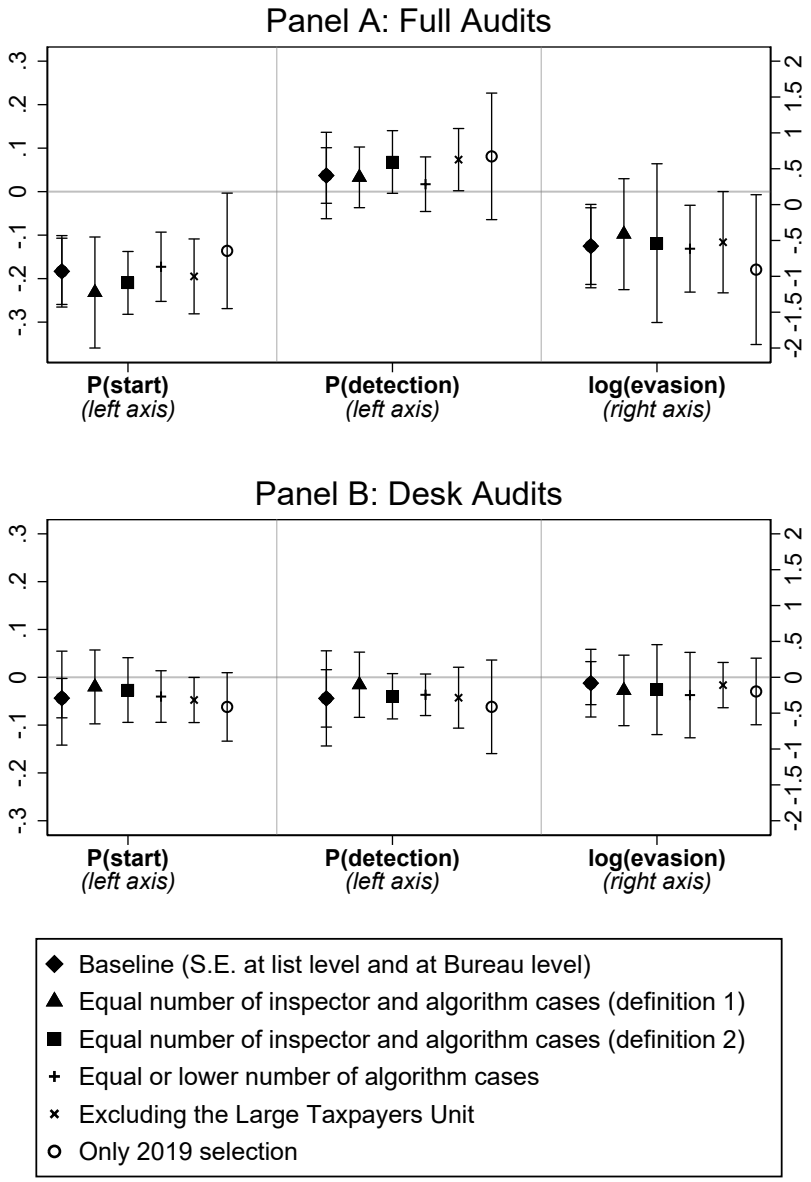
Notes: This figure shows the organizational structure of the tax authority in Senegal (DGID). Tax collection and enforcement is divided into three main branches: the Large Taxpayer Unit (divided into four units specialized by economic activity), the Medium Taxpayer Unit (divided into three units) and the multiple Small Taxpayers Offices, which are specialized by region and oversee small firms and individuals. DGID also has a unit to plan audit activities. This figure is discussed in Section 2.

Figure 2: Audit process at the Senegalese tax authority



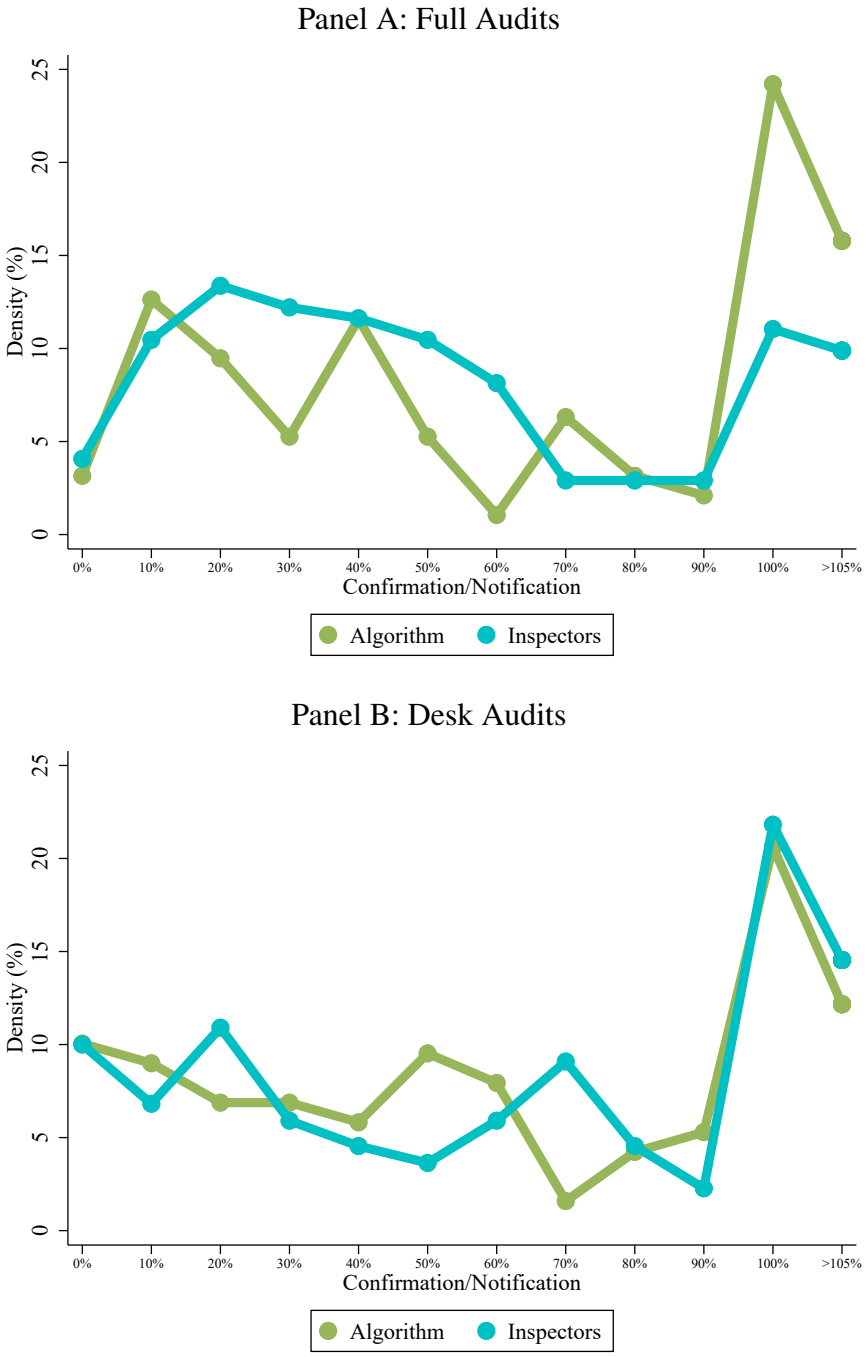
Notes: This figure depicts the steps taken to inspect a taxpayer. At the beginning of the year, inspectors are given a list of audits agreed with their superior in the hierarchy. Upon conducting the audit, they draft an initial notice containing apparent infractions. The initial notice contains the value of presumed evasion and the corresponding fine. The taxpayer can respond to the notice providing evidence that they have complied, which the inspector analyses before sending a final notice. Shortly after the final notice, the taxpayer receives a request to pay the evaded amount plus fines. This figure is discussed in Section 2.

Figure 3: Robustness Across Subsamples



Notes: This figure shows the robustness of our results in different subsamples, for full audits (panel A) and desk audits (panel B), and for the three main outcomes as indicated in the x-axis titles. In each sub-panel, the first coefficient reproduces our main result from Table 4, which we call the baseline. We show standard errors when clustering at the list level, as in our main analysis, and at the bureau level. The remaining coefficients using the same empirical specifications as Table 4, employing inspector fixed-effects for the desk audit estimations, and limit the sample in the following way: 2) lists that feature and equal number of inspector-selected and algorithm-selected cases (as observed by the researchers), 3) lists that feature and equal number of inspector-selected and algorithm-selected cases (as observed by the inspectors), 4) office-year observations where the number of algorithm-selected cases is equal to or lower than the number of inspector-selected cases, 5) small and medium taxpayer offices only (i.e. excluding the large taxpayer office), 6) audit lists for 2019. For the distinction between methods 1 and 2 for equalizing the numbers of algorithm and inspector-selected cases, see footnote 18. Tables **??** to A12 show the details of all these estimations with numbers of observations in table format. This figure is discussed in Section 5.2.

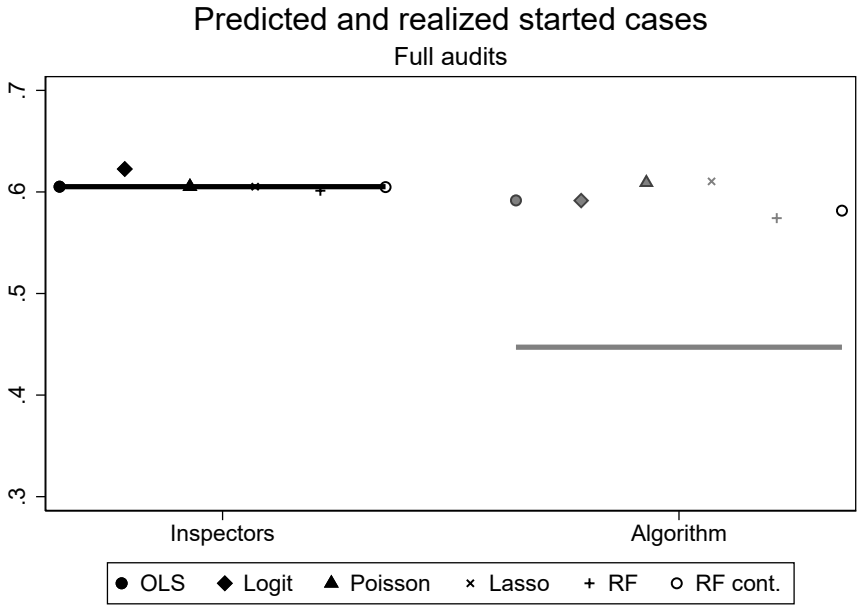
Figure 4: Dispute of Audit Results
Distribution of Confirmed Amount/Notified Amount of Evasion



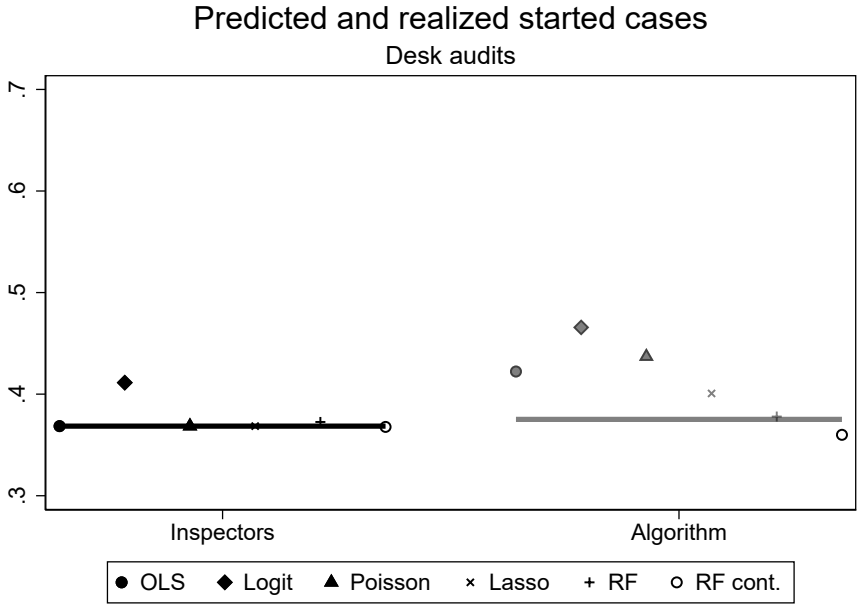
Notes: This figure shows the distribution of the share between confirmation (final notice) amounts and notification amounts for a firm, conditional on the selection method, and on the firm having positive values for confirmation and notification. The share of confirmation to notification measures the intensity with which taxpayers successfully dispute the initial notice. This figure is discussed in Section 5.

Figure 5: Predicted and realized execution of cases

Panel A: Full Audits



Panel B: Desk Audits



Notes: This figure shows the predicted mean execution of audits based on observable characteristics of the firms. Several models are estimated using only inspector-selected firms, and the model is used to predict the execution rate of algorithm cases based on the firms' characteristics. The difference between the realized execution rate and the predicted execution rate represents the part of the algorithm's effect that cannot be explained by firm characteristics. This figure is discussed in Section 6.

A APPENDIX

Audit procedure

Table A1: Steps of Risk-Score Calculation

Step	Description
(1) Prepare database	The tax declarations of each taxpayer are merged across taxes (VAT, CIT, Payroll) and across years. Data from third parties is then merged in (customs, procurement, transaction network).
(2) Calculate inconsistency ratios	Inconsistencies are situations in which a self-reported tax liability can be considered as misreported or incomplete, by comparing different data sources. An example of an inconsistency ratio is third-party reported sales over self-reported sales.
(3) Calculate anomaly ratios	Anomalies correspond to abnormal reporting behavior, compared to peers. Anomalies may be associated with tax evasion, but do not indicate tax evasion behavior with certainty. An example of an anomaly ratio is the inverse of the profit rate.
(4) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons of anomaly ratios are done within clusters.
(5) Transform ratios into risk indicators	For inconsistencies, the magnitude of the ratio is used to assign a value, ranging from one to ten (using deciles). For anomalies, firms within the top decile of a particular ratio within their cluster are assigned a value of one.
(6) Assign weights to indicators	Weights are assigned to each indicator reflecting our beliefs about their relative importance.
(7) Aggregate indicators and years	The weighted risk indicators are first aggregated for each year. Then the yearly scores are summed up to form a total risk score covering the past four years. More recent years are weighted higher than more distant years.
(8) Weigh risk score by declared turnover	The aggregated risk score is weighted by the log of turnover to give more importance to larger firms.

Notes: This table describes the steps taken in calculating the risk score based on which the algorithm selects firms for tax audits. This is discussed in Section 4.1.

Table A2: Tax audit selection methods in selected countries

Country	Discretionary selection	Risk analysis	Random selection
Kenya	Yes ; For all except large taxpayers	Yes ; Only for large taxpayers	No
Senegal	Yes	Yes, Introduced in FY 2018	Introduced in FY 2018
Zimbabwe	Yes; Inspectors rated on selection.	Yes; based on turnover variances	No
Lesotho	No	No	Yes ; Randomly by managers
Tanzania	Abandonned in 2007	Yes	
United Kingdom	Yes; For 55% of audit cases	Yes; Risk scoring	Yes ; Simple random sample
Switzerland	Yes for all cases	No	Yes, periodically for some taxes
United States	No	Yes	
France	Yes; For intelligence gathering	Yes; statistical techniques, data-mining	No
Bulgaria	Yes ; According to set criteria	Yes; Central risk analysis	No
Turkey	No	Yes; Analysis by tax type	Yes ; to collect unbiased data

Notes: This is based on Khwaja, Awasthi, and Loeprick (2011) and our survey of select country tax officials.

A.1 Risk Scoring of Tax Evasion

C.1.1 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 A1 discussed in Section 4.1 summarizes the steps we took in deriving the risk score.

C.1.2 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 **??** 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 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 [24]. 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

[24] 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.

compared to its peers, which might be associated with evasion. Discrepancies and anomalies are aggregated to produce a risk-score for each taxpayer.

C.1.3 Balancing tests

Table A3: Balancing table for randomization of ordering: probability of being on top of the list

	(1) P(top)	(2) P(middle)	(3) P(top)	(4) P(middle)	(5) P(top)	(6) P(middle)	(7) P(top)	(8) P(middle)
Algorithm case	0.00264	0.00390						
	(0.0204)	(0.0139)						
log(Mean Turnover)			0.000889	0.000527				
			(0.00150)	(0.00148)				
log(Mean Tax Liability)					0.00115	0.00100		
					(0.00159)	(0.00156)		
Profit rate							0.0490	-0.104*
							(0.0592)	(0.0594)
N	3675	3675	3675	3675	3675	3675	3675	3675
R2	0.000523	0.000424	0.000609	0.000440	0.000664	0.000521	0.000720	0.00133
Mean outcome	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the coefficients of a regression to predict the position of a case on the inspectors' list, conditional on characteristics. It predicts the probability that the case is located at the top third of the list, or in the middle third of the list. The table shows OLS results with fixed effects at the year X inspector level (tax office level for full audits). Standard errors are clustered at the inspector level (tax office level for full audits). This table is discussed in Section 5.

Table A4: Balancing tests of information intervention

	(1) log(turnover 2017)	(2) profit rate 2017	(3) log(payroll 2017)
Information	0.780	-0.00546	1.408**
	(0.526)	(0.0188)	(0.668)
Overlap	-0.125	-0.0784**	-0.141
	(0.797)	(0.0364)	(0.968)
Overlap x Information	0.271	0.0656	0.825
	(1.444)	(0.0521)	(1.799)
Algorithm	0.0827	-0.0281	-0.991
	(0.578)	(0.0211)	(0.697)
Algorithm x Information	-0.639	-0.00814	-1.870**
	(0.683)	(0.0279)	(0.857)
Random	-0.624	-0.00513	-0.344
	(0.822)	(0.0283)	(1.085)
Random x Information	-0.381	0.0307	1.185
	(0.964)	(0.0333)	(1.180)
N	2494	2275	2494
R2	0.191	0.112	0.219
Mean outcome	16.37	-0.05	9.46

Obs: Only desk audit cases were used in the intervention. OLS results with fixed effects at the year X inspector level. Standard errors are clustered at the inspector level. This table is discussed in Section 6.

C.1.4 Additional results on selection method

Table A5: Evasion Rate as % of Liability

	(1) Full audits	(2) Desk audits	(3) Desk audits	(4) All
Algorithm	0.0446	0.00993	0.00620	0.0220
	(0.0408)	(0.0258)	(0.0252)	(0.0215)
Inspectors x Overlap	0.0463	-0.0257	-0.0771*	-0.0257
	(0.0504)	(0.0461)	(0.0461)	(0.0338)
Algorithm x Random		0.00380	0.00322	-0.00189
		(0.0383)	(0.0305)	(0.0293)
Tax center X Year	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	Yes
N	502	999	979	1481
R2	0.105	0.146	0.287	0.230
Mean outcome	0.36	0.30	0.30	0.32

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A6: Evasion Rate as % of Pre-Audit Mean Turnover

	(1) Full audits	(2) Desk audits	(3) Desk audits	(4) All
Algorithm	0.0211	0.0290	0.0294*	0.0261
	(0.0386)	(0.0212)	(0.0176)	(0.0182)
Inspectors x Overlap	-0.00983	0.0198	0.00640	-0.000482
	(0.0400)	(0.0345)	(0.0308)	(0.0245)
Algorithm x Random		0.0235	0.0196	0.0209
		(0.0349)	(0.0266)	(0.0264)
Tax center X Year	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	Yes
N	501	998	977	1478
R2	0.161	0.153	0.282	0.245
Mean outcome	0.19	0.17	0.16	0.17

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

C.1.5 Additional robustness on main results

Table A7: Results of main outcomes - Cluster S.E. at Tax Office level

	P(start)			P(detect)			log(evasion)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.183***	-0.0456	-0.0436	0.0371	-0.0465	-0.0441	-0.579**	-0.178	-0.0818
	(0.0347)	(0.0392)	(0.0415)	(0.0420)	(0.0435)	(0.0421)	(0.226)	(0.195)	(0.200)
Inspectors x Overlap	0.159***	0.0465	0.0423	0.0809**	-0.0744*	-0.0756**	0.0689	-0.213	-0.0264
	(0.0330)	(0.0471)	(0.0475)	(0.0311)	(0.0366)	(0.0237)	(0.354)	(0.363)	(0.391)
Algorithm x Random		-0.00104	-0.00139		0.00284	0.00534		-0.0921	-0.0836
		(0.0254)	(0.0281)		(0.0472)	(0.0388)		(0.0816)	(0.108)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	944	2731	2731	507	1016	997	453	751	732
R2	0.263	0.227	0.317	0.151	0.324	0.430	0.214	0.271	0.402
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	18.84	17.42	17.39

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes restricting to the lists in which the number of algorithm and discretionary cases were exactly the same. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A8: Results of main outcomes - Equal number of cases (definition 1)

	P(start)			P(detect)			log(evasion)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.232***	-0.0128	-0.0128	0.0329	-0.0380	-0.0206	-0.412	-0.279	-0.290
	(0.0573)	(0.0380)	(0.0377)	(0.0313)	(0.0330)	(0.0334)	(0.347)	(0.233)	(0.263)
Inspectors x Overlap		0.0138	0.000219		0.0329	-0.00949		-1.245***	-0.126
		(0.194)	(0.223)		(0.0383)	(0.0154)		(0.236)	(0.247)
Algorithm x Random		0.471***	0.448***		0.0709	0.0119		-1.447**	-0.662
		(0.0578)	(0.0987)		(0.0494)	(0.0193)		(0.640)	(0.425)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	414	834	834	264	214	204	243	205	195
R2	0.194	0.174	0.235	0.0925	0.0593	0.173	0.176	0.367	0.451
Mean outcome	0.63	0.25	0.25	0.92	0.95	0.95	18.52	17.57	17.56

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes excluding the Large Taxpayer Unit. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A9: Results of main outcomes - Equal number of cases (definition 2)

	P(start)			P(detect)			log(evasion)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.210***	-0.0272	-0.0267	0.0682*	-0.0647**	-0.0397*	-0.537	-0.300	-0.172
	(0.0319)	(0.0339)	(0.0338)	(0.0318)	(0.0280)	(0.0235)	(0.489)	(0.278)	(0.311)
Inspectors x Overlap	-0.120	0.103	0.0896	0.0764***	-0.0815	-0.116	0.709**	-1.811***	-1.410***
	(0.137)	(0.0957)	(0.110)	(0.0145)	(0.0828)	(0.1000)	(0.233)	(0.462)	(0.505)
Algorithm x Random		0.0266	-0.0150		0.0852*	0.0410		-0.122	-0.222
		(0.107)	(0.125)		(0.0429)	(0.0310)		(0.865)	(0.710)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	336	838	838	205	180	173	187	175	168
R2	0.219	0.131	0.192	0.100	0.0659	0.252	0.115	0.261	0.380
Mean outcome	0.61	0.21	0.21	0.91	0.97	0.97	18.22	17.58	17.60

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes only for the year 2019, which we consider the implementation of the experiment to be best. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A10: Results of main outcomes - More algorithm cases

	P(start)			P(detect)			log(evasion)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.173***	-0.0448	-0.0403	0.0171	-0.0537**	-0.0366*	-0.615**	-0.309	-0.248
	(0.0385)	(0.0276)	(0.0271)	(0.0303)	(0.0241)	(0.0217)	(0.292)	(0.266)	(0.297)
Inspectors x Overlap	0.131*	0.0752	0.0716	0.0798*	-0.0790	-0.0904	0.0694	-1.183***	-1.098*
	(0.0696)	(0.0454)	(0.0465)	(0.0465)	(0.0565)	(0.0768)	(0.296)	(0.416)	(0.553)
Algorithm x Random		0.0333	-0.00994		0.0791*	0.0471		-0.00474	-0.0769
		(0.106)	(0.119)		(0.0445)	(0.0360)		(0.877)	(0.706)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	855	1180	1180	474	226	208	428	220	203
R2	0.261	0.123	0.202	0.116	0.0674	0.252	0.208	0.299	0.440
Mean outcome	0.55	0.19	0.19	0.90	0.97	0.97	18.86	17.87	17.81

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes computing the standard errors at the bureau level, as opposed to bureau-year level. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies.

Table A11: Results of main outcomes - Excluding Large Taxpayer Unit

	P(start)			P(detect)			log(evasion)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.195***	-0.0488**	-0.0475**	0.0736**	-0.0486	-0.0427	-0.524	-0.224	-0.109
	(0.0406)	(0.0238)	(0.0238)	(0.0337)	(0.0303)	(0.0321)	(0.333)	(0.153)	(0.159)
Inspectors x Overlap	0.220*	0.00784	-0.00135	0.147*	-0.0803	-0.0817	1.117**	0.182	0.309
	(0.123)	(0.0546)	(0.0554)	(0.0823)	(0.0589)	(0.0536)	(0.470)	(0.339)	(0.352)
Algorithm x Random		-0.00860	-0.00866		0.0131	0.00958		-0.0476	-0.00567
		(0.0332)	(0.0335)		(0.0369)	(0.0352)		(0.185)	(0.177)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	548	2194	2194	319	906	900	285	645	639
R2	0.277	0.222	0.304	0.205	0.307	0.414	0.135	0.152	0.289
Mean outcome	0.58	0.41	0.41	0.89	0.71	0.71	18.32	17.13	17.13

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes computing the standard errors at the bureau level, as opposed to bureau-year level. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies.

Table A12: Results of main outcomes - Only 2019 program

	P(start)			P(detect)			log(evasion)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	-0.136**	-0.0620*	-0.0620*	0.0811	-0.0693	-0.0618	-0.907*	-0.351	-0.198
	(0.0587)	(0.0356)	(0.0355)	(0.0643)	(0.0481)	(0.0485)	(0.462)	(0.239)	(0.230)
Inspectors x Overlap	0.234*	-0.0215	-0.00910	0.143**	-0.146	-0.154*	0.198	-0.266	0.133
	(0.112)	(0.0872)	(0.0885)	(0.0447)	(0.0913)	(0.0832)	(0.314)	(0.401)	(0.440)
Algorithm x Random		0.0554	0.0623*		0.00187	0.00197		-0.0533	-0.00740
		(0.0378)	(0.0364)		(0.0544)	(0.0514)		(0.275)	(0.268)
Tax center X Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector X Year	No	No	Yes	No	No	Yes	No	No	Yes
N	327	909	909	195	547	547	163	299	299
R2	0.165	0.0978	0.242	0.183	0.142	0.281	0.221	0.153	0.317
Mean outcome	0.59	0.60	0.60	0.83	0.54	0.54	18.80	17.14	17.14

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. This table show the results for the main outcomes computing the standard errors at the bureau level, as opposed to bureau-year level. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies.

Main Results

Table A13: Probably of Accepting the Audit Result Based on Notification

	(1) Full audits	(2) Desk audits	(3) All
Algorithm	0.121**	0.0101	0.0641*
	(0.0433)	(0.0519)	(0.0342)
Inspectors x Overlap	-0.0310	0.0291	0.00214
	(0.0651)	(0.1000)	(0.0601)
Algorithm x Random		-0.00779	-0.0434
		(0.0954)	(0.0901)
N	264	372	636
R2	0.171	0.235	0.211
Mean outcome	0.15	0.21	0.19

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A14: log(confirmation/notification)|conf.,notification>0

	(1) Full audits	(2) Desk audits	(3) All
Algorithm	0.157	-0.0321	0.0582
	(0.179)	(0.179)	(0.125)
Inspectors x Overlap	-0.123	0.509	0.183
	(0.414)	(0.415)	(0.299)
Algorithm x Random		0.167	0.0898
		(0.207)	(0.198)
N	260	346	606
R2	0.118	0.275	0.213
Mean outcome	-0.77	-0.62	-0.69

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A15: log(notification)|conf.,notification>0

	(1) Full audits	(2) Desk audits	(3) All
Algorithm	-0.344	-0.0471	-0.181
	(0.202)	(0.175)	(0.133)
Inspectors x Overlap	0.682*	-0.582	0.0489
	(0.388)	(0.420)	(0.289)
Algorithm x Random		-0.133	-0.00408
		(0.270)	(0.263)
N	264	372	636
R2	0.329	0.537	0.507
Mean outcome	19.47	17.98	18.60

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A16: log(confirmation)|conf.,notification>0

	(1) Full audits	(2) Desk audits	(3) All
Algorithm	-0.191	-0.193	-0.186
	(0.258)	(0.236)	(0.173)
Inspectors x Overlap	0.543	-0.0598	0.257
	(0.385)	(0.479)	(0.296)
Algorithm x Random		0.0537	0.0704
		(0.249)	(0.242)
N	260	346	606
R2	0.189	0.509	0.432
Mean outcome	18.68	17.33	17.91

Obs: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of audit outcome on the selection method of the case. The sample includes all cases selected in the audit programs of 2018, 2019, and 2020. The data includes the selection and audits of the Large Taxpayer Unit, Medium Taxpayer Units 1 and 2, Liberal Professionals, and the regional SME units of Dakar Plateau, Grand Dakar, Pikine Guediawaye, and Ngor Almadies. Standard errors are clustered at the tax office x year levels, and inspector x year levels when inspector fixed effects are included.

Table A17: Treatment effect of information intervention

	(1) P(start)	(2) P(detect)	(3) log(evasion)	(4) P(start)	(5) P(detect)	(6) log(evasion)
Algorithm	-0.0548***	-0.0354	-0.0827	-0.0521**	-0.0353	-0.0786
	(0.0204)	(0.0291)	(0.151)	(0.0208)	(0.0292)	(0.151)
Algorithm x Random	0.00192	0.00537	-0.0819	0.000437	0.00537	-0.0822
	(0.0319)	(0.0330)	(0.172)	(0.0317)	(0.0330)	(0.172)
Information	0.0230	-0.0119	0.0255			
	(0.0186)	(0.0259)	(0.118)			
Info. (indicators)				0.0414	-0.0115	0.110
				(0.0266)	(0.0344)	(0.155)
Info. (indicators and data)				0.0131	-0.0122	-0.0400
				(0.0194)	(0.0308)	(0.136)
N	2731	997	732	2731	997	732
R2	0.317	0.429	0.402	0.317	0.429	0.403
Mean outcome	0.37	0.73	17.39	0.37	0.73	17.39

Obs: Only desk audit cases were used in the intervention. OLS results with fixed effects at the year X inspector level. Standard errors are clustered at the inspector level. This table is discussed in Section 6.

C.1.6 Descriptive

Table A18: Count of selected desk audits by year, tax office, and selection method

		Random	Algorithm	Discretion	Overlap	Replacement	Total
DGE	2018	60	72	81	12	0	213
	2019	0	0	0	0	0	0
	2020	0	85	239	76	85	409
CME 1	2018	34	49	52	5	0	135
	2019	14	42	52	10	7	115
	2020	0	0	0	0	0	0
CME 2	2018	49	78	83	6	0	210
	2019	38	83	95	12	16	232
	2020	0	38	38	1	38	114
CPR	2018	59	86	95	10	0	240
	2019	26	66	70	4	12	174
	2020	0	70	70	7	70	210
Dakar P.	2018	0	0	0	0	0	0
	2019	37	71	78	7	15	201
	2020	0	72	72	1	72	216
Ngor A.	2018	0	0	0	0	0	0
	2019	19	57	59	2	10	145
	2020	0	49	59	2	49	157
Pikine G.	2018	0	0	0	0	0	0
	2019	14	26	26	0	8	74
	2020	0	63	63	2	63	189
G. Dakar	2018	0	0	0	0	0	0
	2019	14	10	12	2	8	44
	2020	0	53	53	2	53	159
All	2018	202	285	311	33	0	798
	2019	162	355	392	37	76	985
	2020	0	430	594	91	430	1454
Total		364	1070	1297	161	506	3237

Note: Number of selected desk audits by year and tax office. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion.

Table A19: Count of selected full audits by year, tax office, and selection method

		Algorithm	Discretion	Overlap	Total
DGE	2018	81	94	13	182
	2019	25	96	11	121
	2020	25	75	5	100
CME 1	2018	31	33	2	67
	2019	27	27	1	54
	2020	25	25	0	50
CME 2	2018	25	25	0	53
	2019	20	20	0	40
	2020	25	25	0	50
CPR	2018	15	15	0	32
	2019	15	15	1	30
	2020	20	16	2	36
Dakar P.	2018	0	0	0	0
	2019	14	15	2	29
	2020	7	7	0	14
Ngor A.	2018	0	0	0	0
	2019	11	10	0	21
	2020	8	8	0	16
Pikine G.	2018	0	0	0	0
	2019	8	7	0	15
	2020	8	8	0	16
G. Dakar	2018	0	0	0	0
	2019	9	8	0	17
	2020	8	8	1	16
All	2018	152	167	15	334
	2019	129	198	15	327
	2020	126	172	8	298
Total		407	537	38	959

Note: Number of selected full audits by year and tax office. The sum of the rows is larger than the total because there are overlapping cases between algorithm and discretion.