Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal

November 26, 2026

Pierre Bachas*
World Bank Research
Alipio Ferreira‡
Southern Methodist University

Anne Brockmeyer†
World Bank, IFS, UCL, and CEPR
Bassirou Sarr
Senegal Ministry of Finance

Abstract

Governments are increasingly using algorithms to guide bureaucratic decisions, yet there is little evidence on whether algorithms outperform bureaucrats on complex tasks. Leveraging recently digitized tax data, we conducted a three-year, at-scale field experiment in which half of Senegal's tax audit program was selected by a transparent risk-score algorithm. Relative to inspector-selected audits, algorithm-selected audits are 17 percentage points less likely to be completed, uncover 47% less tax evasion, are less cost-effective, and do not reduce corruption. We then benchmark the original risk score, constructed using economic logic as no training data were yet available, against an ex post optimized machine-learning algorithm trained on newly digitized audit outcomes. Even this optimized algorithm only marginally outperforms bureaucrats at selecting high-yield audits. The findings suggest that automating complex decisions can reduce productivity by crowding out bureaucrats' tacit expertise.

JEL codes: O17, H83, H26, D73
Keywords: Audits, Algorithms, Discretion, Risk-score, Tax Administration, Tax Evasion

*Pierre Bachas: pbachas@worldbank.org; †Anne Brockmeyer: abrockmeyer@worldbank.org; ‡Alipio Ferreira: alipioferreira@smu.edu.

We thank Denis Cogneau, Laurent Corthay, Léo Czajka, Lucie Gadenne, Stephen Hansen, Janet Jiang, Nicola Limodio, Jan Loeprick, Markus Kitzmuller, Justine Knebelmann, Imran Rasul, Dan Rogger, Eduardo Souza-Rodrigues, and Gabriel Zucman for helpful comments and discussions, and seminar/conference audiences at Berkeley Haas, IFC, IFS, IIPF, CESifo Public Economics Week, the CEPR Paris Forum, INSPER, TARC Exeter, WB Tax Conference, CMI TaxCapDev Conference, Paris School of Economics, Oxford Centre for Business Taxation, Oxford Economics Department, PUC Rio, University of Muenster, University of New Mexico, Norwegian School of Economics, NTA Conference, North Texas Economics Conference, RIDGE Public Economics, VATT and Universidad del Pacífico. We thank Senegal's Tax Administration (DGID), in particular, Bassirou S. Niasse, Amadou A. Badiane, Oumar D. Diagne, Hady Dieye, Mor Fall, Serigne M. Fall, and Mathiam Thioub. We thank Samba Mbaye, Assane Sylla, and Medoune Sall from the CRDES for their collaboration; Roldan Enamorado, Oumy Thiandoum, and Rafael Vilarouca for excellent research assistance; and the Paris School of Economics and CEPREMAP for administrative support. The authors are grateful for financial support from the World Bank via the Knowledge for Change (KCP) Program, the Global Tax Program (GTP) and the research groups' research support budget (RSB); UKAID via its Economic and Development Institutions program and the Centre for Tax Analysis in Developing Countries (TaxDev) at the Institute for Fiscal Studies; and the UKRI through Brockmeyer's Future Leaders Fellowship (grant reference MR/V025058/1). The findings, interpretations, and conclusions do not represent the views of the World Bank, its affiliated organizations, its Executive Directors, or the governments they represent, nor of the Government of Senegal. All errors are our own.

1 Introduction

In contexts where administrative data is incomplete or fragmented, bureaucrats often exercise important discretion over policy decisions. For instance, individual bureaucrats may decide which taxpayers to audit, which water treatment facilities to inspect, and which manufacturing plants to monitor for pollution (Khwaja et al., 2011; OECD, 2023). By contrast, in data-rich environments, such decisions are increasingly guided by rules and algorithms. Data-driven decision models leverage available information in a systematic manner but depend on high-quality data. Discretionary decisions rely on bureaucrats' experience and private information but can be vulnerable to bias. There is limited evidence on when replacing bureaucratic discretion with data-driven decision-making improves government performance, especially on complex tasks.

We examine whether algorithms can improve tax audit selection relative to the discretionary choices of tax inspectors, when deployed at scale. We study Senegal's corporate tax audit program, which targets firms with high levels of tax evasion. Senegal is particularly relevant because tax inspectors have traditionally exercised broad discretion over audit selection, tax evasion is estimated to exceed 30% of true liabilities, and recent investments in digital tax infrastructure—including mandatory electronic tax returns and the digitization of third-party information—have substantially expanded the administrative data available for risk-based selection. As many countries invest in tax digitization, Senegal offers an opportunity to test the potential of newly available administrative data and prediction algorithms to improve government decision-making.

In partnership with the Senegalese tax authority, we developed a *risk-score* algorithm to identify firms with high evasion amounts. This initiative marks the first systematic effort to leverage all available data for tax enforcement at scale in Senegal. The algorithm draws on taxpayers' self-assessment declarations across various taxes and on third-party records from customs, public procurement, and firm-to-firm transaction reports. Its design was subject to two key constraints: the selection criteria had to be simple and transparent to communicate, and the algorithm could not rely on historical audit outcome data, which were unavailable in digital format at the onset of the project. As a result, the risk-score algorithm is not a trained machine-learning algorithm, but rather built on intuitive compliance risk concepts and best practices, exploiting the data as thoroughly as possible.

For each firm, the algorithm calculates a risk score composed of multiple indicators designed to proxy for tax evasion. One class of indicators flags inconsistencies, for instance, when a firm's self-reported sales are lower than its third-party reported sales. Another class of indicators flags anomalies, such as abnormally low profit rates compared to other firms of the same size group and sector. The choice of risk indicators is based on consultations with experienced Senegalese tax inspectors, international experts, and other tax administrations.

The risk-score algorithm for audit selection was implemented at scale across tax audit offices in Senegal from 2018 to 2020. Each January, the tax administration approves a list of firms for each office that should be audited during the year. To experimentally assess the impact of data-driven selection on the administration's productivity, each list was split in two: half of the cases were selected at the discretion of tax inspectors, and half were selected by the risk-score algorithm. Within each list, the ordering of cases was randomized to minimize the possibility that cases were treated differently based of how they were selected. By design, inspectors knew how a case was selected. While overlaps between inspector- and algorithm-selected cases were possible, they remained rare in practice.

To quantify the relationship between the selection method and audit outcomes, we compare the mean outcomes of algorithm-selected and inspector-selected audits within each audit list (i.e. tax office by year). The outcome measures stem from three new data sources. First, we digitized the universe of audit results records, which report the audit start and end date, inspectors working on the case, uncovered infractions, and fines to pay. Second, we surveyed a subset of audited firms to capture their perceptions of the audit process and corruption experiences. Finally, we surveyed inspectors to record their opinions and characteristics, such as skills and experience. These unique survey data allow us to examine how tax office managers deploy human resources based on the cases' selection methods.

Our analysis has two parts. We first evaluate the risk-score selection relative to inspector discretion. We then use a machine-learning model trained on newly digitized audit outcomes to explain our findings and benchmark the potential of more advanced prediction algorithms.

In the first part, we evaluate risk-score selection along two dimensions. First, we examine audit outcomes. Despite the protocol requiring full program implementation, the execution rate of the audit program is only 53%. This adds an additional margin of selection: from the list of planned audit cases, inspectors in practice implement a subset. Inspectors are more likely to audit cases which they selected: the execution rate is 17 percentage points lower for algorithm-selected audits. This is partly because algorithm-selected and inspector-selected firms differ on several dimensions: algorithm-selected firms are smaller, older, and less profitable. Yet, even after controlling for firm characteristics, an audit execution gap of 14 percentage points between the two audit case types remains. Conditional on implementation, 90% of audits uncover tax evasion, and the detection rate does not vary between inspector and algorithm-selected cases. However, conditional on detection, inspector-selected cases yield 47% higher assessments of taxes evaded on average. The results are robust across subsamples and time periods.[1]

Second, we put the audit process under the microscope. Using detailed administrative data, we find that algorithm-selected audits use less manpower, are implemented faster by various

[1] A silver lining is that algorithm-selected firms are 13 percentage points less likely to dispute audit results, from a baseline dispute rate of 84%. This narrows but does not eliminate the gap in confirmed evasion.

measures, and are conducted by less experienced teams. The implementation of algorithm-selected audits is evaluated less positively by firms, as captured by their perceptions of auditors' professionalism, competency and efficiency. We do not detect any significant difference in the incidence of corruption, as reported by taxpayers. We can reject differences in corruption of over 0.2 standard deviation. Combining the amount of evasion uncovered and measures of audit costs into a measure of productivity, we find that algorithm-selected audits are significantly less productive than inspector-selected audits on average.

In the second part, we ask whether these shortcomings reflect limitations of the original risk-score algorithm or more fundamental limits to algorithmic selection in this context. Having digitized all audit results, we train a machine-learning model to predict detectable evasion. We use these predictions to explain the experimental findings and to benchmark the potential gains from an ex post optimized selection rule.

We first use these predictions to examine the execution and yield gaps between inspector- and algorithm-selected audits. Inspectors' execution decisions seem strategic: they prioritize cases with higher predicted yields and treat inspector- and algorithm-selected cases similarly conditional on predicted evasion. They therefore do not penalize algorithmic selection per se, but instead appear to equalize expected marginal returns across selection methods. The execution and yield gaps instead reflect the algorithm's inability to consistently identify the highest-yield cases. We also provide evidence that inspectors learn about case quality and adjust their execution decisions accordingly.

Second, we rule out two pre-registered alternative explanations. Inspectors might favor their own selections because the risk score is uninformative about evasion or because they lack relevant information about algorithm-selected cases. Neither explanation is supported by the data. The risk score strongly predicts both audit implementation and detected evasion among inspector-selected cases, confirming that it contains useful information. Moreover, a cross-randomized information treatment in the desk-audit program shows that providing inspectors with descriptions of risk flags and firm-level spreadsheets does not increase audit execution.

Finally, we use our machine-learning predictions to benchmark the gains from optimized audit execution among cases selected for the annual program. Within each tax office, we rank cases by predicted evasion and retain the top cases, holding fixed the number of audits. This suggests aggregate detected evasion could have increased by 18%. While meaningful, these gains are smaller than those reported for machine-learning applications in other settings.

Overall, the results highlight the difficulty of improving bureaucratic decision-making via algorithms in lower-income countries when bureaucrats are highly skilled and engage in a complex task. The simple risk-score algorithm, designed without recourse to comprehensive audit outcome data, lags behind the performance of inspector-selected tax audits, but even the machine-learning algorithm might only improve audit performance moderately.

The risk-score selection of firms performed poorly despite our best efforts to draw on international best-practice, the use of all available digital data, and the fine-tuning of the risk-score in three subsequent years based on lessons learned; and despite the fact that inspectors ultimately chose which program cases to execute, de-facto adding their skill to the risk-score selection. Given the close collaboration with and endorsement from senior management at the tax administration, we consider it unlikely that the results are driven by inspectors' desire to derail the algorithm. In fact, inspectors were highly motivated to use an algorithm for audit selection at the start of the program: in our baseline survey, 80% of them agreed with the statement that audit case selection should be automated and data-driven. Inspectors also have strong financial incentives to uncover evasion, as their bonuses are proportional to the uncovered evasion.

We interpret our results as showing that inspectors' performance in identifying high-potential cases is better than the risk-score and close to the ex-post optimized machine-learning algorithm. Tax inspectors are among the top civil servants, highly trained, experienced, and incentivized.[2] Meanwhile, the risk-score algorithm relies on imperfect data: match rates between datasets are incomplete, key variables are measured with error, and audit outcome data was not digitized before the experiment. Even the machine-learning algorithm is constrained by the limited availability of training data: despite implementation at scale in a medium-sized country, only 500 full audits were completed between 2018 and 2020. The resulting training sample is small relative to the large set of predictors available in administrative tax data. Moreover, given the limited number and high stakes of audits, randomizing audit selection to generate unbiased training data is impractical. As a result, the selected nature of the training sample may limit the algorithm's ability to generalize out of sample.

As with any experiment, our design has limitations. Our empirical approach isolates differences in audit outcomes between algorithm- and inspector-selected cases, conditional on factors affecting implementation. It does not, however, identify the optimal audit strategy, which depends not only on detected evasion but also on medium-term effects on audited firms and deterrence effects on non-audited firms. To shed light on the latter, we provide a back-of-the-envelope calculation of the size-dependent deterrence effects required for the risk-score algorithm to generate higher ex-ante compliance than inspector selection. Assuming firms know their size-specific audit probabilities and that the difference between selection methods is a reallocation of audits toward smaller firms under the algorithm, we find that the inspector selection generates higher aggregate declarations given the concentration of revenue among large firms. The risk-score algorithm dominates only if deterrence effects are negligible for larger firms.

This paper is organized as follows. Section 1.1 reviews the literature. Section 2 describes the context and data, Section 3 our conceptual framework, and Section 4 the intervention. Sections 5 and 6 present results on audit outcomes and the audit process, respectively. Section

[2] Anecdotally, as of 2025, both the current president and prime minister of Senegal are former tax inspectors.

7 examines potential mechanisms and Section 8 benchmarks an ex post optimized algorithm. Section 9 discusses policy implications and concludes.

1.1 Related Literature

Our work contributes to the rapidly growing literature on how algorithms can complement or substitute for human decision-making (see Table F.1 for an overview). Algorithms have been shown to improve decision-making and welfare in a variety of settings, including judicial bail decisions (Kleinberg et al., 2018), the targeting of health, safety and water quality inspections (Johnson et al., 2023; Glaeser et al., 2016; Hino et al., 2018), hiring decisions for teachers and police officers (Chalfin et al., 2016), and the targeting of tax rebates or public credit guarantees to firms (Andini et al., 2018, 2022). However, the evidence is not unequivocally positive. For instance, algorithms can exhibit biases (Obermeyer et al., 2019), and humans perform better when human supervisors rather than algorithms are in charge of task assignment and feedback provision (Margalit and Raviv, 2024; Ferreira and Cermeno-Leon, 2026). These studies focus almost exclusively on high-income countries. Our contribution is to test the performance of an algorithm in a low-income country context, at scale, and on a complex task. As in previous studies, the agents in our context (tax inspectors) are white-collar workers, but the skill difference between them and the average worker in the economy is larger than in previous studies. This might explain why the agents in our setting perform so well compared to the algorithm.

In the context of tax audit selection, Battaglini et al. (2025) and Black et al. (2022) argue that machine-learning algorithms can improve enforcement returns in Italy and the United States. In these high-income settings, audit outcomes and predictors are digitally available, and a share of audits can be selected randomly. In lower-income countries, by contrast, audit outcomes are often not digitized, audit programs are smaller, and audits are rarely randomized.[3] We instead evaluate a risk-score algorithm that may be better suited to lower-income settings than machine-learning approaches. A related study by Knebelmann et al. (2026) shows that an automated rule for assessing property tax liabilities yields more accurate assessments than bureaucratic discretion also in Senegal. Similarly, Haseeb and Vyborny (2022) show that replacing bureaucratic discretion with a proxy means test improved targeting and welfare in the allocation of cash transfers to the poor in Pakistan. Studying the impact of an AI roll-out in customer-support centers, Brynjolfsson et al. (2025) argue that the positive effects were mainly concentrated among low-skill agents. In all these settings, the interventions focused on relatively low-skill workers, often temporarily hired, with limited performance incentives, and with tasks that can be more easily automated. In contrast, in our setting, tax inspectors are elite government agents engaging in complex, multi-dimensional tasks.

[3] An exception is Pakistan, where a legal vacuum led to the randomization of audits, which Best et al. (2021) exploit to show that audits had no causal effect on medium-term compliance. This finding contrasts with Advani et al. (2023) who document a positive compliance impact of audits in the UK, and Kotsogiannis et al. (2024) who find positive medium-term effects for audits in Rwanda.

In addition, our analysis is relevant for the literature that studies how the targeting of inspections and the allocation of inspectors to cases affect compliance. Duflo et al. (2018) show that the discretionary selection of audits lowers pollution offenses by more than random targeting, as random inspections found fewer extreme polluters than discretionary inspections. Building on this finding, we examine whether data-driven selection can further improve upon discretionary targeting. Furthermore, Duflo et al. (2013) show that eliminating discretion may reduce collusion between firms and auditors, who may hide detectable offenses in their reports. Reliance on an algorithm may hence influence audit outcomes by removing this potential conflict of interest.Our study is the first experiment to introduce a new selection method at scale in an enforcement context, allowing us to isolate the effect of case selection on a bureaucracy's performance from potential confounding factors.

Finally, our work contributes to the literature on state capacity and development. One strand has documented the value of information for improving tax compliance (Kleven et al., 2011; Naritomi, 2019; Pomeranz, 2015) and the potential for new technologies to generate and process information for tax enforcement, with mixed results (Okunogbe and Santoro, 2023; Okunogbe and Tourek, 2024; Brockmeyer and Somarriba, 2025). Another strand has focused on the power of tax administration (Basri et al., 2021) and the role of bureaucrats in building state capacity (Besley et al., 2022; Finan et al., 2017). This strand has documented the value of discretion and autonomy for bureaucrats' performance (Bandiera et al., 2021; Rasul and Rogger, 2018), but has also highlighted the associated risks, e.g. favoritism (Szucs, 2024). For tax collectors specifically, the literature has quantified the impact of performance incentives (Khan et al., 2015, 2019) and the value of local information on taxpayers (Balán et al., 2022; Dzansi et al., 2022). Our study pushes the frontier by comparing the systematic use of digitized data (in the form of the algorithm) with the value of the tax inspectors' own knowledge and experience. Unlike lower-level tax collectors studied in other papers, whose performance could be enhanced by information provision, the inspectors in our study cannot be easily outperformed by the systematic use of available information, consistent with their high qualifications.

2 Context and Data

2.1 Tax Policy and Enforcement

Senegal has low tax revenue and experiences widespread tax evasion. Tax revenue as a share of GDP was only 16.4% in 2018, similar to levels observed in other lower-income countries. The tax gap stands at about 5-6% of GDP since 2017 (World Bank 2024). Senegal's Ministry of the Economy estimates evasion rates of 35% for the personal income tax, 30% for Value Added Tax (VAT), and 26% of the Corporate Income Tax (CIT) (Faye et al. 2022). Importantly, these estimates are based on the share of evasion that is realistically detectable through enforcement,

not on a full-compliance scenario, which would yield much higher evasion estimates.

Senegal's tax structure is typical of a lower-middle-income country. The Value Added Tax is the largest source of revenue, representing slightly more than 29% of total tax revenue in 2022, followed by corporate and personal income taxes (29% of total tax revenue) and customs duties (15%). The VAT applies monthly at a standard rate of 18%, with a reduced rate for selected activities (e.g. tourism). The Corporate Income Tax is paid annually at a rate of 30% of profits or 0.5% of turnover, whichever is larger. Small firms with a yearly turnover of under 50 million CFA Francs (about 100,000 USD) are eligible for a simplified tax (*Contribution Globale Unique*, CGU) on turnover, which replaces VAT and CIT and has rates varying from 1% to 8% depending on economic sectors and turnover. Firms withhold the Personal Income Tax (PIT) at source for their employees, also referred to as Paye-As-You-Earn (PAYE).

The tax administration (*Direction Générale des Impôts et des Domaines*, or DGID) is tasked with enforcing the tax code. Its main enforcement tool is the annual audit program. Audits are implemented by 26 different tax offices, with taxpayers segmented by size, region, and sector (Figure A.1). There are two types of audits: full audits, carried out by a team of inspectors at the taxpayer's premises; and desk audits, conducted by correspondence by an individual inspector using the firm's tax returns and third-party reported data. Most audits are targeted at firms.

At the beginning of each calendar year, audit offices prepare a list of taxpayers to audit based on a target provided by the tax administration. This list is reviewed and approved by senior management, usually without any changes. Before our intervention, the selection of cases for tax audits was discretionary, meaning it did not follow an explicit rule and no rationale was required for desk audit selection. Management sometimes requested a rationale for the selection of full audits (e.g., an analysis of the firm's audit history or of its sales and profit margin), but such the rationale could take different forms and was not systematically required. Thus, the criteria used for case selection varied across units.

Figure A.2 outlines the steps of the audit process. An audit starts with a notification letter to the taxpayer. After examining a case, inspectors list the detected irregularities and corresponding penalties, sending a notification letter to the taxpayer. They may also request additional information. Upon receiving the notification, taxpayers have 30 days to respond. The inspector then reviews the response and has 60 days to issue a final confirmation with the irregularities, penalties, and the final amount due. A revenue order is then generated for the tax collection unit, requiring the taxpayer to make payment within ten business days.

2.2 Data

During the last decade, DGID invested in digitizing its tax information and required taxpayers to file electronically. As a result, data availability has expanded substantially. Our study draws on data from three administrative sources and two surveys. The administrative data include the

self-assessment declarations filed by taxpayers, third-party reports, and audit reports. The tax declarations and third-party data were used in the calculation of the risk scores and had just been digitized at the onset of the project. The audit reports were hand-collected and digitized by our research team after the end of the intervention in 2021 and 2022. To study how audits were implemented, we complement the administrative data with a tax inspector survey and a taxpayer survey designed by the research team. Only aggregated results from these surveys were shared with the tax administration.

Tax Declarations. Table 3, Panel A, provides an overview of the available tax declarations covering the period of 2014-2019. The CIT data covers about 5,000-7,000 firms per year, and the VAT data twice as many.[4] The PAYE data provides the number of employees and the aggregate wage bill for each firm. Around 2,000 firms file tax under the simplified regime CGU. About 150 financial institutions pay the *Taxe sur les Activités Financières* (TAF), a VAT-substitute for the financial sector.

Third-Party Reports. Table 3, 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 digitally since 2017. We aggregate these transaction-level data at the firm-year level to merge with the tax declarations.

Audit Reports. We collect audit process and outcome data from two sources (Table 3, Panel C). First, we digitized all audit reports for 2017-2020. These reports cover all process steps from audit announcement to notification, confirmation, and payment request. They identify the inspectors conducting the audit, the taxes verified, infractions detected, evaded amounts, applicable penalties, and the dates of each audit stage. Second, we asked inspectors to complete pre-filled Excel sheets listing their audit cases. These provide information not recorded in the audit reports, such as the number of days inspectors spent on each case.

Tax Inspector Survey. In 2017, prior to our intervention, we conducted a detailed survey among 97 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 sample includes almost all tax inspectors involved in tax audits at the time, except for those who were unavailable throughout December 2017 and January 2018. Descriptive statistics are presented in Section 2.3.

Taxpayer Survey. After the completion of the 2018 and 2019 audit programs, we surveyed 742 firms in the Dakar region, most of which had been selected for audit as part of the program. We conducted the survey in two waves, from October to December 2020 and from March to May 2021. The survey allowed us to elicit taxpayers' experience with the audit process, including perceptions of corruption, audit risk, and their view of the tax administration. The first

[4]Self-employed individuals and unincorporated firms may file VAT but not CIT.

survey wave focused on the 2018 audit program. We sampled all full audit cases of our program; a matching number of desk-audit cases, randomly sampled and stratified across offices to preserve the relative distribution of planned audits across offices; and a matching number of non-audited cases, similarly stratified. This yielded a total of 1226 targeted firms. For the second survey wave, we targeted an additional 702 firms, all selected by stratified random sampling. We asked surveyors to prioritize interviews of full audit cases (27% of the sample), and the response rate for these cases was slightly higher (42% vs 37% for other cases).

2.3 Tax Inspectors

The key bureaucrats in the tax enforcement process are the tax inspectors and their managers, who are themselves former tax inspectors. Tax inspectors form an elite corps in the public administration and often proceed to hold influential political roles. Selection into the tax administration function is highly competitive. Most inspectors have completed a Master's degree from the *Ecole Nationale d'Administration* (ENA), and over 60% hold a doctorate degree. According to our survey, the mean and median inspector has ten years of professional experience.

Tax inspectors are very well-paid. Annual base salaries are typically 5-6 million FCFA in the period we study, and stand at 7-8 million FCFA (approximately 12,000 to 13,500 USD) in 2024. In addition, inspectors receive a bonus that depends on the amount of evasion uncovered in audits, and that is usually higher than the base salary. The median (mean) bonus for inspectors in our data was 7 million FCFA (10.6 million FCFA). The size of the bonus varies across inspectors, e.g. with seniority (details in Appendix A.2), but quantifying the exact strength of incentives is difficult, as it depends on the endogenous assignment of inspectors to cases. There is no institutional reason for incentives to be correlated with the case selection mechanism.

According to the survey, when selecting cases for audits, inspectors typically start from a long list of cases to consider, roughly half of which they chose with no particular suspicion. Inspectors then examine the long list of cases to derive a short list. Figure 1, Panel A, shows inspectors' reported objectives in audit selection. The most common objectives were the diversity of audits (i.e. creating the perception of a non-zero audit probability for most taxpayers, which we interpret as an interest in the deterrence function of audits, discussed in Section 6.4) and the detected evasion amount. A list experiment embedded in the survey revealed that 45% of inspectors prefer to audit larger firms, even compared to smaller firms with the same level of evasion but a higher evasion rate. Larger firms are perceived to yield higher evasion amounts and to have better material (books, receipts etc.) that allow documenting taxable transactions.

Panel B shows that there is a clear ranking in the types of information used in audit selection. The most commonly used data sources are the self-assessed declarations, followed by third-party data (70% of cases), with soft information being a distant third. Over 60% of inspectors report finding quantitative information more useful than soft information. Most inspectors use

Excel to analyze taxpayers' quantitative information. In response to the survey's open-ended question about proxies of evasion, frequently-mentioned indicators include low turnover compared to other variables, frequent losses, and high VAT credits. Respondents almost unanimously agree that turnover is the single most important variable they consider.

Inspectors are critical of the status quo: 55% of inspectors consider that they do not have sufficient data for audit selection, and almost all want more advice on case selection. Concretely, 85% of respondents agree that a more systematic analysis of data would be beneficial and 70% support an automated selection based on risk indicators.[5]

To summarize, tax inspectors are highly-trained elite civil servants, strongly incentivized to maximize detected evasion in audits, and motivated to use new tools to improve audit selection.

3 Conceptual Framework

This section presents a framework for thinking about how the algorithm might impact the audit process and audit outcomes. We start by stating the assumptions required for a fair comparison of algorithm and inspector-selected cases. First, the algorithm must maximize the same objective as the government (social planner): the audit yield, within each tax office.[6] We explain below which elements in the algorithm design ensure this.[7] Second, inspectors' incentives (bonus, sanctions for bribery etc) and the audit process must be the same for both case types. There was indeed no difference in monetary or non-monetary performance incentives across case types and all audits are based on the same regulatory framework. Third, we assume that inspectors are not inherently biased against the algorithm, consistent with their reported enthusiasm for an algorithmic selection method.

Note that, while the algorithm and the social planner both maximize yield, the inspectors (whose case suggestions constitute the government program) maximize their private returns: the expected bonus plus bribes minus effort costs. This difference in objectives can drive a wedge between the social planner's selection and the inspector's selection. The algorithm would eliminate this wedge even if it had the same data and processing model as inspectors.

Given the above-stated assumptions, we lay out in Table 1 how the algorithm may affect the audit yield along the different steps of the audit process. We consider the different determinants of the audit yield (case riskiness, inspector effort, corruption and bias) in the columns, and

[5] This is consistent with the fact that inspectors report spending a non-trivial 27% of their time deciding which cases to audit, while they spend 52% of their time on the actual audits, and the rest on administrative tasks and complaints. An "effortless" selection via the algorithm would free more time to implement audits.

[6] This is equivalent to maximizing the net yield, i.e. the total yield - total cost, as costs are essentially fixed: inspectors are tenured civil servants and cannot easily be redeployed. In the analysis, we consider cost-effectiveness as an additional outcome and find that the results are qualitatively the same as the results for yield.

[7] Roughly, the algorithm uses dummy indicators for proxies for evasion, then weighs them by the size of the detected gap/issue (e.g. misreporting rate), and then re-weights firms by size to give extra weight to large firms.

separate impacts by the stage in the audit process (selection or implementation) in the rows. In each cell, we only focus on the direct effect of the algorithmic selection, ignoring indirect effects (e.g. those mediated by changes in other cells in the table).

The primary channel through which we expect the algorithm to affect audit yield is by changing the riskiness (detectable evasion level) of an audit case (column 1). The algorithm processes all machine-readable information in a systematic manner, while the inspector selection relies on a combination of codified information and private information processed through the inspector's own mental model. The algorithm raises the average riskiness of cases on the annual program if i) the data quality is sufficiently high, and/or ii) the value of inspectors' private information, and/or iii) the quality of their mental processing model is sufficiently low. In addition, algorithm selection may affect audit start and implementation. If inspectors' prior knowledge of a case, for instance of the risk flags, facilitates audit implementation, the algorithm may adversely impact the audit implementation and hence yield.

The second channel through which the algorithm may affect audit yield is inspector effort (column 2). Because algorithmic selection requires no inspector time (after the fixed cost to design the algorithm in incurred), inspectors can reallocate effort from case selection to implementation (or to other tasks). However, by reducing inspector autonomy, the algorithm may lower motivation and effort in implementation.

Third, the algorithm can affect audit outcomes through changes in corruption (column 3). While algorithmic selection eliminates opportunities for corruption in case selection, firms may still bribe inspectors to avoid the implementation of a planned audit or to reduce detected evasion. Whether such corruption increases or decreases under algorithmic selection is ambiguous, as both firms' willingness to offer bribes and inspectors' willingness to accept them may differ by case selection method. Finally, the algorithm eliminates biases inherent in human selection (e.g. discrimination against specific types of firms, owners or sectors), and instead introduces biases driven by data limitations (column 4). It is ex-ante unclear which of these two types of biases would drive a bigger wedge between the socially optimal and the actual audit selection.

The link between the conceptual framework and our empirical analysis is indirect because we cannot directly observe concepts such as true case riskiness, inspector effort, or biases. We therefore organize the empirical analysis around observed outcomes and their underlying mechanisms. Taken together, the evidence allows us to assess the validity of the framework's implicit hypotheses—that is, the different cells of the framework. We return to this in the conclusion.

Table 1: Conceptual Framework for Algorithm Impact on Audit Process and Audit Outcome

	Determinants of Audit Yield (Detected Evasion)			
Audit Process Stage	**Riskiness of Cases**	**Inspector Effort**	**Corruption**	**Inspector Bias**
Selection	Algorithm increases or lowers average riskiness of cases depending on i) quality of data, ii) value of inspectors' private information, iii) quality of inspectors' mental processing model	Algorithm reduces time spent on selection, allowing for more time (effort) spent on implementation	Algorithm eliminates corruption	Algorithm replaces human bias with bias due to data limitations; effect on yield is ex-ante ambiguous
Implementation	Algorithm hampers implementation if inspector knowledge of case (e.g. risks) is relevant	Algorithm may reduce motivation and effort, as it reduces inspector autonomy	Algorithm can increase or decrease corruption	No change in bias (only indirect change due to change in pool of cases)

Notes: This table presents a conceptual framework for how the algorithm would impact the determinants of audit yield (in the columns) along the different steps in the audit process (rows). The riskiness of the case is the true detectable evasion level, which can differ from the ultimately detected evasion, e.g. due to variation in effort or bribery. Table 2 summarizes our results using the framework presented in this table.

4 Experimental Design

4.1 Intervention: Risk-Based Audit Selection

We collaborated with Senegal's tax administration to design a risk-score algorithm for audit selection and to deploy it in a controlled manner, thus allowing for a rigorous evaluation of its impact. The objectives of the algorithm were a) to ensure that audit selection followed objective and quantifiable criteria and b) to increase the detected evasion amounts.

We faced two constraints when designing the algorithm. 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 audit outcome data, which was available only for a small and selected subset of audits. Accessing and digitizing all paper records of audit results, which were held in individual folders in each tax office, came in a variety of formats, and were considered sensitive information by the administration, took several years. Given the constraints, we designed a "best-practice" risk score based on intuitive indicators discussed with and validated by the tax administration. The choice of indicators drew on technical assistance work from the World Bank in Pakistan and Turkey, best practices shared by the Danish tax administration, and feedback from experts at the World Bank and the IMF. Our strategy is broadly applicable to other lower-income countries where similar constraints on audit selection likely bind.

The algorithm generates a risk score for each firm, allowing us to rank firms within a tax office. The risk score combines two classes of risk indicators: inconsistencies and anomalies (Table B.2). 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, constructed as the sum of exports, procurement contracts, and purchases declared by other firms. In contrast, anomaly indicators are *across-firm indicators* that flag outlying behavior potentially associated with tax evasion relative to the firm's peers. For example, one anomaly indicator flags firms with a low profit margin relative to firms of the same economic sector and similar size.

Each inconsistency and anomaly is captured in the form of a ratio (e.g. third-party reported over self-reported turnover). We assign points from 0 to 10 based on the deciles of the ratio to take into account the severity of the irregularity. We then aggregate the points using importance weights by indicator. We weighted inconsistencies higher than anomalies as inconsistencies were more likely to reflect non-compliance, while anomalies could reflect temporary economic difficulties or poor management. Finally, we reweigh the aggregate firm-level risk scores by the logarithm of firm size (turnover). This is because the algorithm is intended to maximize detectable evasion *amounts*, which the decile-based indicators capture only imperfectly.[8] Appendix B describes the construction of the risk score in detail.

4.2 Study Design

To evaluate the performance of the risk-score algorithm, we introduced algorithm-selected audits in all audit lists. We asked each office to report the total number of planned audits for the year. Inspectors were then tasked to select half this number of cases following their discretionary method, while the algorithm selected the other half of the cases. To obtain the algorithm selection, we ranked firms by risk score within a tax office and selected the top cases on the list until the required number of cases was reached. We had access to the inspector-selected cases before running the algorithm selection, allowing us to tag overlapping cases selected by both methods.

Table 4 displays the number of cases selected by tax office and selection method. Table B.3 provides a more detailed count for each of the years 2018-2020.[9] Each year, the algorithm's indicators and weights were updated based on lessons learned. A fair comparison between algorithm- and inspector-selected cases requires that both methods select the same number of cases per office $\times$ year. While the numbers are similar, they are not always identical. Section 5.4 discusses those deviations in detail and shows that our results are not sensitive to them.

[8] Having amounts-based scores for each indicator is impractical due to the high dispersion and noise in the data, so we chose to do the size-based correction after aggregating the indicator-specific scores.

[9] The intervention started with the four large and three medium taxpayer offices in 2018 and included four regional tax offices in 2019. Additional regional tax offices joined the program in 2020, but we were not able to digitize audit outcome data for these offices. Our evaluation hence focuses on the eleven offices with complete data.

An alternative research design would have been to randomize the use of the algorithm across tax offices. This was infeasible due to political constraints and the small number of tax offices conducting auditsmany of which joined our program only in the third year. The combination of both selection methods is also the more realistic policy, as fully replacing discretionary selection risks reducing revenue and could breed resistance among inspectors. Instead, with our design, both inspectors and managers were initially enthusiastic to use the algorithm, as evidenced in workshops we held with all tax offices. Our design effectively allows us to study the combination of algorithm selection with inspector selection: although the annual audit program was supposed to be fully implemented, inspectors actually picked only a subset of cases from the list to audit. We discuss this incomplete execution in Section 5.

We randomized the order of case types displayed on each list and asked inspectors to adhere to the proposed ordering. We intended to ensure that both case types were treated similarly in terms of timing and effort exerted to complete the audit. Inspectors were provided with a protocol emphasizing the importance of conducting all audit cases with the same rigor.

In the desk audit program, which has more cases than the full audit program, a subset of cases was selected randomly, drawing from the pool of all firms in the relevant tax office. Both randomly and algorithm-selected cases appeared in the lists as having been selected by the "new method", so that inspectors would not de-prioritize randomly-selected cases.

4.3 Empirical Strategy

Our empirical exercise compares outcomes for algorithm-selected vs inspector-selected cases within the same audit list. We estimate the following Ordinary Least Squares model:

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

where $y_{i\ell}$ is the outcome of an audit for firm i selected in audit list ℓ, the dummies $Algorithm_{i\ell}$, $Overlap_{i\ell}$ and $Random_{i\ell}$ indicate algorithm-selected, overlap and randomly-selected cases, γ_ℓ denotes the list fixed effects and $\varepsilon_{i\ell}$ is a conditional mean zero error term. The audit lists are tax office-year specific for full audits and can be inspector-year specific for desk audits. The dummy for random selection is applicable only for desk audit cases. Note that overlap cases appeared on the lists sent to inspectors as inspector-selected cases. The list fixed effects control for any office and year-specific factors, such as the identity of the office manager, the fact that more experienced inspectors are more likely to be assigned to the large and medium taxpayers offices, or that inspectors in regional offices spend a larger share of their time on non-audit administrative tasks, such as taxpayer service and information campaigns. The main coefficient of interest is β_1. We use robust standard errors for inference.

Any difference in audit outcomes between algorithm-selected and inspector-selected cases we observe can be due to a combination of a change in the type of taxpayer selected and a possible

change in inspector type or skill set. This is because tax office managers assign full audit cases to teams of inspectors. Although we emphasized that all audit cases should be treated the same way, we cannot preclude the possibility that managers take into account the case type when assigning inspectors with different skill sets to cases. We investigate this further below.[10]

4.4 Descriptive Statistics of Firms Selected for Audits

Table 5 examines the characteristics of firms selected for a full audit. Across all tax offices, about 150 firms per year were selected by inspectors, and a similar number was selected by the algorithm.[11] Panel A compares the characteristics of firms selected by inspectors to the general population of firms, pooling the years 2018-2020. Inspector-selected firms are orders of magnitude larger than the average firm appearing in the administrative tax data (columns 1 and 2), more likely to engage in import or export (column 4), and older (column 6). Inspector-selected firms do not significantly differ from other firms in their location (columns 5).

Panel B compares firms selected by the algorithm to those that were selected by inspectors, using equation 1. Algorithm-selected firms are smaller than firms selected by inspectors, though still much larger than the average formal firm (columns 1-2).[12] They also have lower profit rates than inspector-selected firms but are more likely to engage in international trade (columns 3 and 4) and tend to be older (column 6). The difference in profitability appears by construction, due to the choice of risk indicators. Our survey data also indicate that algorithm-selected firms have a higher share of sales in cash than inspector-selected firms (column 8), a characteristic which the algorithm does not observe but inspectors may. A high perceived share of cash sales may lead them to avoid auditing a firm, as it would be hard to document tax evasion on cash sales.

5 Audit Outcomes

We examine the association between algorithm selection and audit outcomes. We first study audit execution, given that only a subset of programmed audits was implemented. We then analyze evasion detection, evasion amounts, and taxpayer disputes. Finally, we assess the robustness of our results to potential sources of bias.

5.1 Audit Execution

In the inspectors' audit reports, we observe all audits that are implemented, including those that do 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 execution rate improved slightly from 2018 to

[10] We do not expect a change in taxpayer behavior, as they were not informed how they were selected for audit.
[11] Table B.4 shows the characteristics of firms selected for desk audits.
[12] The firm-size-selection gradient is present both within and across tax offices (results available upon request).

2019, but then worsened in 2020 during the COVID-19 pandemic.[13] Our analysis considers all audits reported until the end of 2021, hence including audits that were executed in the year(s) following their original selection.

Table 6, columns 1-3, shows how the execution of algorithm-selected cases compares to inspector-selected cases. Full audit cases selected by the algorithm are 17 percentage points less likely to be implemented than inspector-selected cases, corresponding to a 32% reduction compared to the mean implementation rate (column 1). "Overlap" cases, which were selected by both the inspectors and the algorithm are 16 percentage points more likely to be implemented than cases selected only by inspectors.

The results for the desk audit program are qualitatively similar, but the differences are much smaller (columns 2 and 3). Algorithm-selected desk audits are about 4 percentage points less likely to be implemented, which corresponds to 13% of the mean. Our preferred specification is the one controlling for inspector fixed effects (column 3), as desk audit cases are selected and conducted by the same inspector. Overlap cases are again more likely to be implemented, though the coefficient is not statistically significant. Random cases (only included in desk audits lists) are no less likely to be implemented than algorithm-selected cases, presumably because the two case types appear in the same way to inspectors (as selected by "new methods").[14]

5.2 Audit Yield

Table 6, columns 4-6, show the relationship between the selection method and the probability of evasion detection, conditional on execution.[15] Overall, 89% of executed full audits and 73% of desk audits detect evasion, with no significant differences across selection methods. Only the coefficient on overlap cases in column 4 is economically meaningful and statistically significant at the 10% level. Columns 7-9 show the results for the amount of detected evasion, conditional on positive detection. Algorithm-selected full audits detect significantly smaller amounts of tax evasion plus fines. For desk audits, the difference is not statistically significant once we control for inspector fixed effects (column 9).[16]

[13] As shown in Section 5.4, execution rates are higher in 2019, when the project team devoted more resources in monitoring implementation and adherence to the experimental protocol, and lower in 2020, when the COVID-19 pandemic disrupted audit execution. Despite the incomplete implementation of the 2018 audit program, the tax administration emphasized their goal to fully implement each annual program. As a result, the number of cases was reduced by a third in 2019 for large taxpayer offices, and by 20% for medium taxpayer offices (Table B.3).

[14] On the other hand, replacement cases, which appear at the bottom of the list sent to inspectors and are explicitly marked as replacements for algorithm-selected cases that turned out to be void (e.g. taxpayers that had become inactive) are 47% less likely to be implemented than other algorithm-selected cases (coefficients not shown). These cases are labeled as replacement cases (rather than algorithm x replacement) on the list sent to inspectors, but audit lists did not previously include replacement cases. So it is reasonable to assume that inspectors perceive these cases as algorithm-selected cases of lower priority ranking.

[15] Intent-to-treat analyses are shown in Table C.1. In these regressions, evasion detection and evasion amounts are zero for non-executed cases, which reduces the coefficient on "algorithm".

[16] Similarly, we do not detect statistically significant differences by selection method 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 (Table C.2).

Since the analysis in columns 4-6 and 7-9 is conditional on audit execution and detection of evasion, the results are affected by non-random attrition: it is unclear what the results would look like if all planned audits had been conducted. To examine whether our results are affected by sample selection, we conduct a sensitivity analysis by computing upper and lower bounds on these effects as proposed by Lee (2009). This analysis assesses scenarios by eliminating the highest or lowest values in the group with fewer missing observations (algorithm or inspector-selected cases). When the outcome is the binary detection dummy, we randomly trim a share of cases among the algorithm-selected firms with detection (without detection) to obtain a lower (upper) bound. When the outcome is log(evasion), we rank cases by the outcome variable among algorithm-selected cases, and trim log(evasion) at the top (bottom) cases to obtain the lower (upper) bound estimates. The share of cases to trim is list-specific, calculated based on the list-specific attrition rates. The estimated bounds are noted in Table 6 within square brackets below the main point estimate on algorithm selection. These are consistent with our main results: we cannot reject the null of no correlation between the selection method and detection of evasion, but we find a negative and statistically significant association between algorithm selection and the amount of evasion for full audits.

To summarize, algorithm-selected audits are less likely to be executed, similarly likely to detect evasion when implemented, and uncover smaller amounts of evasion. Results are starker for full audits compared to desk audits, which is consistent with the fact that full audits involve higher stakes for both the administration and the taxpayer.

5.3 Dispute of Audit Outcomes

Once an audit concludes with the detection of tax evasion, the taxpayer receives a notification and the possibility 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 (e.g. because of errors in how the audit was conducted, uncertainty over the exact amount evaded, or collusive behavior between taxpayers and inspectors to alleviate the penalties).

Algorithmic case selection might reduce uncertainty and collusive behavior. To test this hypothesis, we express the confirmation amount as a share of the notification amount, and plot its distribution for inspector- and for algorithm-selected cases, separately for full and desk audits (Figure 2).[17] The confirmation amount matches the notification amount in only 22% of completed full audits and 29% of desk audits. On average, the confirmation amount is only 40% of the notification for full audits, and 62% for desk audits. For full audits, inspector-selected cases are subject to more push-back against the notification amount, leading to a lower confirmation-to-notification ratio.

[17]The figure excludes cases with no confirmation, which represent about a third of these cases.

Table C.3 shows these results in regression form. Among full audits for which both a confirmation and a notification in our data, algorithm audits are 12 percentage points more likely to have matching confirmation and notification amounts. This reduction in dispute lowers the gap in (confirmed) detected evasion between algorithm and inspector-selected audits. For full audits, algorithm audits have a 35% lower notified amount and only a 19% lower confirmed evasion amount than inspector-selected audits (columns 3 and 5).[18] Although the lower likelihood of disputes for algorithm audits slightly improves the relative return of algorithm-selected audits, it does not close the gap in detected evasion between algorithm and inspector-selected audits.

5.4 Robustness

We assess the robustness of our results. 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 decided that 30% of the full audit cases would be selected by the algorithm, instead of 50%. Other deviations occurred when managers adjusted the number of cases selected at discretion slightly upwards or downwards. If both inspectors and the algorithm had ranked cases by perceived risk and had selected cases following the rank order, the average riskiness of selected cases would be higher when a smaller number of cases was selected. We hence rerun our results for the following sub-samples: 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 weakly smaller than the number of inspector-selected cases. The latter subsample is less restrictive than the first two and useful from a robustness perspective, as it would give the algorithm cases an advantage, hence going against our main results.[19]

Second, tax offices and years differ substantially. The Large Taxpayer Unit (LTU) has the best inspectors and the lowest caseload per inspector. Although large firms are more complex, inspectors are more familiar with them, and they are audited every four to five years, more frequently than smaller firms. The association between audit selection and outcomes may therefore differ in the LTU. We therefore rerun our analysis excluding the LTU.

[18] The result is similar when we control for firm size. For desk audits, there are no significant differences in dispute of outcomes between inspector and algorithm-selected audits. The point estimates in this section are slightly different from those in Table 6, which does not impose the requirement that all audit cases have both a notification and a confirmation. We find no statistically significant association between the size of the confirmation-notification discrepancy and taxpayer perceptions of corruption or audit efficiency, as analyzed in Section 6.3.

[19] The difference between sub-samples a) and b) comes from a methodological change in how overlap cases (chosen by both the inspector and the algorithm) were dealt with. In 2018, we ran the algorithm to select a number of cases equal to the number of inspector-selected cases, allowing for overlap cases. The overlap cases appear to inspectors as inspector-selected cases. Thus, the number of "pure" algorithm cases appearing on their list is weakly lower than the number of inspector-selected cases. In 2019 and 2020, we added an additional algorithm case for each overlap case. Overlap cases still appear on the inspectors' lists as inspector-selected cases.

Third, the experiment's implementation differed in each of the three years. The year 2018 was the first one and 2020 was affected by the pandemic. The 2019 implementation is the cleanest from a research perspective, as we carefully prepared and tightly monitored the execution of the intervention. We thus re-estimate the results for the 2019 program only.

Figure 3 plots the coefficients on the algorithm selection indicator from Equation 1 for our three main outcomes: audit execution, detection of 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 estimates are stable across subsamples, and inference methods, confirming that our findings are qualitatively and quantitatively robust. For full audits, algorithm cases are less likely to be conducted, equally likely to detect evasion (though point estimates are positive), and uncover significantly smaller evasion amounts. For desk audits, the estimates are almost always small and statistically indistinguishable from zero. The negative association between algorithm selection and audit execution is confined to the full sample and disappears when we restrict the analysis to lists with more comparable numbers of algorithm and inspector-selected cases.

6 The Audit Process

6.1 Time and Human Resource Costs

We next examine how human resource costs and audit length vary by selection method among executed audits. Table 7, Panel A1, shows the analysis for the number of inspectors working on a case and for various proxies of audit length for full audits (Table C.4 shows desk audits). The most accurately captured measure of audit cost is the number of inspectors working on an audit, which is always reported in the administrative data. Recall that managers assign inspectors to cases. Column 1 shows that algorithm-selected full audits are conducted by teams that are about 10% smaller than the average team of 2.9 inspectors. In addition, teams for algorithm-selected audits have lower seniority, as measured by age and experience (Table C.5).[20]

Columns 2-4 of Table 7, Panel A, show the association between audit selection and three measures of audit length. All measures suggest that algorithm-selected full audits are shorter. In column 2, the outcome is the audit duration as reported by firms subject to a full audit from the taxpayer survey. The mean duration is 28 days, and algorithm-selected audits are approximately nine days shorter on average, a quantitatively large difference. In column 3, we use the audit duration as measured by the difference between the date of notification and the audit start

[20] These differences in team composition are consistent with the fact that, for desk audits, inspectors with above-median experience are less likely to favor using an algorithm for audit selection (Table C.6). Estimating inspector-quality fixed effects, we find large heterogeneity in inspector skills (Figure C.1), suggesting the algorithm may be more beneficial to low-skilled inspectors. We do not detect a correlation between overall list-level execution rates and algorithm execution rates (Figure C.2). However, we find that the number of inspectors on a (full) audit team mediates the association between algorithm selection and audit outcomes (results available upon request).

date. This measure captures the length of an audit but not the time inspectors actively spend on the case, as they work on several cases simultaneously. The average full audit takes 160 days until notification. Algorithm-selected audits are completed 30 days faster.

Finally, column 4 shows the self-reported number of days inspectors work on a case. For full audits, we multiply the reported days by the number of inspectors working on a case. Full audits require 177 inspector-days on average. Algorithm-selected audits require 27 less inspector-days, though the estimate is not statistically significant. This is unsurprising given the smaller sample and measurement error. More generally, the duration measures in columns 2–4 are not consistently available for all completed cases. However, Panel A2 rejects differential outcome availability across case types for all but one measure (column 4).[21]

In addition to the detected evasion amount, audit reports indicate the specific year of the infraction and a short description.[22] To test if the scope of inspections varied by selection method, we run regressions using as outcomes the number of years covered in the audit report, the number of detected infractions, and the share of fines to evasion amount, which indicates the severity of detected infractions. The results show that in algorithm cases, inspectors systematically reported fewer years with any infraction and fewer infractions, both for full and desk audits (Table C.7, columns 1-4). All differences are statistically significant at conventional levels. There is little evidence of systematic differences in the fine to evasion ratios (Table C.7, columns 5-6).[23]

6.2 Audit Productivity

Given the differences in audit outcomes and audit costs, the question of whether algorithm-selected audits may be more cost-effective than inspector-selected audits arises. Table 7, Panel B1, shows the results for our main specification in Equation 1, using audit productivity measures as outcomes (similar results for desk audits are shown in Table C.4). We consider the log of the ratio of uncovered evasion over different audit cost measures used in Panel A1: the number of inspectors working on a case (column 5), the audit length in number of days from the survey data (6), the number of days from opening to closing a case from the administrative data (7), and the self-reported number of days the inspectors worked on the case (8). The association between algorithm selection and audit productivity is negative in all estimations. The coefficient on algorithm selection is statistically significant when using the number of inspectors as a cost measure or the number of days from the administrative data. Together, these results imply that, although algorithm audits use fewer resources, the difference in evasion is larger than the difference in audit cost, leading to lower productivity for algorithm audits.[24]

[21] The results in Table 7 are almost unchanged when we control for turnover.

[22] By Senegalese law, an audit can investigate tax declarations up to four years prior to the audit start date.

[23] Panels B and C of Table C.7 show that there are also no statistically significant differences between algorithm and inspector-selected audits on a range of other outcomes that capture the severity of the detected infractions.

[24] Panel B2 reports the availability of outcome measures across audit selection methods, showing small but significant negative point estimates in columns (7) and (8). The lower availability of outcome data for algorithm-

The averages captured by the regression coefficients in Table 7, Panel B1, mask heterogeneity across cases. Figure C.3 shows distributions of the three main audit productivity measures for algorithm and inspector-selected cases. The productivity distributions for inspector-selected cases are shifted slightly to the right for all measures. For our preferred measure of productivity — the evasion amount divided by the number of agents on the case — the two distributions are statistically significantly different as per the Kolmogorov-Smirnov test. Yet the two distributions overlap substantially, suggesting that many algorithm-selected cases have higher productivity than the marginal inspector-selected case. This rationalizes the fact that execution of the inspector-selected set of audits is incomplete and algorithm audits are also implemented. Note that audit productivity is positively correlated with firm size, consistent with the preference for larger firms in audit selection and execution (results available upon request).

6.3 Taxpayers' Perceptions

An alternative measure of inspectors' effort and performance comes from the taxpayer's point of view. We surveyed around 600 firms that had been selected for audits and asked them about their perceptions of interactions with the tax authority. The main objective was to consider the taxpayer's perspective of the audit process, in terms of efficiency and incidence of corruption.

We construct two indices capturing the perceived efficiency of audits and perceived prevalence of corruption during audits. Each index combines answers to three questions following the procedure in Anderson (2008). To measure efficiency, we combine taxpayers' assessment of the auditor's (i) technical knowledge, (ii) efficiency during the audit, (iii) capacity to uncover all evaded taxes. Inspectors obtain an average score of 6.4/10 on honesty, lower than their knowledge score of 7.3/10, but equal to their efficiency score. For corruption, we combine responses on (i) the auditor's dishonesty during the audit, (ii) the frequency of bribes paid to inspectors among firms similar to the respondent's, (iii) a preferential treatment for firms connected with the tax administration. Obtaining information on corruption incidence is challenging (Sequeira, 2012). Yet, taxpayers report that bribes are paid in 16% of audits (64% response rate), and 27% believe that connected firms receive preferential treatment (89% response rate).

Table 8 shows the regression results for the association between taxpayer survey responses and algorithm selection. Panel A uses the sample of interviewed firms that were part of the annual audit program and reported being audited, while Panel B conditions on firms confirmed as audited in our administrative audit reports.[25] Columns (1) to (3) show the results of taxpayers'

selected cases is consistent with inspectors spending less time on these cases (including documenting them and completing basic paperwork, e.g. an audit announcement letter or audit closure report). Indeed, our results in Panel A, column (4) and Panel B, columns (7) and (8) would be even stronger if we replaced the resource and productivity outcomes by zero for audit cases that were implemented but have no outcome data.

[25] Firms that were actually audited as per the administrative data account for half of the surveyed firms reporting an audit. This could mean that firms perceive routine interactions with the tax administration as audits or that the administrative data is incomplete, e.g. because inspectors fail to report cases with no infraction.

evaluation of audit efficiency. For full audits (column 1), firms algorithm-selected firms report a significantly lower audit efficiency score than inspector-selected firms. The magnitude is large: the efficiency index is 0.38 standard deviations lower. The coefficient is less negative and no longer significant when conditioning on confirmed audits (Panel B). Firms selected both by inspectors and the algorithm reported higher inspector evaluations. The lower perceived efficiency does not apply to desk audits, maybe as they involve less interaction (column 2).

Columns (4) to (6) display the association between audit selection and perceived corruption. We don't find evidence of differential corruption perception across selection methods. For full audits, we can reject that the algorithm improved corruption perceptions by more than 0.2 standard deviations, both for firms reporting an audit and for firms receiving an audit based on the administrative data.[26] The results do not lend support to the idea that algorithm-selection affects corruption levels.

6.4 Could Deterrence Effects Justify Using the Algorithm?

In addition to audit productivity, selection rules may affect economy-wide compliance through deterrence. Firms expecting a higher audit probability may comply ex ante, even if they are not audited. We examine whether algorithm-based selection could foster greater ex-ante compliance than discretionary selection.

Deterrence effects depend on perceived audit probabilities, which are challenging to measure (Bergolo et al., 2023). We therefore do not go all the way to examining the revenue-maximizing audit policy. Instead, we ask how large the deterrence effects would need to be for the risk-score selection to increase revenue collection. To answer this question, we estimate the deterrence effect under each selection method, and compute counterfactual reported turnover in scenarios where the audit program is either fully selected by the risk-score or fully selected by inspectors. We summarize our analysis below and detail it in Appendix D.

We focus on audit probability differences by reported turnover, the seemingly most important determinant of audit selection. Risk-score selected firms are on average 1.78 log points smaller in turnover than inspector-selected firms (Table 5). Audit probabilities rise sharply with firm size under both selection methods, but less steeply under the algorithm. Our analysis requires two strong assumptions: first, that firms know their turnover-specific audit probabilities; and second, that evasion rates by turnover in the population equal those observed among audited firms. Detected evasion rates decline with firm size, from 50% for the median firm to 20% for the largest firms (Figure D.2). We assume for simplicity that evasion takes the form of turnover underreporting rather than cost underreporting. The key parameter is the deterrence elasticity: the percentage reduction in evasion induced by a one-percent increase in audit probability.

[26] As the questions asked about interactions in "full audits", it is unsurprising that the coefficients in the desk audit sample are mostly small and statistically insignificant. For the same reason, it is difficult to explain the significant negative coefficient on algorithm-selection in Panel B, Column 5.

Considering different scenarios for the value of the deterrence elasticity, our analysis yields two insights. First, if the deterrence elasticity is uncorrelated with firm size, the inspector's selection schedule–which prioritizes larger firms–always yields higher tax revenue. This stems from the very high concentration of turnover at the top of the firm-size distribution. Second, if the deterrence elasticity is inversely related to firm size (i.e., smaller firms react more to audit threats than larger firms), then the algorithm selection could improve reporting behavior and revenue collection. Even then, the potential reporting gain peaks at just 0.2% of baseline reported turnover and only if the deterrence elasticity varies sharply with firm size.

To assess the plausibility of the latter scenario, we estimate evasion elasticities with respect to audit probability by exploiting differences in audit probabilities among firms of the same size arising from their (conditionally quasi-random) assignment to different tax offices. We find substantial deterrence elasticities that increase with firm size. Yet applying these estimates to counterfactual turnover reporting still implies larger deterrence effects under the inspector schedule than under the algorithm (Figure D.3).[27] Thus, ex-ante deterrence does not appear to favor algorithmic selection. We nevertheless emphasize that deterrence elasticities remain highly uncertain and that our study was not designed to identify them.

7 Mechanisms

The analysis so far has shown that algorithm-selected cases are less likely to be executed and are less cost-effective. We now ask whether these shortcomings reflect limitations of the original risk-score algorithm or more fundamental limitations of algorithmic selection in general, as observed audit outcomes are endogenous to inspectors' execution and effort. Using newly digitized audit outcomes, we train a machine-learning model to predict detectable evasion. We use these predictions to explain the experimental findings, showing that inspectors' execution decisions mainly reflect expected audit yield, and rule out alternative explanations.

7.1 Predicting Audit Yield

We start by leveraging the newly digitized audit outcome data to train a random forest model to predict detected evasion (audit yield) based on firms' observable characteristics and prior behavior (Breiman, 2001). As predictors of execution, we consider three lags of turnover, productivity (turnover/wage bill), the profit rate, firm age, sector of activity, the tax office and distance to it (measured in travel time and Euclidean distance). We also include a dummy indicating whether the case was selected by the algorithm. We tune the hyper-parameters (node size, number of features in split, number of trees) using semi-structured grid search, and choose

[27] We estimate a semi-elasticity of -2: a 1 percentage point increase in audit probability reduces evasion by 2%. Under this calibration, the inspector schedule raises reported turnover by 2% over the current mixed selection.

parameters that maximize out-of-sample RMSE, which are also those that minimize out-of-sample R2. Table F.2 and Figure F.1 document the performance of our prediction models.

Figure F.2 shows the importance ranking of predictors. For detection of evasion in full audits, turnover is consistently the most important predictor, followed by firm age and the profit rate. For the amount of evasion, the tax office is the most important predictor, followed by turnover and the profit rate. Firm age and distance to the tax office also play role, but a much smaller one than the firm size measures. The predictors of detection and evasion are similar in the case of desk audits, although firm age and distance measures are relatively more important here.

7.2 Inspectors' Decisions are Strategic

We can now use our prediction of detectable evasion to examine whether inspectors are strategic, i.e. they maximize detectable evasion when deciding which audits to execute. Figure 4, Panel A, plots our analysis for full audits. For cases that were audited, the distribution of the predicted and realized evasion amounts are similar, consistent with the high quality of our prediction. For non-audited cases, however, the distribution of the expected detectable evasion is shifted leftwards, indicating that these cases have lower predicted potential on average.[28] The figure also shows that the realized execution rate is upward-sloping in predicted evasion amounts. Panel B shows that the level and slope of the realized execution rate in predicted evasion is very similar for algorithm and inspector-selected cases. Hence, inspectors do not seem to prioritize inspector-selected cases, conditional on the amount of predicted evasion. The higher implementation rate among inspector-selected cases arises because the distribution of detectable evasion is shifted outward relative to that of algorithm-selected cases.

Figure E.2 provides additional evidence. Panel A shows that more inspectors are allocated to audits of firms with higher predicted evasion, for both inspector and algorithm-selected cases. The confidence intervals for the slopes overlap across selection methods, indicating similar treatment of cases conditional on predicted evasion. Panels B and C show that audit duration is also similar for inspector- and algorithm-selected audits and varies little with case potential, likely because high-potential cases are assigned larger teams. Inspectors thus allocate audit resources according to expected detectable evasion, irrespective of the selection mechanism. [29]

7.3 Inspectors Learn About Case Quality

While the previous section documented strategic behavior, a fully rational tax inspector with a perfect prediction of detectable evasion would exhibit a step function execution rule: all cases

[28] The same holds when considering separately inspector- and algorithm-selected audits (Figure E.1), part I.

[29] Conversely, Table E.1 shows that simply controlling for firm size only reduces the association between audit execution (or evasion amounts) and algorithm selection by a third, and additional firm-characteristics controls make little difference. Similarly, when we use only firms' turnover ranking to predict execution, the simulated choice of cases to execute is still far from the realized execution choice, as shown in Table E.2.

with predicted evasion above a threshold (determined by the case number) would be audited, while all cases with a lower prediction would not. The observed slope suggests that either the prediction or inspectors' adherence to the ranking of predicted evasion is imperfect. Both of these explanations would imply that inspectors might learn and adjust their behavior over time.

Table E.3 provides evidence consistent with this mechanism. We exploit the fact that the ordering of cases on each office's audit list was random, conditional on the audit selection mechanism, and that the rank order of a case on the list predicts execution. That is, cases at the top of the list are more likely to be audited than cases towards the bottom of the list. We regress list-level audit execution rates on list characteristics, focusing on an indicator for high-potential cases (high predicted evasion) among the first few cases on the list. Having a high-potential algorithm case near the top of the list (or a higher share of high-potential cases at the top of the list) is associated with a significantly higher execution rate among algorithm cases, consistent with positive updating about the algorithm among teams that conduct a high-potential algorithm audit.[30]

7.4 Ruling out Alternative Explanations

We next rule out two pre-specified explanations for why inspectors execute algorithm-selected cases at lower rates.

7.4.1 The Risk Score Predicts Audit Performance

One potential explanation for inspectors' execution decisions is a possible lack of association between the risk score and audit outcomes. To test for this possibility, Table 9, Panel A, reruns our main estimation (Table 6, column 1), adding the risk score as an additional regressor. We allow the association between the risk score and audit outcomes to differ between inspector-selected and algorithm-selected cases, as by construction, variation in the risk score for algorithm-selected cases is limited. We find that inspector-selected cases with a higher risk score are more likely to be audited (column 1), are slightly but not statistically significantly more likely to yield a detection of evasion (columns 3 and 4), and that they exhibit an economically and statistically higher recovered evasion amount (columns 5 and 6). For both full and desk audits, a one standard deviation increase in the risk score is associated with 38% more tax evasion detected. The results are not driven by firm size. Indeed, the results are almost unchanged when we control for baseline turnover (Table 9, Panel B) or use the raw risk score (without turnover weighting) as the regressor (Table 9, Panel C).[31]

[30] By contrast, low-potential algorithm cases at the top of the list are not associated with differential execution, suggesting they confirm inspectors' priors. Nor is there evidence of belief updating for inspector-selected cases.

[31] Table E.4 shows that individual risk-score components have mainly the expected association with audit outcomes, but we cannot detect with statistical confidence which components of the risk score are most predictive.

7.4.2 Providing Information does not Increase Audit Performance

Another factor that may determine inspectors' audit implementation is their information about a case. Inspectors may have limited information about the compliance risks among algorithm-selected cases and hence struggle to implement these audits. We test this hypothesis through a randomized information treatment among desk audit cases. For a third of the cases, inspectors were provided with the risk flags from the algorithm (e.g. abnormally low profit rate, turnover lower than third-party reported turnover). For another third of the cases, inspectors were provided with both the risk flags and a user-friendly spreadsheet containing the full information available from the firms' tax declarations and third-party data. For the remaining cases, inspectors were not provided any additional information by the research team.[32]

Table 10 shows that the information treatment did not affect the probability of starting a case, nor subsequent audit outcomes. Columns 1-3 bundle the two treatment arms into one, while columns 4-6 show the effects of the two sub-treatments "risk indicators only" and "risk indicators and data spreadsheet". Columns 1-3 show that the negative association of algorithm selection with the main audit outcomes remains significant, whereas the interaction with the information treatment has a small and statistically insignificant coefficient. Columns 4, 5, and 6 show that neither of the treatments had any statistically or economically meaningful effect on audit outcomes. The lower execution rate for algorithm-selected cases hence does not seem to be due to a lack of information about these cases.

8 Potential Performance of a Machine-Learning Algorithm

We now ask whether a more sophisticated machine-learning algorithm would have improved audit performance. We leverage our previous random-forest predictions of evasion to simulate the revenue potential of audits selected by a machine-learning algorithm. Concretely, we first train a two-stage random forest to predict evasion on the sample of implemented audits. In step one, we predict detection of any evasion; in step two, we predict evasion amounts, conditional on detection. We assign tenfold higher weights to cases in the top quartile of the evasion distribution to better match aggregate revenue (detected evasion and fines) than an unweighted prediction.[33] We then apply the resulting model to predict detectable evasion for all firms in the annual audit programs - those that could have faced an audit - and rank firms by predicted evasion within audit lists. Finally, we calculate predicted aggregate revenue from alternative execution scenarios, holding the number of audits per office constant.

Table 11 presents the results of our preferred simulations. Panel A, column 1, reports realized

[32] Panels B and C of Table B.5 show that taxpayer characteristics were balanced across treatment groups. For full audits, inspectors were always provided with the risk flags for all cases.

[33] Alternative specifications of the weights do not achieve as close a match between realized and predicted aggregate revenue. Table F.2 and Figure F.1 document the performance of our prediction models.

aggregate revenue from full audits across all tax offices in our study. Column 2 reports predicted revenue from the same audit program, obtained by summing predicted revenue across all implemented audits. Predicted revenue tracks realized revenue but is slightly lower, so we use it to benchmark our simulations. Column 3 reports the potential revenue gain from our main simulation, which reallocates execution across all cases in the annual audit program. Concretely, we rank cases based on their predicted evasion within each tax office and select the top-ranked N cases, where N is the realized number of audits. Aggregate revenue is predicted to increase by 18.1%. This should be viewed as an upper bound, as it requires full inspector adherence to the case selection, which based on our intervention appears unlikely in practice.[34]

Panel B shows that the potential revenue gains from optimization are higher for desk audits than for full audits (31.0% versus 18.1%). This is consistent with inspectors investing less in desk audit selection and selecting some cases at random. For the risk-score algorithm, too, desk audit selection might be further away from the optimum. This is because firms with the highest risk scores are assigned to full audits, while desk audits target firms with medium scores. Lower-score firms are possibly a mix of genuinely low-risk firms and firms with poor or missing data. Table F.3 shows the results by tax office. Consistent with lower inspector skill in small taxpayer offices, the revenue gains from optimization are larger there (Panel D).

While the revenue gains from optimizing with a machine-learning algorithm are not negligible, they are smaller in our context than in most other evaluations of machine-learning applications (see Figure F.3 and Table F.1). Our limited optimization gains are consistent both with the inspectors' skills and incentives and the challenges of building a machine-learning algorithm for corporate audits in Senegal. Inspectors are highly trained and experienced civil servants, with strong financial incentives to maximize detected evasion. The algorithm, however, is trained on just a few hundred audit cases per year, with over a hundred potential predictors from tax returns and third-party data.[35] This contrasts with the ideal machine-learning setting, which combines a large training sample with few predictors (Athey and Imbens, 2019).

In addition, Senegal's newly digitized tax data still face quality issues, including missing identifiers and inconsistent firm names that hinder cross-dataset matching. Several potentially informative sources, such as contract registries, business-to-business payment and withholding records, were digitized only after our intervention. Czajka et al. (2024) show the value of such data for tax enforcement, suggesting their inclusion could strengthen the algorithm.

[34] We also perform the optimization using rank-based predictions, but obtain smaller gains (results available upon request). We do not extend the exercise to all firms, as this would require predicting evasion for a population that differs substantially from the audit sample and suffers from more extensive missing data.

[35] As the list of potential predictors is very large compared to the size of the training dataset, we manually selected predictors from the list for our random forest. Figure F.2 shows the importance ranking of predictors.

9 Conclusion

This paper examined whether an algorithm using newly digitized data improves tax audit outcomes in Senegal. Our results (summarized in Table 2) suggest that building a simple and transparent rule that outperforms tax inspectors at selecting high-evasion firms is difficult. Inspectors prioritize the audits they selected over those chosen by the risk-score algorithm and their selection yields better results. We rule out a number of pre-specified explanations for this differential execution decisions, and show that inspectors behave strategically, prioritizing firms with higher amounts of predicted evasion. Even an ex-post machine-learning algorithm trained on outcome data would likely yield only moderate revenue gains.

Our findings contrast with a growing literature showing a large potential for rule-based decisions to improve upon the work of humans. Our analysis points to the importance of contextual features in shaping the potential for machine-learning-based enhancement. Corporate tax audits are complex and multi-dimensional tasks that require knowledge of law, accounting, data science, and context-specific information. The agents whose performance we attempted to enhance are elite civil servants with strong financial incentives. The performance of tax inspectors in Senegal is consistent with evidence on the powerful role of specialized tax administrators in generating higher tax compliance (Basri et al., 2021). By contrast, the data fed to our algorithm suffered from quality limitations, as is common for administrative datasets in lower-income countries. Even at scale in a mid-sized country, building a sufficiently large training sample would require years of audit outcomes, given the many potential predictors of evasion.

Our results also highlight the benefits of bureaucratic autonomy. In other contexts, discretion has limited the effectiveness of algorithms, as agents may pursue different objectives than the algorithm (Kim et al., 2024; Stevenson and Doleac, 2024). In our setting, however, agent discretion may have limited the cost of an under-performing algorithm. The strong performance of overlap cases suggests that inspector judgment and algorithmic selection capture complementary information. How to best combine bureaucrats' expertise with systematic data analysis remains an important question for future research. More broadly, as demand for algorithms and the expectations surrounding them grow, understanding when and where they can enhance organizational performance remains a key challenge.

Table 2: Realized Effects of the Algorithm on the Audit Process and Audit Outcomes

	Determinants of audit outcome / success / yield			
Audit process stage	**Riskiness of case (true evasion)**	**Inspector effort**	**Corruption**	**Inspector bias**
Selection	Algorithm cases are less risky than inspector-selected cases	Inspectors have more time for implementation, but invest this time disproportionately in inspector-selected cases	No evidence for change in corruption (combining corruption at different stages in audit process)	Data-based biases (errors/noise) are more important than human biases in driving a wedge between optimal and actual selection (implementation)
Start	Algorithm cases less likely to be implemented, due to lower riskiness	No evidence that lack of information about algorithm cases (risks) explains lower implementation		
Implementation	Algorithm cases have lower yield, given lower riskiness	No evidence for reduction in effort conditional on case riskiness		

Notes: This table summarizes our main results, following the conceptual framework laid out in Table 1. This table is discussed in Section 9.

References

Advani, Arun, William Elming, and Jonathan Shaw (2023) "The Dynamic Effects of Tax Audits," *The Review of Economics and Statistics*, 105 (3), 545–561, 10.1162/rest_a_01101. 6, OA20

Anderson, Michael L. (2008) "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103 (484), 1481–1495, https://ideas.repec.org/a/bes/jnlasa/v103i484y2008p1481-1495.html. 22, 38

Andini, Monica, Michela Boldrini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Andrea Paladini (2022) "Machine learning in the service of policy targeting: The case of public credit guarantees," *Journal of Economic Behavior & Organization*, 198, 434–475, https://doi.org/10.1016/j.jebo.2022.04.004. 6, OA26

Andini, Monica, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Viola Salvestrini (2018) "Targeting with machine learning: An application to a tax rebate program in Italy," *Journal of Economic Behavior & Organization*, 156 (C), 86–102, https://EconPapers.repec.org/RePEc:eee:jeborg:v:156:y:2018:i:c:p:86-102. 6, OA26

Athey, Susan and Guido W. Imbens (2019) "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11 (Volume 11, 2019), 685–725, https://doi.org/10.1146/annurev-economics-080217-053433. 28

Balán, Pablo, Augustin Bergeron, Gabriel Tourek, and Jonathan L. Weigel (2022) "Local Elites as State Capacity: How City Chiefs Use Local Information to Increase Tax Compliance in the Democratic Republic of the Congo," *American Economic Review*, 112 (3), 762–97, 10.1257/aer.20201159. 7

Bandiera, Oriana, Michael Best, Adnan Qadir Khan, and Andrea Prat (2021) "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats*," *The Quarterly Journal of Economics*, 136 (4), 2195–2242, https://EconPapers.repec.org/RePEc:oup:qjecon:v:136:y:2021:i:4:p:2195-2242. 7

Basri, M. Chatib, Mayara Felix, Rema Hanna, and Benjamin A. Olken (2021) "Tax Administration versus Tax Rates: Evidence from Corporate Taxation in Indonesia," *American Economic Review*, 111 (12), 3827–71, 10.1257/aer.20201237. 7, 29

Battaglini, Marco, Luigi Guiso, Chiara Lacava, Douglas L. Miller, and Eleonora Patacchini (2025) "Refining public policies with machine learning: The case of tax auditing," *Journal of Econometrics*, 249, 105847, 10.1016/j.jeconom.2024.105847. 6, OA26

Bergolo, Marcelo, Rodrigo Ceni, Guillermo Cruces, Matias Giaccobasso, and Ricardo Perez-Truglia (2023) "Tax Audits as Scarecrows: Evidence from a Large-Scale Field Experiment," *American Economic Journal: Economic Policy*, 15 (1), 110–153, 10.1257/pol.20200321. 23, OA15, OA20

Besley, Timothy, Robin Burgess, Adnan Khan, and Guo Xu (2022) "Bureaucracy and development," *Annual Review of Economics*, 14, 397–424. 7

Best, Michael, Jawad Shah, and Mazhar Waseem (2021) "Detection without deterrence: Long-run effects of tax audit on firm behavior," *Working Paper*. 6, OA20

Black, Emily, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho (2022) "Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1479–1503, 10.1145/3531146.3533204. 6

Breiman, Leo (2001) "Random Forests," *Machine Learning*, 45 (1), 5–32. 24

Brockmeyer, Anne and Magaly Sáenz Somarriba (2025) "Electronic Payment Technology and Tax Compliance: Evidence from Uruguay's Financial Inclusion Reform," *American Economic Journal: Economic Policy*, 17 (1), 242–272, 10.1257/pol.20220434. 7

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond (2025) "Generative AI at work," *The Quarterly Journal of Economics*, 140 (2), 889–942. 6

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016) "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review*, 106 (5), 124–27, 10.1257/aer.p20161029. 6, OA26

Czajka, Léo, Bassirou Sarr, and Mattea Stein (2024) "Fraud Detection Under Limited State Capacity: Experimental Evidence From Senegal," Working Paper No. 731, Centre for Studies in Economics and Finance. 28

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan (2013) "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India," *The Quarterly Journal of Economics*, 128 (4), 1499–1545. 7

——— (2018) "The value of regulatory discretion: Estimates from environmental inspections in India," *Econometrica*, 86 (6), 2123–2160. 7

Dzansi, James, Anders Jensen, David Lagakos, and Henry Telli (2022) "Technology and Tax Capacity: Evidence from Local Governments in Ghana," Working Paper 29923, National Bureau of Economic Research, 10.3386/w29923. 7

Faye, Issa, Abdoulaye Sow, and Sambane Yade (2022) "Evaluation du Potential Fiscal du Secteur Informel au Sénégal,"Technical report, Ministere de l'Economie, du Plan et de la Coopération'. 7

Ferreira, Alipio and Sandra Cermeno-Leon (2026) "Automated Assignment and Bureaucrat Productivity: Evidence from a Field Experiment in Peru," *Working Paper*. 6

Finan, Frederico, Benjamin A. Olken, and Rohini Pande (2017) "The Personnel Economics of the Developing State," in *Handbook of Economic Field Experiments*, 2, 467–514: Elsevier. 7

Glaeser, Edward L, Andrew Hillis, Scott Duke Kominers, and Michael Luca (2016) "Crowdsourcing city government: Using tournaments to improve inspection accuracy," *American Economic Review*, 106 (5), 114–18. 6, OA26

Haseeb, Muhammad and Kate Vyborny (2022) "Data, discretion and institutional capacity: Evidence from cash transfers in Pakistan," *Journal of Public Economics*, 206 (C), S0047272721001717. 6

Hino, M., E. Benami, and N. Brooks (2018) "Machine Learning for Environmental Monitoring," *Nature Sustainability*, 1, 583–588. 6, OA26

Johnson, Matthew S., David I. Levine, and Michael W. Toffel (2023) "Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA," *American Economic Journal: Applied Economics*, 15 (4), 30–67, 10.1257/app.20200659. 6, OA26

Khan, Adnan Q, Asim I Khwaja, and Benjamin A Olken (2015) "Tax farming redux: Experimental evidence on performance pay for tax collectors," *The Quarterly Journal of Economics*, 131 (1), 219–271. 7

Khan, Adnan Q., Asim Ijaz Khwaja, and Benjamin A. Olken (2019) "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings," *American Economic Review*, 109 (1), 237–70, 10.1257/aer.20180277. 7

Khwaja, Muwer Sultan, Rajul Awasthi, and Jan Loeprick (2011) *Risk-based tax audits: approaches and country experiences*: The World Bank, 10.1596/978-0-8213-8754-2. 2

Kim, Hyunjin, Edward L Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca (2024) "Decision authority and the returns to algorithms," *Strategic Management Journal*, 45 (4), 619–648. 29

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018) "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133 (1), 237–293, 10.1093/qje/qjx032. 6, OA26

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," *Econometrica*, 79 (3), 651–692. 7

Knebelmann, Justine, Victor Pouliquen, and Bassirou Sarr (2026) "Discretion versus algorithms: bureaucrats, tax equity and acceptability," Working Paper 26/34, Institute for Fiscal Studies, 10.1920/wp.ifs.2026.3426. 6

Kotsogiannis, Christos, Luca Salvadori, John Karangwa, and Theonille Mukamana (2024) "Do tax audits have a dynamic impact? Evidence from corporate income tax administrative data," *Journal of Development Economics*, 170, 103292. 6

Lee, David S. (2009) "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76 (3), 1071–1102, 10.1111/j.1467-937X.2009.00536.x. 18

Margalit, Yotam and Shir Raviv (2024) "When Your Boss is an Algorithm: The Effect of Algorithmic Management on Worker Performance," *Available at SSRN*, 10.2139/ssrn.4776355. 6

Naritomi, Joana (2019) "Consumers as tax auditors," *American Economic Review*, 109 (9), 3031–72. 7, OA20

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan (2019) "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366 (6464), 447–453. 6

OECD (2023) *Tax Administration 2023: Comparative Information on OECD and other Advanced and Emerging Economies*: OECD Publishing, Paris, 177, 10.1787/900b6382-en. 2

Okunogbe, Oyebola and Fabrizio Santoro (2023) "The Promise and Limitations of Information Technology for Tax Mobilization," *The World Bank Research Observer*, 38 (2), 295–324, 10.1093/wbro/lkac008. 7

Okunogbe, Oyebola and Gabriel Tourek (2024) "How Can Lower-Income Countries Collect More Taxes? The Role of Technology, Tax Agents, and Politics," *Journal of Economic Perspectives*, 38 (1), 81–106, 10.1257/jep.38.1.81. 7

Pomeranz, Dina (2015) "No Taxation Without Information: Deterrence and Self-Enforcement in the Value Added Tax," *American Economic Review*, 105 (8). 7, OA20

Rasul, Imran and Daniel Rogger (2018) "Management of bureaucrats and public service delivery: Evidence from the nigerian civil service," *The Economic Journal*, 128 (608), 413–446. 7

Sequeira, Sandra (2012) "Advances in Measuring Corruption in the Field," in *New Advances in Experimental Research on Corruption*, 15, 145–175: Emerald Group Publishing Limited, 10.1108/S0193-2306(2012)0000015008. 22

Stevenson, Megan T and Jennifer L Doleac (2024) "Algorithmic risk assessment in the hands of humans," *American Economic Journal: Economic Policy*. 29

Szucs, Ferenc (2024) "Discretion and Favoritism in Public Procurement," *Journal of the European Economic Association*, 22 (1), 117–160, 10.1093/jeea/jvad017. 7

World Bank (2024) "Public Finance Review Tool," knowledge tool. 7

Tables

Table 3: Number of Firms by Data Source (Digitized Data)

		2014	2015	2016	2017	2018	2019	2020
A: Self reported	CIT	5121	5954	6224	6575	6701	7210	6817
	VAT	11179	11905	12703	13355	13960	14190	13530
	PAYE	7055	7498	7870	8493	8757	8982	8619
	CGU	1589	1829	2033	2203	2649	2672	2799
	TAF	86	105	112	125	121	114	117
B: Third party	Imports	8965	12404	13070	11850	13556	13676	10587
	Exports	1409	1727	1872	1830	1687	1660	1563
	Procurement	816	730	1387	1343	1904	1893	1690
	VAT annexes	6	9	21	804	3582	3199	NA
C: Audits data	Digitized	NA	NA	1	3291	2756	2946	3717
	Self-reported (Excel)	NA	NA	NA	102	559	674	57

Notes: This table shows the number of unique taxpayers (firms) by year of available data in digital format, in the main datasets used to construct the risk-score algorithm and analyze its performance. The available data covers the years 2014 to 2020, and the experiment was conducted in the years 2018, 2019, and 2020. The self-reported data (Panel A) include the self-assessment declarations for the Corporate Income Tax (CIT), Value-Added Tax (VAT), Pay-As-You-Earn withheld personal income tax (PAYE), the simplified Senegalese tax for small enterprises called CGU, and the Senegalese tax for financial services enterprises called TAF. The third-party datasets (Panel B) include transactions data from customs (imports and exports), procurement records, and VAT annexes. The audit data (Panel C) come from a) an effort to digitize audit reports from 2017 onwards, covering all audits conducted under our intervention, and b) an Excel spreadsheet which the research team requested tax inspectors to fill out to report audit outcomes and audit process information. This table is discussed in Section 2.2.

Table 4: Number of Audit Cases by Selection Method and Tax Office (2018-2020)

	Full Audits			Desk Audits			
	Inspectors	Algorithm	Total	Inspectors	Algorithm	Random	Total
Large Taxpayer Unit	265	131	396	320	157	60	537
Medium Taxpayer Unit	155	153	308	320	290	135	745
Liberal Professions	46	50	96	235	222	85	542
SME (Regional)	71	73	144	422	401	84	907
Total	537	407	944	1297	1070	364	2731

Notes: Number of selected cases by tax office, audit type, and selection method. 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 4.2. A more detailed count by year and regional tax office is provided in Table B.3.

Table 5: Firm Characteristics of Algorithm vs Discretionary Selection

A: Characteristics of Inspector-Selected Cases Relative to the Full Population of Firms

Data Source:	Tax Declarations				Admin. Audit Data			Taxpayer Survey	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log(Turnover)	log(Payroll)	Profit Rate	P(Trade)	Duration Trip	Firm Age	Employees	% Sales in Cash	Audit Frequency
Inspectors' Selection	3.25***	1.37***	0.01	0.18***	0.01	0.91*	-5.44	-7.41	-0.01
	(0.83)	(0.28)	(0.01)	(0.02)	(0.04)	(0.40)	(6.36)	(5.36)	(0.18)
N	22571	7420	5914	61513	60884	52035	697	703	640
R2	0.23	0.13	0.05	0.02	0.06	0.10	0.09	0.02	0.04
Mean outcome	10.81	14.82	-0.15	0.41	1.94	5.91	31.23	50.28	2.77

B: Comparison of Inspector versus Algorithm Selected Cases

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Algorithm	-1.68***	-1.54***	-0.06**	0.06*	0.14	0.18***	16.13	12.63**	-0.05
	(0.42)	(0.41)	(0.03)	(0.03)	(0.09)	(0.06)	(18.25)	(4.96)	(0.25)
Inspectors x Overlap	0.15	1.23**	0.01	0.09	0.35	0.10	-95.42	10.81	0.37
	(1.11)	(0.48)	(0.04)	(0.08)	(0.23)	(0.19)	(59.47)	(22.25)	(0.77)
N	1182	1016	899	1318	1296	1318	266	275	265
R2	0.12	0.12	0.09	0.14	0.05	0.06	0.28	0.13	0.12
Mean outcome	18.37	16.48	-0.02	0.51	2.04	2.13	44.71	45.50	2.80

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows OLS regression coefficients comparing the characteristics of inspector-selected and algorithm-selected firms for full audits (Table B.4 shows the same exercise for desk audits). Panel A compares the characteristics of firms selected for full audits by inspectors, to the general population of firms, pulling the 2018-2020 audit selection years. Panel B compares the characteristics of firms selected by the algorithm relative to those that were selected by the inspectors (all selected full audit cases 2018-2020). In Panel A, columns 1-6, the regressions show the difference in the firms' characteristics using a dummy variable that indicates that the firm was selected for full audits by inspectors, at some point during the years 2018-2020. The sample is the entire set of firms with tax declarations in Senegal. The regressions control for fixed effects at the tax office level. Robust standard errors are shown in parentheses (Huber-White formula). Data on 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 has exports or imports. We use the value for the year before the firm was selected for audit. We use data from the firm registry on the firm's age and the distance between its location and DGID, in minutes of travel time (computed using GoogleMaps for a Monday at 3 PM). 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 Sections 4.4 and mentioned in Section 6.4.

Table 6: Algorithm Selection and Audit 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	-0.17***	-0.05**	-0.04**	0.02	-0.04	-0.04	-0.65***	-0.24*	-0.12
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.18)	(0.13)	(0.14)
				[-0.05 **,0.04*]		[-0.07***,-0.02]	[-0.77***,-0.52***]		[-0.31***,0.01]
Inspectors x Overlap	0.16**	0.05	0.05	0.08	-0.05	-0.04	0.29	-0.54*	-0.37
	(0.07)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.46)	(0.31)	(0.37)
Algorithm x Random		-0.00	-0.00		-0.02	-0.02		-0.25	-0.25
		(0.03)	(0.03)		(0.04)	(0.04)		(0.19)	(0.19)
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	508	1011	992	457	747	728
R2	0.26	0.23	0.32	0.11	0.33	0.43	0.32	0.29	0.42
Mean outcome	0.53	0.37	0.37	0.89	0.73	0.73	19.26	17.76	17.73

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of a regression of the main audit outcomes on the selection method of the case, for full and desk audits separately. Specifications for full audits control for tax office x year fixed effects, which corresponds to the selection list for full audits. For desk audits, we show specifications with tax office x year fixed effects, and inspector x year fixed effects, the latter corresponding to the list for desk audits. The first outcome (Columns 1 to 3) is the probability that the case on the list was actually executed by the inspectors. The second outcome (Columns 4 to 6) is the probability that the audit resulted in a penalty (initial or final notice), conditional on the case being conducted. The third outcome (Columns 7 to 9) is the log amount of the evasion plus penalties uncovered, conditional on the case ending in positive penalty. The evasion amount is obtained from the final notice, which we complement with the initial notice when the final notice is not available. The sample includes all cases selected in the 2018, 2019, and 2020 audit programs. 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. Robust standard errors (Huber-White) are shown in parentheses, and are clustered at the inspector x year level when inspector fixed effects are included. Lee bounds are shown in brackets for the coefficient on the algorithm-selection dummy, computed based on the attrition between algorithm and inspector cases for variables after audit execution. Table C.1 shows intent-to-treat effects, Table C.2 shows the results when using the evasion rate (instead of the evasion amount) as the outcome. Figure C.1 shows the distribution of audit list (=inspector×year) fixed effects for columns 3, 6 and 9. This table is discussed in Sections 5.1 and 5.2 and mentioned in Sections 5.3 and 7.4.1.

Table 7: Resources Allocated to Audits and Audit Productivity, Full Audits

	A1: Resource Outcomes				**B1: Productivity Outcomes**			
	Number of Agents	Duration in Days (Taxpayer Survey)	Days from Start to Conf. (Admin. Data)	Days Working on Case (Self-Reported)	Evasion/ Number of Agents	Evasion/Duration (Taxpayer Survey)	Evasions/Duration (Admin. Data)	Evasion/Days Working (Self-Reported)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Algorithm	-0.26***	-8.54**	-30.18**	-27.26	-0.55***	-0.12	-0.50*	-0.74
	(0.08)	(3.72)	(13.75)	(30.60)	(0.18)	(0.39)	(0.26)	(0.49)
Inspectors x Overlap	0.04	-25.57	-9.83		0.30	0.96	-0.01	
	(0.17)	(21.89)	(27.66)		(0.44)	(2.20)	(0.57)	
N	508	214	282	51	457	98	253	40
R2	0.21	0.25	0.30	0.25	0.31	0.24	0.25	0.40
Mean outcome	2.87	28.06	160.07	177.07	18.19	15.90	15.00	13.62
	A2: Data Availability for Resource Outcomes				**B2: Data Availability for Productivity Outcomes**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Algorithm		-0.01	0.01	-0.05**	0.02	-0.00	-0.08**	-0.04**
		(0.04)	(0.01)	(0.02)	(0.03)	(0.04)	(0.04)	(0.02)
Inspectors x Overlap		-0.03	-0.02	-0.01	0.08	-0.01	0.07	0.00
		(0.06)	(0.04)	(0.04)	(0.05)	(0.06)	(0.07)	(0.03)
N		508	508	508	508	508	508	508
R2		0.22	0.96	0.47	0.11	0.22	0.49	0.41
Mean outcome	1.00	0.21	0.55	0.10	0.89	0.20	0.50	0.08

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results of regressions of outcomes on the selection method of the case and tax office $\times$ year fixed effects. The sample includes all full audit cases selected and conducted under the 2018, 2019, and 2020 audit programs, for which we had data on the outcome variables. Panel A shows the results for the resource and productivity outcomes, and Panel B shows the results of regressions where data availability dummies are the outcomes. Column 1 shows the results for the number of agents involved in the audit, which is obtained from the audit records. Column 2 shows the results for audit duration as reported by taxpayers in the taxpayer survey. The survey asked them about their last full audit experience. Column 3 uses the difference between the date of initial notice and the date of the start of the audit (either verification announcement or information request or start date of audit) as a proxy for the duration of the audit. This measure is available only for cases that had an initial notice. We drop negative values, trim the data at the 99th percentile, and control for dummies indicating availability of date of initial notice and final notice. These controls are helpful because the definition of "start date" may vary depending on the availability of information. Column 4 measures the duration of the audit using the self-reported number of days inspectors worked on a case, reported on a spreadsheet that inspectors needed to fill out during the 2019 audit program. Columns 5-8 use as outcomes the log of the ratio of evasion over the different cost measures used in columns 1-4. We multiplied the answer by the number of inspectors that worked in the case. Robust standard errors (Huber-White formula) are shown in parentheses. Table C.4 shows that same analysis for desk audits. This table is discussed in Sections 6.1 and 6.2.

Table 8: Algorithm Selection, Perceived Audit Efficiency and Corruption

A: Surveyed Firms (Self-Reporting a Recent Audit)

Outcome:	Efficiency Index			Corruption Index		
	Full Audits	Desk Audits	All Audits	Full Audits	Desk Audits	All Audits
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-0.41***	0.01	-0.17	0.05	-0.05	-0.02
	(0.15)	(0.14)	(0.10)	(0.15)	(0.14)	(0.10)
Inspectors x Overlap	0.45**	0.04	0.16	-0.56*	-0.12	-0.18
	(0.18)	(0.26)	(0.21)	(0.29)	(0.27)	(0.20)
Algorithm x Random		-0.09	-0.05		-0.08	-0.04
		(0.19)	(0.18)		(0.18)	(0.16)
N	199	273	472	199	272	471
R2	0.10	0.02	0.03	0.11	0.10	0.05
Mean outcome	0.00	0.00	0.00	-0.07	0.05	0.00

B: Only Audited Firms (as per Administrative Audit Data)

	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-0.36	0.37	-0.02	0.29	-0.24	0.02
	(0.26)	(0.23)	(0.17)	(0.24)	(0.22)	(0.17)
Inspectors x Overlap	0.19	0.40	-0.07	-0.33	-0.13	0.09
	(0.24)	(0.35)	(0.29)	(0.48)	(0.31)	(0.29)
Algorithm x Random		-0.01	0.08		-0.03	-0.04
		(0.27)	(0.24)		(0.26)	(0.23)
N	89	120	211	88	122	212
R2	0.10	0.08	0.03	0.16	0.13	0.05
Mean outcome	0.00	-0.06	-0.02	-0.14	-0.03	-0.08

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. OLS results with fixed effects at the year and tax office level for outcomes from the taxpayer survey. Robust standard errors are in parenthesis. The table shows the association between audit selection method and taxpayers' perception of the efficiency of audits (columns 1-3), and of the incidence of corruption during audits (columns 4-6). The sample of surveyed taxpayers is a subsample of the audit cases selected by the yearly audit programs between 2018 and 2020. Panel A shows the results for all firms interviewed in the survey, conditional on them self-reporting an audit in recent years. Panel B is restricted to firms for which we can confirm that an audit part of the yearly audit program took place, according to the administrative audit reports. The two outcomes are (1) an index of perceived efficiency of audits, and (2) an index of perceived corruption of audits. Each measure is constructed by combining three questions into a normalized index following Anderson (2008). To measure efficiency, we combine the questions on taxpayers' assessment of the auditor's (i) technical knowledge, (ii) efficiency during the audit, (iii) capacity to uncover all evaded taxes. For corruption, we combine the questions on the perception of (i) the auditor's dishonesty during the audit, (ii) the frequency of bribes paid to inspectors (among firms similar to the respondent's), (iii) the existence of a preferential tax audit treatment for firms connected with the tax administration. The coefficients are measured in terms of standard deviation of the respective indexes. Evidence that respondents' likelihood of answering the questions is not associated with the audit selection method is available upon request and part of the previous working paper version. This table and the survey questions the indices are based on are discussed in Section 6.3.

Table 9: Risk Score and Audit Outcomes

A: Baseline Risk Score	P(Execution)		P(Detection|Execution)		log(Evasion)|Detection	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.29***	-0.07***	0.01	-0.08*	-1.09***	-0.26
	(0.05)	(0.03)	(0.04)	(0.04)	(0.28)	(0.21)
Risk score	0.05**	-0.00	0.00	0.01	0.36***	0.38***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.13)
Alg. x Risk score	-0.02	0.02	0.00	0.01	-0.29*	-0.43***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.15)	(0.13)
N	944	2731	508	992	457	728
R2	0.27	0.32	0.11	0.43	0.33	0.43
Mean outcome	0.53	0.37	0.89	0.73	19.26	17.73
B: Controlling for Firm Size						
Algorithm	-0.24***	-0.07**	0.03	-0.08*	-1.04***	-0.26
	(0.05)	(0.03)	(0.04)	(0.04)	(0.28)	(0.21)
Risk score	0.02	-0.02	-0.01	0.01	0.34***	0.38***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.13)
Alg. x Risk score	0.01	0.04**	0.01	0.02	-0.28*	-0.43***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.15)	(0.13)
Lagged log(turnover)	0.02***	0.01***	0.01**	0.01***	0.02	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)
N	944	2731	508	992	457	728
R2	0.33	0.33	0.13	0.44	0.34	0.43
Mean outcome	0.53	0.37	0.89	0.73	19.26	17.73
C: Unweighted Risk Score						
Algorithm	-0.25***	-0.08***	0.10*	-0.07*	-0.74**	-0.26
	(0.06)	(0.03)	(0.05)	(0.04)	(0.33)	(0.23)
Unw. Risk score	0.06**	0.00	0.01	0.03	0.31*	0.07
	(0.03)	(0.02)	(0.02)	(0.02)	(0.17)	(0.17)
Alg. x Unw. Risk score	-0.04	0.04	-0.05	-0.00	-0.34	0.03
	(0.04)	(0.02)	(0.03)	(0.03)	(0.22)	(0.19)
N	944	2731	508	992	457	728
R2	0.27	0.32	0.12	0.43	0.33	0.42
Mean outcome	0.53	0.37	0.89	0.73	19.26	17.73

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. Panel A shows the correlation of the algorithm risk score with the three main audit outcomes: the probability of starting the selected audit, the probability of detecting evasion among started cases, and the log of detected evasion amount plus fines. The risk score is a continuous variable computed within "clusters" of firms of similar economic activity and size. We standardize the risk score values at the cluster-year level, i.e., we subtract the mean and divide by the cluster's standard deviation. We do that because the risk score scale changed in the three iterations of the experiment. The coefficients show how the risk score correlates with the outcomes for discretionary cases (non-interacted term) and for algorithm cases (interacted term). The results stem from OLS regressions with fixed effects at the list level (year X tax office for full audits, and year x inspector for desk audits). Robust standard errors (Huber-White) are shown in parentheses. Panel B is identical to Panel A, except that we add lagged log turnover as an additional regressor. Panel C is identical to Panel A, except that we use the raw risk score, prior to reweighting by turnover, as a regressor. Table E.4 shows an additional robustness test for these results. This table is discussed in Section 7.4.1.

Table 10: Treatment Effect of Information Intervention

	(1) P(Execution)	(2) P(Detection| Execution)	(3) log(Evasion)| Detection	(4) P(Start)	(5) P(Detection| Execution)	(6) log(Evasion)| Detection
Algorithm	-0.05	-0.08*	-0.19	-0.05	-0.08*	-0.19
	(0.04)	(0.05)	(0.22)	(0.04)	(0.05)	(0.22)
Algorithm x Random	-0.00	-0.01	-0.29	-0.00	-0.01	-0.29
	(0.03)	(0.04)	(0.19)	(0.03)	(0.04)	(0.19)
Information	0.02	-0.04	-0.24			
	(0.04)	(0.04)	(0.20)			
Algorithm x Information	0.01	0.07	0.28			
	(0.04)	(0.06)	(0.26)			
Info. (indicators)				0.03	-0.05	-0.09
				(0.04)	(0.05)	(0.23)
Info. (indicators+data)				0.01	-0.04	-0.37
				(0.04)	(0.05)	(0.24)
Alg. x Info. (Indicators)				0.02	0.07	0.04
				(0.05)	(0.07)	(0.31)
Alg. x Info. (Indicators+data)				0.00	0.07	0.49
				(0.05)	(0.06)	(0.31)
N	2136	892	628	2136	892	628
R2	0.32	0.41	0.42	0.32	0.41	0.43
Mean outcome	0.42	0.70	17.68	0.42	0.70	17.68

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows estimates for the effect of the randomized information treatment on audit execution and subsequent audit outcomes, for desk audits. The treatment was cross-randomized across algorithm and inspector cases in 2018 and 2019, but only used for algorithm cases in 2020. Inspector-selected cases for 2020 are hence excluded. Columns 1-3 show the results for a specification containing a dummy indicating whether the case was treated. Columns 4-6 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. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 7.4.2.

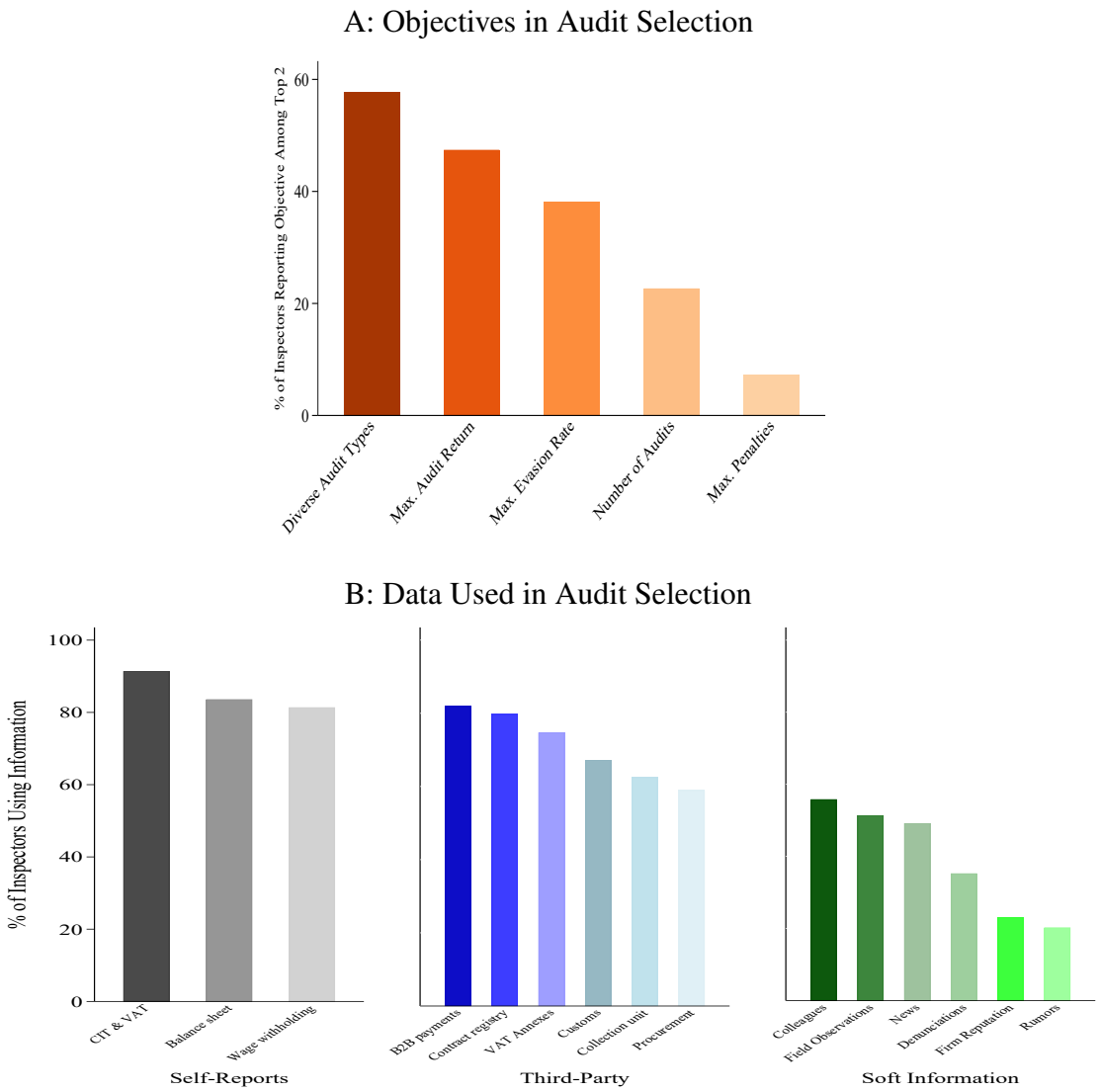
Table 11: Optimization Gains from Machine-Learning Algorithm Trained on Outcome Data

Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Overlap Between Optimized and Realized Audit Program
(1)	(2)	(3)	(4)
A: Full Audits			
21.24	21.14	+ 18.09%	65%
B: Desk Audits			
19.00	18.73	+ 31.01%	61%

Notes: This table shows the results from our optimization exercise in which we use the audit results from the realized set of audits to train a machine-learning (random forest) algorithm to predict evasion. We optimize which audits were conducted in 2018 and 2019. Column 1 shows the realized revenue (evasion and fines) from audits across all tax offices. Column 2 shows the predicted revenue, based on the random forest model, for firms that experienced an audit. Column 3 shows the revenue increase (in percent, compared to the predicted revenue in column 2) when using the random forest for audit selection. Concretely, we use our estimated random forest to predict evasion for all cases in the annual audit program list, then rank cases by predicted evasion within tax office and year, and pick the top cases such that the number of audits by tax office and year is the same as the realized number of audits. Column 4 shows the share of cases audited in the realized program that were also chosen to be audited under the optimized program. This table is discussed in Section 8. Table F.3 shows the same results disaggregated by tax office.

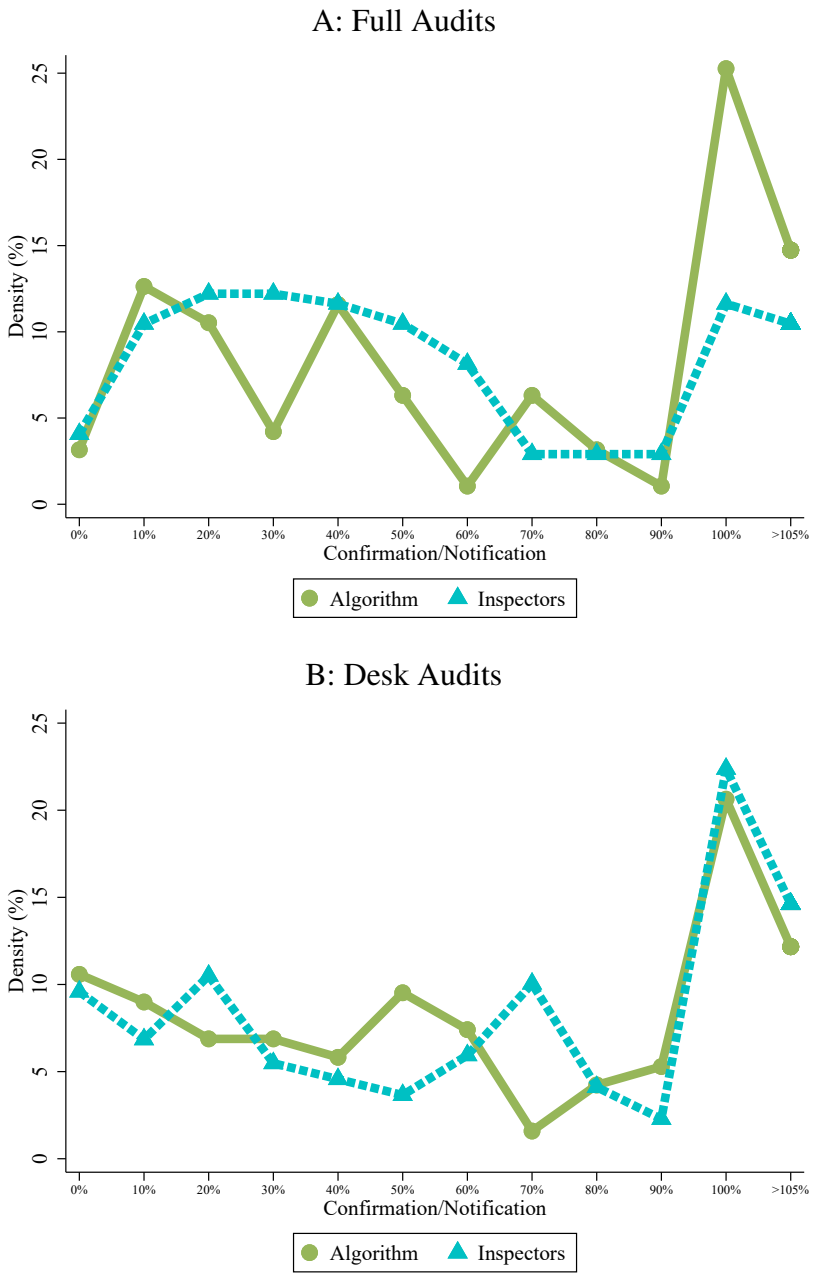
Figures

Figure 1: Tax Inspector Objectives and Practices in Audit Selection



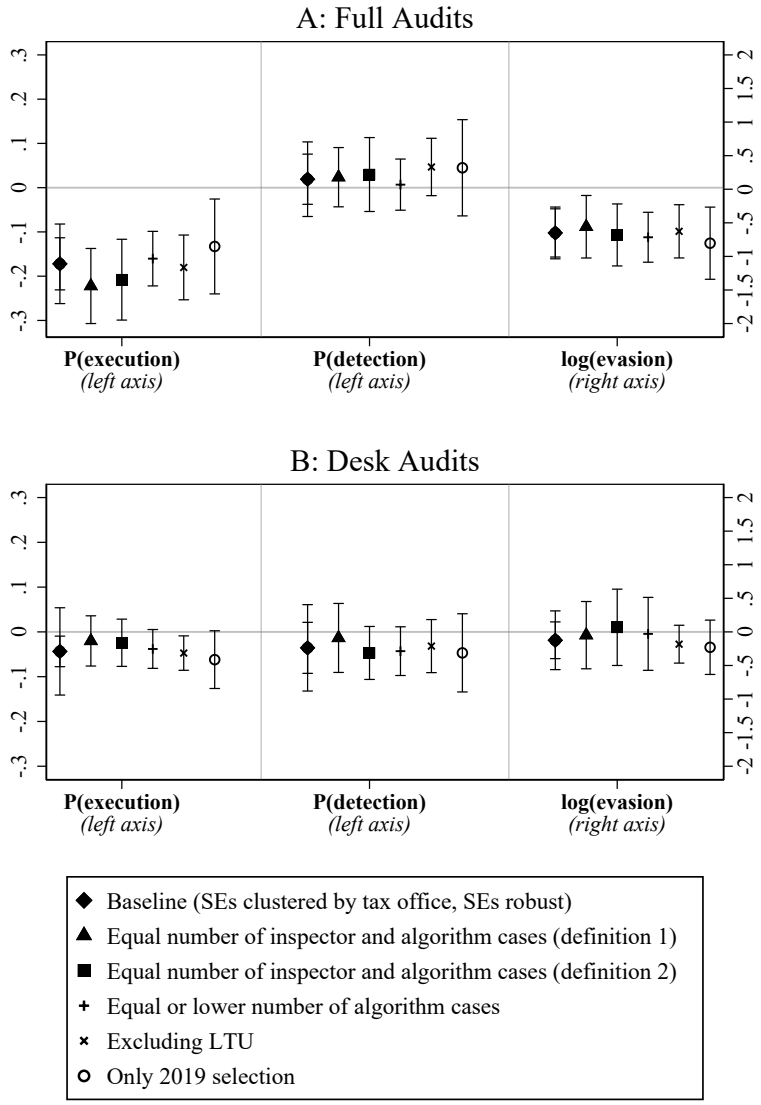
Notes: This figure shows inspectors' objectives and the information they use for audit selection. The data is collected at baseline (December 2017-January 2018) from our survey of 97 tax inspectors actively involved in conducting audits. Panel A shows the top two reported personal objectives of audits, from the following choice set: diversity of audit types, maximize audit returns, maximize the evasion rate, conduct the largest number of audits, and maximize the assessed penalties. Results are similar when reporting the top objective. Panel B shows the frequency of using different data types for audit selection, classifying data into three categories: firms' self-reports, third-party reported data, and soft information which is hard to codify. This figure is discussed in Section 2.3.

Figure 2: Dispute of Audit Results
Distribution of Confirmed Amount/Notified Amount of Evasion



Notes: This figure shows the distribution of the confirmed amount over the notified amount of evasion, for each selection method and audit type. The share is calculated for audit cases where both a non-zero notified amount and a non-missing confirmed amount are reported. The share of confirmation to notification reflects both the extent to which taxpayers bargain to lower the detected evasion and fines due and the quality of the initial assessment on taxes due by the inspectors. This figure is discussed in Section 5.3.

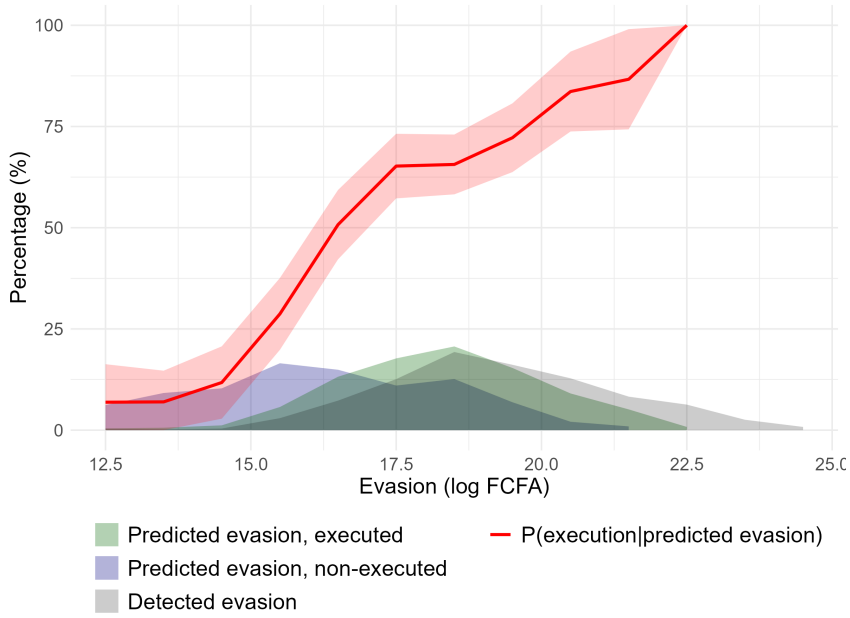
Figure 3: Algorithm Selection and Audit Outcomes: Robustness of Estimates Across Samples



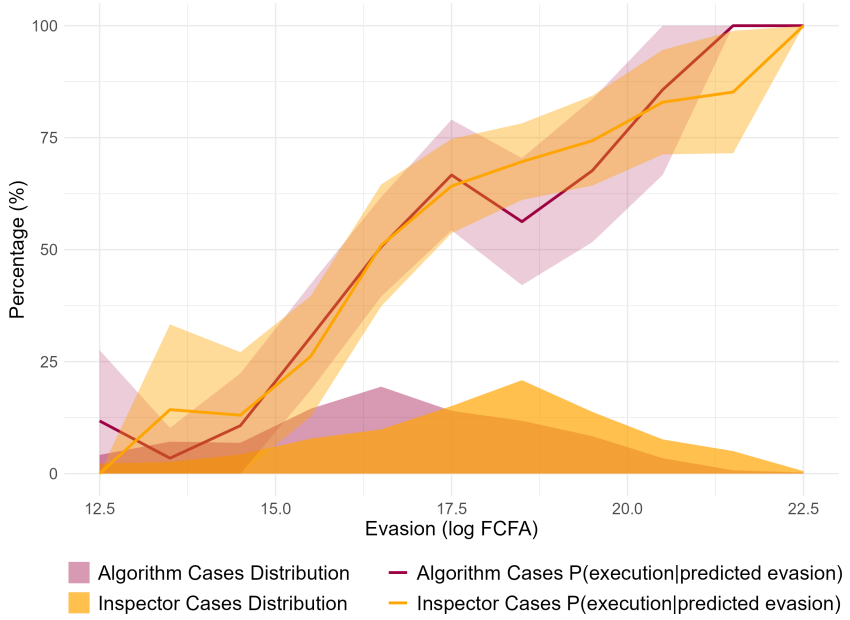
Notes: This figure shows the robustness of our results on the probability of execution, detection and evasion amounts uncovered by the algorithm selection relative to the inspector selection, in different subsamples. Panel A shows results for full audits, and Panel B shows results for desk audits. In each sub-panel, the first coefficient reproduces our main result from Table 6, which we call the baseline. We show robust standard errors, as in our main analysis, and standard errors clustered at the tax office $\times$ year level. The remaining coefficients use the same empirical specifications as Table 6, employing tax office $\times$ year fixed effects for full audits and inspector $\times$ year fixed-effects for the desk audits. We limited the sample in the following way: 2) lists that feature an equal number of inspector-selected and algorithm-selected cases (as observed by the researchers), 3) lists that feature an 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 only for the year 2019 (which we consider as the cleanest execution year for the program). For the distinction between methods 1 and 2 for equalizing the numbers of algorithm and inspector-selected cases, see footnote 19. The detailed table outputs underlying these robustness estimates are available upon request. This figure is discussed in Section 5.4.

Figure 4: Predicted Evasion and Audit Execution, Full Audits

A: Predicted Evasion Amount and Realized Execution Rate, All Cases



B: Realized Execution Rate for Algorithm Cases and Inspector Cases



Notes: Panel A plots the distribution of detected evasion for all full audits (log of FCFA amounts, grey), as well as the results from a random forest model predicting detected evasion based on firms' observable characteristics. The blue distribution is predicted evasion for cases that were part of the annual audit program but not actually audited. The green distribution is predicted evasion for cases that were audited. The red solid line plots the realized audit execution rate by level of predicted evasion. Panel B shows the distribution of predicted evasion for all algorithm-selected cases (pink) and for all inspector-selected cases (orange). The solid lines plot the realized audit execution rates by level of predicted evasion. This figure is for full audits only. This figure is discussed in Section 7.2. Figure E.1 replicates panel A for inspector-selected and algorithm-selected audits separately (panels A and B), and replicates both panels of this figure for desk audits (panels C and D).

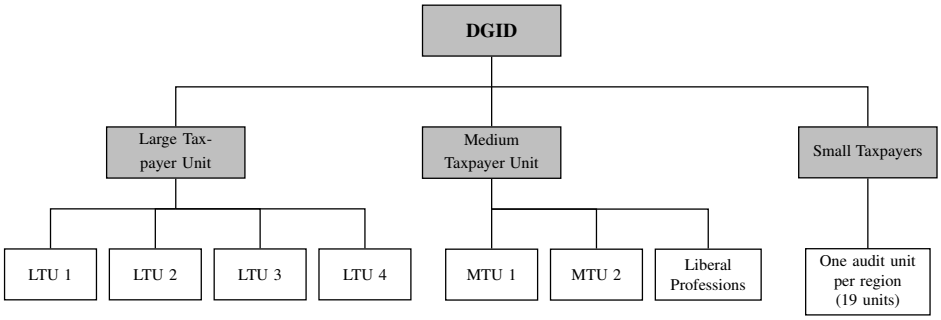
Online Appendix: Not for Publication

This appendix contains additional information and analyses. Appendix A provides further information on the context of our study. Appendix B presents the risk-score algorithm and study design. Appendix C presents additional results on audit outcomes and the audit process. Appendix D examines deterrence effects of audits. Appendix E presents additional results on mechanisms. Appendix F presents additional results on the machine-learning optimization exercise and reviews the relevant literature.

A Context Appendix

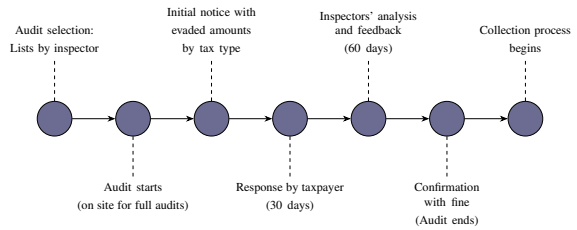
A.1 Tax Policy and Enforcement in Senegal

Figure A.1: Tax Offices Involved in Conducting Audits



Notes: This figure shows the units within the tax administration (DGID) in charge of conducting audits. The Large Taxpayer Unit (LTU) oversees firms with a turnover greater than 3 billion CFA francs (approximately 5.3 million USD). The LTU is divided into four units specialized by economic activity (mining and energy, financial services and telecommunications, real estate, all other sectors). The medium taxpayer unit (MTU) oversees firms with a turnover between 100 million CFA francs and 3 billion CFA francs (MTU1) and those with a turnover between 100 million and 50 million CFA francs (MTU2), as well as regulated liberal professions (lawyers, notaries etc). Other taxpayers are overseen by one of the 19 small taxpayer offices, organized by region. This figure is discussed in Section 2.1.

Figure A.2: Audit Process at the Senegalese Tax Authority



Notes: This figure shows 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 2.1.

A.2 Tax Inspector Performance Incentives in Senegal

Tax inspectors have a strong financial performance incentives, as mentioned in Section 2.3. **Their remuneration consists of a base salary, a share from a "common fund", and an individual bonus. The latter two are proportional to the fines that non-compliant taxpayers pay.** The fines are typically 50% of the recovered amount of evasion.

10% of the audit fines revenue is set aside for the **individual performance bonus** for auditors who worked on the specific case, distributed among the inspectors (6%), the unit manager (2%), and the recovery agents (2%). The remaining 90% of the fines enters the **common fund**, which is distributed among all agents in the relevant tax administration offices. An agent's **share of the common fund** depends on their function and level of seniority in the administration ("statutory points") and on their performance ("performance points"), as evaluated quarterly by their manager.In general, more senior inspectors have stronger performance incentives as they receive a larger share of the common fund.

B Intervention and Study Design

B.1 The risk-score algorithm

Starting in 2017, the research team held consultations with DGID leadership and former tax inspectors to map the compliance risks of firms and to exploit all available data sources to assess this risk. The algorithm's risk score is based on a parametric formula that aggregates individual indicators suggestive of non-compliance. The score relies on two types of risk indicators: discrepancies and anomalies, as explained in Section 4.1. The definition of comparison clusters for the anomaly indicators is based on the administrative division by DGID and economic sectors. Discrepancies and anomalies are aggregated to produce a risk-score for each taxpayer. Table B.1 summarizes the steps taken in calculating the risk-score.

The computation of the risk score changed in each edition of the algorithm. In 2018, we computed the ratios of the indicators and used these ratios as the risk score, adding them up for each firm. In 2019, we computed deciles of the ratios and of the absolute deviations. For example, a firm would have 10 points if it was in the bottom decile of profit rate within its cluster, plus 7 points if the total profits were in the third lowest decile in absolute values. Finally, in 2020, we reduced the number of indicators and computed the risk score by flagging the top decile of each anomaly indicator, and adding a flag for each inconsistency.

The indicators used in the risk-score are summarized in Table B.2, with details here.

Table B.1: Steps of Risk-Score Calculation

Step	Description
(1) Prepare database	Merge tax declarations (VAT, CIT, Payroll) and third-party data (customs, procurement, transaction network) across taxpayers and across years.
(2) Define comparison clusters	Clusters regroup firms in the same economic sector and of comparable size. Peer comparisons of anomaly ratios are done within clusters.
(3) Compute inconsistency ratios	Within-firm indicators denoting inconsistent reporting across declarations. Example: third-party reported sales larger than self-reported sales.
(4) Compute anomaly ratios	Across-firm, indicators suggesting abnormal behavior, based on within-cluster peers' distribution. Example: being at the bottom decile of the cluster's profit rate distribution.
(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 B.2: Indicators Used in the Risk-Score Algorithm

Group	Indicator
Inconsistencies	I1: Treasury payments plus exports exceed declared turnover I2: VAT turnover exceeds CIT turnover I3: Max(payroll+imports, total sales from VAT annexes) exceeds declared turnover I4: Max(treasury paymts+exports, treasury paymts+imports) $>$ FCFA 50 mio, firm $\in$ CGU I5: Inconsistent filing I6: Imports reported in VAT declarations exceed imports at customs I7: Imports from low-tax jurisdictions
VAT anomalies	VAT A1: Turnover divided by net VAT liability VAT A2: Non-taxable domestic turnover divided by domestic turnover VAT A3: Reduced-rate turnover divided by domestic turnover VAT A4: Deductible VAT divided by domestic turnover VAT A5: VAT refund claims divided by deductible VAT VAT A6: Gross VAT liability divided by turnover
CIT anomalies	CIT A1: Turnover divided by net CIT liability CIT A2: Taxable profit divided by net CIT liability CIT A3: Gross CIT liability divided by net CIT liability CIT A4: Deductions divided by turnover CIT A5: Amortization divided by turnover CIT A6: Gross CIT liability divided by turnover
TAF anomalies	TAF A1: High ratio of domestic turnover to TAF liability TAF A2: High ratio of non-taxable turnover to domestic turnover TAF A3: High ratio of turnover taxed at reduced rate to turnover taxed at normal rate TAF A4: High ratio of gross deductions to TAF liability

Notes: This table lists the risk indicators used in the risk score. The table is discussed in Section 4.1.

B.2 Deviations from Pre-Analysis Plan

Our analyses in Sections 5-7 follow the pre-analysis plan, with small exceptions:

- We pre-specified the use of detected evasion and audit costs as outcomes, but not the ratio of the two, i.e. productivity of audits. We still consider audit productivity in Section 6.2, as it is an important outcome from a policy perspective.
- We pre-specified but do not use self-reported difficulty of the audit, challenges encountered, and certain process variables, such as whether inspectors requested additional information or a revised tax return. Inspectors' compliance with the reporting requirement for these variables (submitted in pre-filed excel files) was limited, so the data points are too few and selected to draw meaningful conclusions.
- We do not report heterogeneity, spillovers and learning effects of the information treatment for desk audits, as the treatment had no detectable effect on the main outcomes.
- The analysis in Sections 7.2 and 7.3 was not pre-specified. We think this additional analysis is justified as it does not consider additional outcomes or heterogeneity dimension beyond the pre-analysis plan. Rather, the sections shed more light on the underlying reasons for the main results presented in Sections 5 and 6.

Table B.3: Selected Audits by Office, Year, and Method

		Full audits				Desk audits					
		Algorithm	Discretion	Overlap	Total	Algorithm	Discretion	Random	Overlap	Replacement	Total
DGE	2018	81	94	13	182	72	81	60	12	0	213
	2019	25	96	11	121	0	0	0	0	0	0
	2020	25	75	5	100	85	239	0	76	85	409
CME 1	2018	31	33	2	67	49	52	34	5	0	135
	2019	27	27	1	54	42	52	14	10	7	115
	2020	25	25	0	50	0	0	0	0	0	0
CME 2	2018	25	25	0	53	78	83	49	6	0	210
	2019	20	20	0	40	83	95	38	12	16	232
	2020	25	25	0	50	38	38	0	1	38	114
CPR	2018	15	15	0	32	86	95	59	10	0	240
	2019	15	15	1	30	66	70	26	4	12	174
	2020	20	16	2	36	70	70	0	7	70	210
Dakar P.	2018	0	0	0	0	0	0	0	0	0	0
	2019	14	15	2	29	71	78	37	7	15	201
	2020	7	7	0	14	72	72	0	1	72	216
Ngor A.	2018	0	0	0	0	0	0	0	0	0	0
	2019	11	10	0	21	57	59	19	2	10	145
	2020	8	8	0	16	49	59	0	2	49	157
Pikine G.	2018	0	0	0	0	0	0	0	0	0	0
	2019	8	7	0	15	26	26	14	0	8	74
	2020	8	8	0	16	63	63	0	2	63	189
G. Dakar	2018	0	0	0	0	0	0	0	0	0	0
	2019	9	8	0	17	10	12	14	2	8	44
	2020	8	8	1	16	53	53	0	2	53	159
All	2018	152	167	15	334	285	311	202	33	0	798
	2019	129	198	15	327	355	392	162	37	76	985
	2020	126	172	8	298	430	594	0	91	430	1454
Total		407	537	38	959	1070	1297	364	161	506	3237

Notes: Counts by year and tax office. Method counts can exceed totals because algorithm and discretion selections can overlap; desk audits also include random and replacement selections. This table is discussed in Sections 4.2 and 5.1.

Table B.4: Firm Characteristics of Algorithm vs Discretionary Selection, Desk Audits

A: Characteristics of Inspector-Selected Cases									
Data source:	Tax Declarations				Administrative			Taxpayer Survey	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log(Turnover)	log(Payroll)	Profit Rate	P(Trade)	Duration Trip	Firm Age	Employees	% Sales in Cash	Audit Frequency
Inspectors' Selection	4.68***	0.99***	0.04*	-0.27***	-1.02***	2.37***	0.03	1.62	0.11
	(0.81)	(0.18)	(0.02)	(0.07)	(0.22)	(0.49)	(8.13)	(2.25)	(0.16)
N	22571	7420	5914	61513	60884	52035	697	703	640
R2	0.27	0.13	0.05	0.07	0.14	0.11	0.09	0.02	0.04
Mean outcome	10.81	14.82	-0.15	0.41	1.94	5.91	31.23	50.28	2.77
B: Comparison Among Selected Cases									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Algorithm	1.04***	-0.64**	-0.01	0.08***	0.58***	0.51***	1.00	3.59	-0.23
	(0.28)	(0.30)	(0.02)	(0.02)	(0.05)	(0.04)	(10.75)	(4.27)	(0.22)
Inspectors x Overlap	1.48***	-1.11*	-0.01	0.13***	0.28**	0.30***	5.24	1.59	-0.53
	(0.57)	(0.61)	(0.03)	(0.04)	(0.12)	(0.10)	(8.50)	(17.22)	(0.78)
Algorithm x Random	-0.81**	-0.11	-0.01	-0.10***	-0.29***	-0.25***	-5.21	4.72	0.04
	(0.39)	(0.45)	(0.03)	(0.03)	(0.09)	(0.07)	(7.74)	(5.71)	(0.32)
N	6422	4019	3292	13653	13427	13653	649	650	574
R2	0.29	0.17	0.15	0.26	0.24	0.33	0.30	0.22	0.21
Mean outcome	14.55	14.53	-0.06	0.17	1.43	1.33	22.59	51.15	2.71

Notes: This table is the same as Table 5, but focuses on desk audits. This table is mentioned in Section 4.4.

Table B.5: Balancing Tests

A: Randomization of Case Order on Audit 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	-0.0153	0.0110						
	(0.0208)	(0.0169)						
log(Mean Turnover)			0.00155	-0.000622				
			(0.00167)	(0.00176)				
log(Mean Tax Liability)					0.00133	0.00113		
					(0.00170)	(0.00168)		
Profit rate							-0.0126	-0.0176
							(0.0663)	(0.0614)
N	3034	3034	3034	3034	3034	3034	3034	3034
R2	0.00301	0.00217	0.00305	0.00209	0.00297	0.00219	0.00279	0.00207
Mean outcome	0.35	0.33	0.35	0.33	0.35	0.33	0.35	0.33

B: Information Intervention: Firm Characteristics

	Profit Rate (1)	log (Turnover) (2)	log (Payroll) (3)	log(N. Employees) (4)	log (Exports) (5)	log(Tax Liability) (6)	log(Firm's Age) (7)	log(Distance to Firm) (8)
Indicators	-0.02	-0.07	-0.11	-0.17	-0.25	-0.27	0.05	0.01
	(0.02)	(0.19)	(0.56)	(0.26)	(0.28)	(0.75)	(0.04)	(0.00)
+ data spreadsheets	0.01	0.12	0.86	0.05	0.05	-0.28	-0.07*	0.00
	(0.02)	(0.19)	(0.54)	(0.26)	(0.30)	(0.75)	(0.04)	(0.00)
N	1172	1889	1451	1408	231	1419	1840	1531
R2	0.17	0.14	0.20	0.23	0.39	0.38	0.20	0.14
Mean outcome	-0.06	1.81	13.80	2.84	27.33	4.38	2.44	0.32

C: Information Intervention: Availability of Observations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Indicators	0.01	0.00	0.01	0.01	-0.02	-0.01	-0.03	0.02
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
+ data spreadsheets	-0.00	0.00	-0.02	-0.02	0.01	0.02	0.01	0.00
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
N	2136	2136	2136	2136	2136	2136	2136	2136
R2	0.26	0.12	0.18	0.17	0.27	0.09	0.10	0.15
Mean outcome	0.54	0.88	0.68	0.66	0.12	0.66	0.86	0.72

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. Panel A shows coefficients from OLS regressions predicting whether a case is located in the top or middle third of the inspectors' list, conditional on characteristics. The regressions include fixed effects at the year X inspector level (tax office level for full audits). Panels B and C show results from OLS regressions of firm characteristics and the availability of observations, respectively, on the information treatments. The sample only includes desk audit cases for 2018, 2019, and 2020, which were the years used in the intervention. The treatment was cross-randomized across algorithm and inspector cases in 2018 and 2019, but only used for algorithm cases in 2020. Therefore, we exclude the inspector-selected cases in 2020, which stemmed from a different selection method that is not directly comparable with the algorithm cases. The outcomes are based on tax declarations or firm registry data. For the tax data, we use information from the year before the firm was selected for audit. The outcomes in Panels B and C are defined as follows: the profit rate is defined as total profits divided by total sales, obtained from CIT declarations (Column 1), log(turnover) is the natural logarithm of total sales (plus 1 to avoid dropping observations with zero turnover) obtained from CIT and VIT declarations (Column 2), log(payroll) is the log of total payroll (plus 1) obtained from the Pay-As-You-Earn declarations (Column 3), log(N. employees) is the log of the number of employees obtained from the Pay-As-You-Earn declarations (Column 4), log(exports) is the log of total value of exports (plus 1) obtained from customs data (Column 5), log(tax liability) is the log of the sum of VAT, CIT, and PAYE liability as computed from the tax declarations (Column 6), log(Firm's age) is the log of the age in years of the firm obtained from the firm's date of creation (Column 7), and log(distance to firm) is the log of the distance from the tax office's location to the firm's premises in minutes computed using Google Maps on a Monday afternoon (Column 8). The regressions in Panels B and C include fixed effects at the list level (inspector X year). Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 7.4.2.

C Additional Results on Audit Outcomes and Audit Process

Table C.1: Algorithm Selection and Audit Outcomes, Intent-to-Treat Analysis

A: Linear Regression						
	P(Execution)		P(Detection | Execution)		log(Evasion) | Detection	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-0.17***	-0.04**	-0.14***	-0.04**	-3.54***	-0.81**
	(0.03)	(0.02)	(0.03)	(0.02)	(0.60)	(0.33)
Inspectors x Overlap	0.16**	0.05	0.19***	0.04	3.80***	0.82
	(0.07)	(0.04)	(0.07)	(0.04)	(1.46)	(0.67)
N	944	2731	944	2731	893	2467
R2	0.26	0.32	0.24	0.19	0.27	0.26
Mean outcome	0.53	0.37	0.48	0.27	9.85	5.37
B: Poisson Regression						
Algorithm	-0.32***	-0.12**	-0.30***	-0.17***	-0.37***	-0.15**
	(0.06)	(0.05)	(0.07)	(0.06)	(0.06)	(0.06)
Inspectors x Overlap	0.27***	0.14	0.34***	0.15	0.32***	0.16
	(0.10)	(0.10)	(0.12)	(0.12)	(0.11)	(0.12)
N	925	2504	925	2493	874	2230
Pseudo-R2	0.07	0.11	0.08	0.08	0.18	0.17
Mean outcome	0.54	0.40	0.49	0.29	10.07	5.94

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table is identical to Table 6 discussed in Section 5.1, except it shows intent-to-treat effects, where outcomes are set to zero for audits that are not implemented. Panel A shows ordinary least squares results, and Panel B shows Poisson Pseudo-Maximum Likelihood estimates. Robust standard errors are shown in parentheses (Huber-White formula). The coefficients on the Algorithm x Random interaction are not shown. This table is discussed in Section 5.2.

Table C.2: Algorithm Selection and the Evasion Rate

	% of Liability			% of Pre-Audit Turnover		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Desk audits	Full audits	Desk audits	Desk audits
Algorithm	0.01	-0.03	-0.03	0.02	0.01	0.01
	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
Inspectors x Overlap	0.18**	-0.07	-0.13**	0.10	-0.02	-0.04
	(0.08)	(0.06)	(0.06)	(0.07)	(0.03)	(0.03)
Tax center x Year	Yes	Yes	Yes	Yes	Yes	Yes
Inspector x Year	No	No	Yes	No	No	Yes
N	499	976	956	496	974	953
R2	0.08	0.19	0.33	0.11	0.14	0.30
Mean outcome	0.65	0.50	0.50	0.25	0.21	0.21

Notes: * 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 is similar to Table 6, columns 7-9, but considers the evasion rate as the outcome. The coefficients on the Algorithm x Random interaction are not shown. These results are mentioned in Section 5.2.

Table C.3: Dispute of Audit Outcomes

	P(Confirmation= Notification)		log(Notification)| Both>0		log(Confirmation)| Both Not. and Conf. >0	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	0.12***	-0.00	-0.35	-0.12	-0.23	-0.21
	(0.04)	(0.05)	(0.22)	(0.18)	(0.26)	(0.22)
N	264	371	260	346	260	346
R2	0.17	0.23	0.34	0.54	0.20	0.50
Mean outcome	0.16	0.21	19.43	17.97	18.66	17.35

Notes: * 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). To construct the outcomes we needed data for each case on both the initial notification of the evasion plus penalty amount, and its confirmation. Thus, the sample for the first outcome includes all cases with both non-missing confirmation and notification. For the second and third outcome, the sample is conditional on both notification and confirmation having a non-zero amount. Robust standard errors (Huber-White) are shown in parentheses. This table is discussed in Section 5.3.

Table C.4: Resources Allocated to Audits and Audit Productivity, Desk Audits

A: Resource and Productivity Outcomes						
	Duration in Days (Taxpayer Survey)	Days from Start to Conf. (Admin. Data)	Days Working on Case (Self-Reported)	Evasion/Duration (Taxpayer Survey)	Evasions/Duration (Admin. Data)	Evasion/Days Working (Self-Reported)
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	-1.08	17.94	3.25	-0.42	-0.47	0.32
	(2.12)	(19.11)	(4.62)	(0.48)	(0.34)	(0.44)
Inspectors x Overlap	13.61	22.35	-3.51	0.64	-1.49	0.26
	(12.54)	(40.58)	(3.67)	(0.96)	(1.15)	(1.28)
N	294	247	115	89	167	68
R2	0.28	0.40	0.54	0.43	0.58	0.41
Mean outcome	20.21	110.68	9.88	14.90	13.60	15.87
B: Data Availability						
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	0.01	0.00	-0.03	0.01	-0.02	-0.04*
	(0.03)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)
Inspectors x Overlap	-0.10*	-0.00	-0.01	-0.05	-0.10*	-0.00
	(0.05)	(0.00)	(0.04)	(0.05)	(0.05)	(0.04)
N	992	992	992	992	992	992
R2	0.19	0.98	0.46	0.21	0.29	0.29
Mean outcome	0.16	0.27	0.12	0.11	0.20	0.07

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table is similar to Table 7, but focuses on desk audits. The regressions control for inspector $\times$ year fixed effects. We do not consider the number of agents working on an audit or evasion / number of agents as outcomes, as desk audit lists are selected and executed by individual inspectors or pairs of inspectors. There is hence no variation in the number of inspectors working on a case within a desk audit list. This table is discussed in Sections 6.1 and 6.2.

Table C.5: Inspector Characteristics by Case Selection Method, Full Audits

A: Full-Audit Cases in Which Some Inspector Reported Information

	(1) Mean Age	(2) Max Age	(3) Mean Years of Experience	(4) Max Years of Experience	(5) Share with Masters/PhD	(6) Max Education	(7) Share in Favor of Algorithm
Algorithm	-0.31	-0.95**	-0.04	-0.39	0.04	-0.09*	-0.02
	(0.33)	(0.42)	(0.21)	(0.24)	(0.03)	(0.05)	(0.03)
N	463	463	421	421	421	421	421
R2	0.35	0.39	0.36	0.37	0.31	0.34	0.35
Mean outcome	37.11	40.56	8.15	9.34	0.81	3.23	0.83

B: Only Cases in Which All Inspectors Reported Information

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algorithm	-0.23	-0.65	0.19	-0.09	0.01	-0.02	-0.00
	(0.37)	(0.48)	(0.35)	(0.36)	(0.04)	(0.04)	(0.05)
N	226	226	115	115	115	115	115
R2	0.38	0.31	0.30	0.18	0.59	0.88	0.19
Mean outcome	37.54	42.09	9.24	11.31	0.76	3.35	0.76

C: Probability that All Inspectors Reported Information

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algorithm	-0.03	-0.03	0.01	0.01	0.01	0.01	0.01
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	508	508	508	508	508	508	508
R2	0.27	0.27	0.40	0.40	0.40	0.40	0.40
Mean outcome	0.45	0.45	0.23	0.23	0.23	0.23	0.23

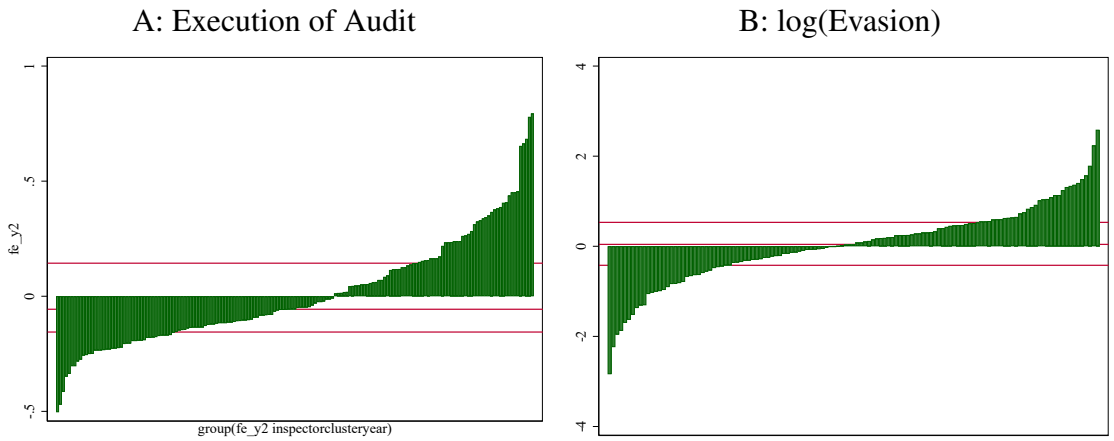
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the OLS estimation results for the mean characteristics of the team members working on cases. The regression is done at the case level for implemented full audits. Tax offices managers assign inspectors (or teams of inspectors) to full audit cases. Since there is no assignment of cases to inspectors at the selection phase, we can only run this regression for the cases that were effectively conducted. The RHS variables are obtained from the inspector survey, and are averaged across the inspectors that worked on the case. Columns 1 and 2 show the results using as outcomes the mean and max age of the team members. Columns 3 and 4 use the share of team members with a Master's/PhD and the max of a categorical education variable that goes from 1 (only high school) to 4 (Masters/PhD). Column 5 uses the mean response of team members to whether they would favor the use of an algorithm to automate selection of cases. Columns 7, 8, and 9 use the share of highly experienced (more than median) team members in the team, the mean experience, and the max experience among team members. The regressions control for fixed effects at the tax office-year level. Robust standard errors are shown in parentheses (Huber-White formula). Coefficients on the inspector $\times$ overlap dummy are included in the regression but omitted from the display. Panel A includes all audit cases for which we have demographic information from at least one inspector on the team. Panel B zooms in on the subsample of cases for which we have information on all inspectors on the team. Panel C considers the probability of having the information as an outcome. This table is discussed in Section 6.1.

Table C.6: Inspector Characteristics and Performance, Desk Audits

	(1) # cases Executed	(2) P(execution)	(3) P(detection|execution)	(4) log(evasion) (Mean)	(5) Executed Alg. cases	(6) P(execution| Alg.)	(7) Executed Alg. > Executed Insp.	(8) P(In favor of Alg.)
Masters/PhD	-0.82	-0.02	0.08	1.23	-0.10	0.00	0.12	0.11
	(0.71)	(0.04)	(0.07)	(1.32)	(0.44)	(0.06)	(0.09)	(0.11)
Above median age	0.50	0.05	0.03	0.34	0.14	0.05	-0.05	-0.08
	(0.52)	(0.04)	(0.04)	(0.78)	(0.35)	(0.06)	(0.09)	(0.12)
Above median experience	-0.37	-0.03	-0.02	-0.68	0.24	-0.00	0.11	-0.16
	(0.44)	(0.04)	(0.04)	(0.92)	(0.30)	(0.05)	(0.09)	(0.10)
N	86	86	76	76	86	86	86	86
R2	0.82	0.53	0.57	0.56	0.74	0.53	0.45	0.09
Mean outcome	4.75	0.32	0.90	16.87	2.52	0.29	0.36	0.77

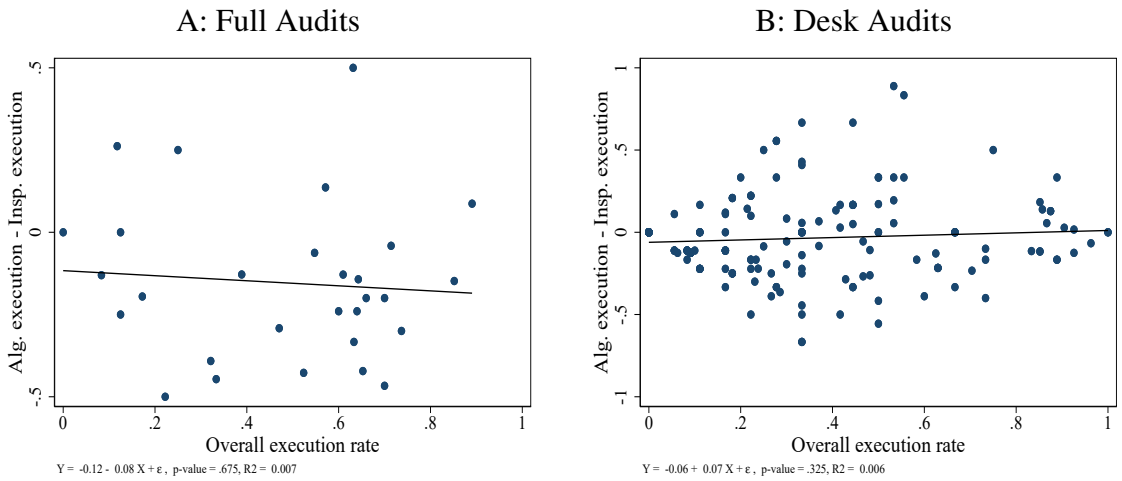
Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the estimation of the OLS regressions of outcomes of inspectors on inspector characteristics. The analysis is done at the list (inspector x year) level and only includes the lists of desk audits. These lists were assigned individually to inspectors. The regressors are three individual characteristics of the inspectors collected using the inspector survey (whether the inspector had a Master's/PhD and whether they had higher than median experience working at the tax authority) and administrative data (age). Column 1's outcome is the total started cases for each list, Columns 2-4 are averages within list of the main outcomes (see Table 6), Column 5 uses the number of started algorithm cases, Column 6 the share of started algorithm cases among algorithm-selected cases, Column 7 uses an indicator of whether the number of started algorithm cases was larger than the number of started inspector cases in the list, and Column 8 uses the response of the inspector to whether they would favor the automation of case selection using an algorithm. The regressions control for fixed effects at the tax office-year level. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 6.1, footnote 20.

Figure C.1: Inspector Quality Fixed Effects for Desk Audits



Notes: This figure shows the distribution of audit list (= inspector × year) fixed effects γ_l from estimating a version of Equation 1 where we additionally control for office fixed effects. This ensures that the list (i.e. inspector) fixed effects pick up variation within rather than across tax offices. The red lines mark p25, p50 and p75 of the distributions. The figure documents substantial heterogeneity in inspector skills, with long tails on either end of the distribution. To gauge the magnitude of the variation, consider that the mean outcome is 0.37 in Panel A and 17.74 log points in Panel B. An improvement in the selection of audit cases through an algorithm would be most valuable for inspectors in the bottom of the skill distribution. This figure is mentioned in Section 6.1, footnote 20.

Figure C.2: Absence of Correlation Between Audit Execution Rate and Execution Rate for Algorithm-Selected Cases



Notes: This figure plots outcomes at the audit-list level (i.e. the tax office × year level for full audits and the tax inspector × year level for desk audits). The X-axis shows the mean execution rate for all cases on the annual list. The Y-axis shows the difference in the execution rate between algorithm-selected cases and inspector-selected cases. This figure is mentioned in Section 6.1, footnote 20.

Table C.7: Algorithm Selection and Scope of Audits

A: Main Scope Outcomes

	N. infractions		N. years		Fine/Evasion	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	-1.24***	-0.46***	-0.32**	-0.24*	0.01	0.03***
	(0.29)	(0.15)	(0.13)	(0.13)	(0.01)	(0.01)
Inspectors x Overlap	0.14	-0.29	0.29	-0.14	0.00	-0.00
	(0.93)	(0.55)	(0.25)	(0.29)	(0.03)	(0.02)
Algorithm x Random		0.33		0.34**		-0.00
		(0.23)		(0.17)		(0.02)
N	457	728	457	728	454	712
R2	0.33	0.33	0.14	0.28	0.21	0.28
Mean outcome	4.91	3.08	3.53	2.65	0.41	0.44

B: Additional Scope Outcomes

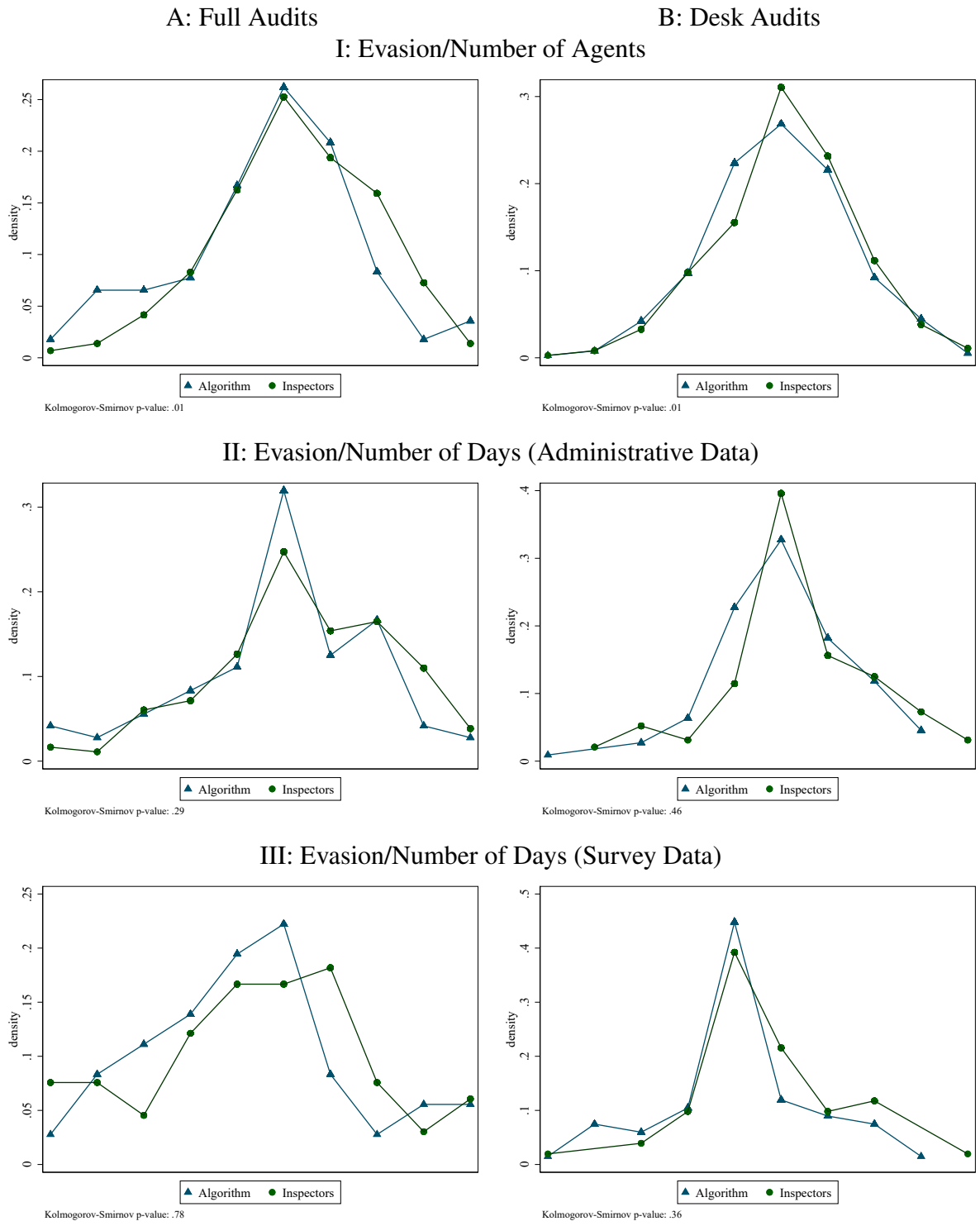
	Share severe infractions (main taxes)		Share severe infractions (all taxes)		P(infraction in main taxes)	
	(1)	(2)	(3)	(4)	(5)	(6)
	Full audits	Desk audits	Full audits	Desk audits	Full audits	Desk audits
Algorithm	1.79	2.25	2.27	4.36	0.01	-0.04
	(3.71)	(3.75)	(3.19)	(3.03)	(0.02)	(0.03)
Inspectors x Overlap	-2.84	4.70	-13.54**	5.62	0.05**	0.04
	(7.81)	(8.13)	(5.38)	(6.54)	(0.03)	(0.06)
Algorithm x Random		4.95		3.82		-0.01
		(5.32)		(4.32)		(0.05)
N	457	728	457	728	457	728
R2	0.18	0.31	0.29	0.37	0.17	0.28
Mean outcome	60.67	64.15	60.71	71.38	0.94	0.86

C: P(severe infraction in main taxes)

	(1)	(2)
	Full audits	Desk audits
Algorithm	-0.02	-0.01
	(0.04)	(0.04)
Inspectors x Overlap	-0.03	0.05
	(0.09)	(0.09)
Algorithm x Random		0.04
		(0.06)
N	457	728
R2	0.11	0.26
Mean outcome	0.78	0.71

Notes: * 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. Panel A reports the three main scope outcomes: 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. Panels B and C provide additional outcomes for the same analysis: the share of severe infractions in main taxes, the share of severe infractions in all taxes, the probability of finding an infraction in main taxes, and the probability of finding a severe infraction in main taxes. Robust standard errors are shown in parentheses (Huber-White formula). This table is discussed in Section 6.1.

Figure C.3: Distribution of Audit Productivity



Notes: This figure plots density distributions of audit productivity outcomes, as per the panel titles. The outcomes are demeaned at the list level. The data is split into ten equally-spaced bins for each audit outcome and type. This figure is discussed in Section 6.2.

D The Deterrence Effects of Tax Audits

- Evasion rates (the share of turnover left unreported) are higher among smaller firms, but turnover, profits, and tax payments are concentrated among larger firms.
- If the (semi-)elasticity of evasion with respect to audit probability is constant across firm sizes, inspector-selected audits generate larger deterrence effects than algorithm-selected audits. If the semi-elasticity declines with firm size, replacing all inspector-selected audits with the risk-score would increase reported turnover by at most 0.2%.
- We estimate an average semi-elasticity of about -2: a 1 percentage point increase in audit probability reduces the evasion rate by 2%. Under this estimate, pure inspector selection would increase reported turnover by 2%, relative to the current mixed-selection system, highlighting an additional advantage of inspector-selected audits.

Objective. Audits generate revenue by uncovering evaded amounts and by inducing tax compliance in the general population. The former effect may induce higher compliance of audited firms in the future ("ex-post compliance"), and the latter may improve the degree of truthful declarations, both among audited and non-audited firms ("ex-ante compliance", see Paradisi and Sartori (2024)). The deterrence effect depends on perceived audit probabilities, which are notoriously difficult to measure (Bergolo et al., 2023). The section examines the following question: how large would deterrence effects have to be for the risk-score selection to improve revenue collection compared to the discretionary selection?

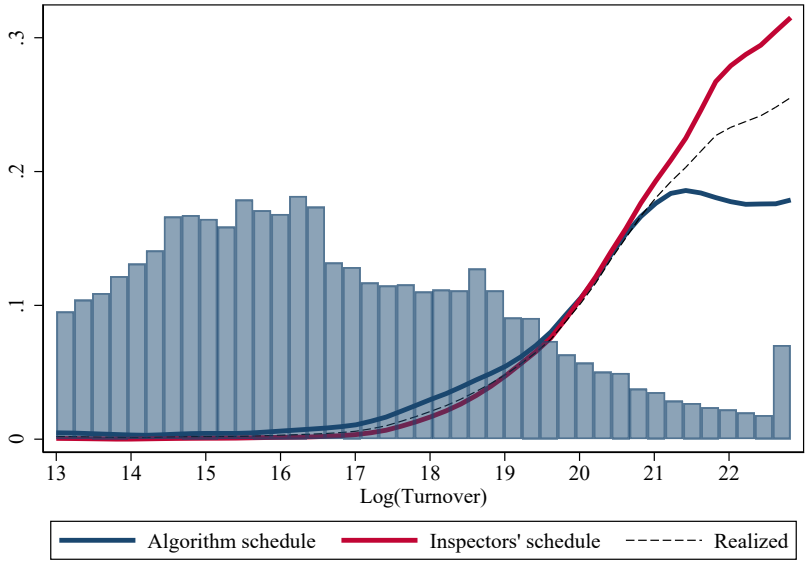
As shown in Table 5, Panel B, risk-score selected firms differ from inspector-selected firms along several dimensions (Section 4.4). We focus on firm size, the most salient difference; risk-score-selected firms are, on average, 1.78 log points smaller in turnover. Figure D.1 plots audit probabilities by firm turnover under the two selection methods. As expected, audit probabilities increase with firm size under both schedules, and are very low for the bottom half of the turnover distribution. The risk score selects medium-sized firms at a higher rate than the inspectors, while inspectors select relatively more firms at the top of the distribution. Consequently, replacing inspector selection with the risk score would raise audit probabilities for upper-mid-sized firms and lower them for the largest firms.

To understand the role of selection methods on deterrence, we simulate firms' reporting decisions under two alternative audit schedules. We assume firms know the audit probability associated with their size, so perceived and actual audit probabilities coincide, and that the evasion rate (i.e., the share of unreported turnover) depends on this probability. Using estimates of the sensitivity of evasion to audit probability, we compute counterfactual aggregate turnover declarations if all audits were selected by the risk score or, alternatively, entirely by inspectors.

To estimate the effect of audit probability on misreporting, we exploit the fact that firms with

similar reported turnover may be arbitrarily assigned to different tax centers—for example, because they are located in different regions—, effectively facing different audit probabilities. This quasi-random variation in audit probability, conditional on firm size, identifies the effect of audit probability on the evasion rate. We then use these estimates to compute the counterfactual.

Figure D.1: Audit Probabilities by Firm Size



Notes: This figure shows the probability of a full audit as a function of turnover using the full sample of firms with tax declarations. The realized probabilities, based on the 2018-2020 audits, are shown in the dashed line. The blue line shows the audit rate for the risk-score selection and the red line shows the audit rate for the inspector selection. The blue bars show the distribution of firms by size, winsorized at the top 5%.

Empirical Analysis. The analysis proceeds in three steps, using the administrative data on tax declarations and audit reports of the main paper. First, we compute the audit probability conditional on firm size and tax office, non-parametrically, using the executed full audits in 2018-2020. Let $P(\text{audit}_i)$ be a dummy variable indicating whether firm i has been audited or not. We estimate the following equation using a kernel and local linear regression:

$$P(\text{audit}_i) = f(\text{turnover}_i, \text{tax office}_i) + \xi_i \tag{D.1}$$

This first step yields a predicted audit probability $\hat{p}_i$ for every firm (audited or not), based on firm size and tax office. $\hat{p}_i$ ranges from 0 to 1.

Second, we estimate the effect of audit probability on the evasion rate. Evasion is defined as evaded taxes divided by 30%, the average corporate income tax rate. The evasion rate is evasion/(evasion + turnover). We assume that firms mainly evade taxes by under-reporting turnover, thereby reducing reported profits proportionally. We regress evasion rates on the audit probabilities to estimate the effect on compliance.

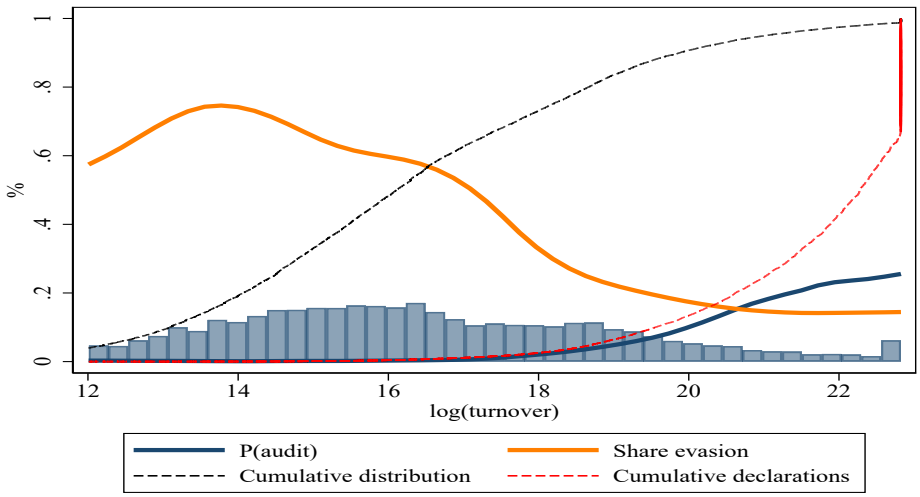
We estimate a two-stage least squares model using tax office dummies as instruments, controlling for firm turnover. Identification comes from similarly sized firms assigned to different tax offices—for example, a firm in the large taxpayer unit may face a higher audit probability than an equally large firm in a medium taxpayer unit. The second stage estimates the semi-elasticity of the log evasion rate with respect to audit probability, and we allow this semi-elasticity to vary with firm size by interacting audit probability with log turnover.

Equation D.2 formally represents this step, using the sample of audited firms in 2018-2020:

$$\begin{aligned} \hat{p}_i &= \alpha_0 + \alpha_1 \log \text{turnover}_i + \text{Office}_i + \epsilon_i \\ \log(\text{evasion rate}_i) &= \beta_0 + \beta_1 \log \text{turnover}_i + \beta_2 \hat{p}_i + \beta_3 \hat{p}_i \times \log \text{turnover}_i + \varepsilon_i \end{aligned} \tag{D.2}$$

Deterrence estimates Figure D.2 plots the evasion rate by firm size. Evasion falls sharply with size: audited firms at the median turnover underreport over 60% of turnover, compared with 25% at the 75th percentile and 17% at the 90th percentile. The dark dashed line shows the cumulative distribution of firms by declared turnover, while the red dashed line shows the cumulative share of declared turnover. More than 30% of total declared turnover is concentrated among the largest 5% of firms.

Figure D.2: Evasion Rates, Audit Probability, and Distributions



Notes: This figure shows estimated evasion rates (orange line) as a function of log(turnover) using audited firms in 2018-2020, full audit probabilities (blue line) as a function of turnover using the full sample of firms with tax declarations, the distribution of firms by size (blue bars), the cumulative distribution of firms by size (black dashed line), and the cumulative declared turnover (red dashed line). The figure is discussed in Section 6.4.

Table D.1 reports estimates of Equation D.2. As expected, the estimated semi-elasticity of evasion is negative, at -2.25 (column 4): a 1 percentage point increase in audit probability reduces the evasion rate by about 2.25%. The positive interaction between audit probability and log turnover (column 2) implies that this semi-elasticity declines with firm size, so larger

firms respond less to increases in audit probability.

To conduct the counterfactuals, we impute evasion rates for non-audited firms based on firm size. We then use the estimated semi-elasticity to compute the change in each firm's reported turnover resulting from changes in audit probability (the deterrence effect). We assume that size-specific evasion rates in the full population are the same as those observed among audited firms, which likely overstates baseline evasion.[36]

Table D.1: Evasion Rate and Audit Probability

	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Audit probability	-24.41***	-37.04**	-5.48***	-2.25*
	(7.31)	(18.80)	(1.52)	(1.30)
Audit probability X Log(Turnover)	1.04***	1.51*		
	(0.39)	(0.81)		
Log(Turnover)	-0.39***	-0.08***	-0.15***	-0.10***
	(0.09)	(0.01)	(0.03)	(0.01)
N	385.00	329.00	385.00	329.00
R2	0.19		0.18	

Notes: * 0.10 ** 0.05 *** 0.01 significance levels. This table shows the results of a regression of the evasion rate on the audit probability, following Equation D.2. Columns 1 and 3 show OLS regressions, while columns 2 and 4 show 2SLS estimations with tax office dummies as instruments for the audit probability.

Counterfactual reported turnover under alternative audit schedules Finally, we compute counterfactual aggregate reported turnover under the risk-score and inspector audit schedules (Figure D.1). Relative to the realized schedule, the risk-score schedule raises audit probabilities for upper-mid-sized firms but lowers them for the largest firms; the opposite holds under the inspector schedule. Figure D.3 shows the resulting changes in total reported turnover for different values of the semi-elasticity of evasion with respect to audit probability. The estimate from our data is marked by the vertical dashed line.

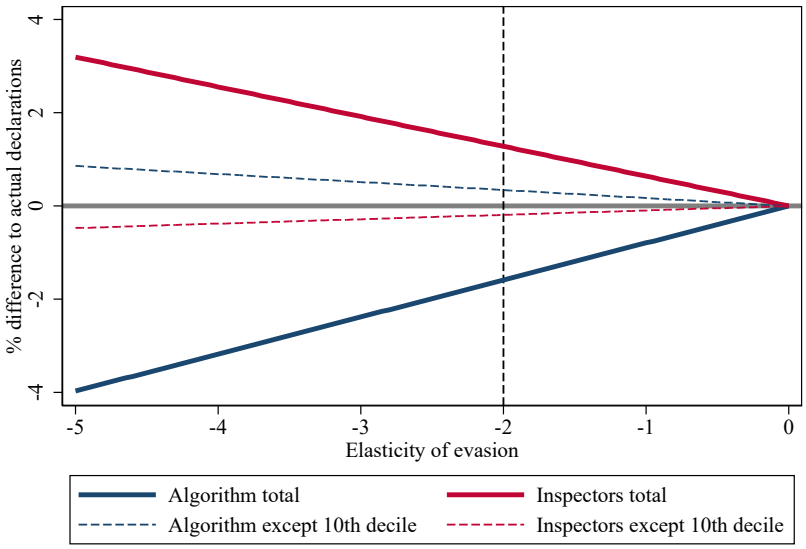
Panel A assumes a constant semi-elasticity (Table D.1, column 4), while Panel B allows it to decline (in absolute value) with firm size (column 2). The solid lines show the change in total reported turnover due to deterrence, relative to the realized audit schedule (a mix of risk-score and inspector selection). The red line corresponds to all executed audits following the inspector schedule, and the blue line to all audits following the risk-score schedule.

For low semi-elasticities, the inspector schedule generates more reported turnover than the realized schedule in both panels, reflecting the strong concentration of turnover among the largest firms. With a constant semi-elasticity (Panel A), this advantage persists for all values. When the semi-elasticity declines with firm size (Panel B), the risk-score schedule performs

[36] As a robustness check, we adjust baseline evasion by firm size using either firms' risk scores (conditional on size) or the machine-learning predictions applied to the full population. While these adjustments affect aggregate evasion, the qualitative results are unchanged. Results are available upon request.

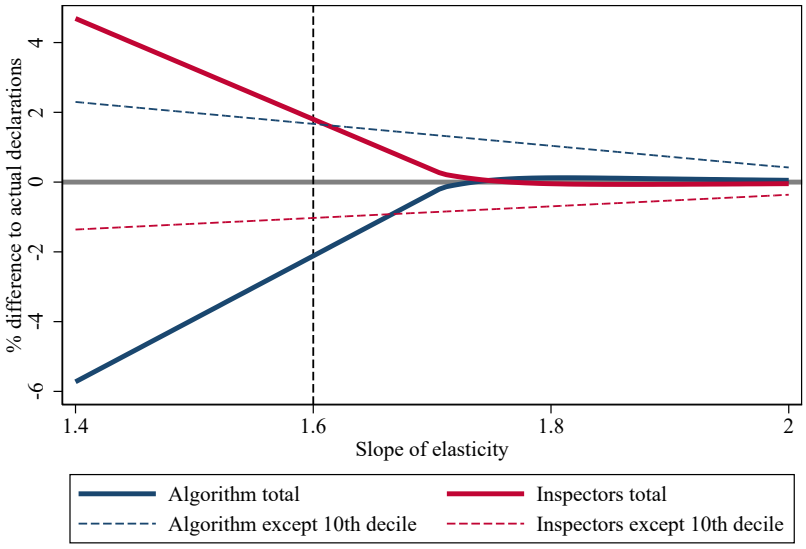
Figure D.3: Counterfactual Turnover Declarations Relative to Realization

A: Constant Semi-Elasticities by Firm Size



Elasticity ε assumed to be constant for all levels of turnover

B: Variable Semi-Elasticities by Firm Size



Elasticity ε assumed to follow ε=-39+slope x log(turnover)

Notes: This figure shows the results of simulation exercises varying the semi-elasticity of evasion rate with respect to percentage point variations in the probability of audit. Evasion rate is defined as the share of underreported turnover in tax declarations. In the figures, we show how aggregate turnover would change relative to the realized aggregate turnover in 2020 in two scenarios: the entire audit program is determined by the inspectors' schedule (red line) or by the algorithm's schedule (blue line). The dashed lines show the relative change in total turnover accumulated up to the 9th decile of turnover. Panel A shows how the total turnover (and accumulated up to the 9th decile) would differ from the realized turnover for different values of the semi-elasticity of evasion rate, assuming that all firms have the same semi-elasticity. Panel B shows the same exercise assuming that the semi-elasticity follows the formula $\varepsilon = -41 + \text{slope} \times \log turnover$ and we vary the slope on the X-axis. In both figures, the vertical dashed lines indicate the estimated values based on Equation D.2. This figure is discussed in Section 6.4.

better once the slope exceeds 1.7. Even then, however, the maximum gain is modest: total reported turnover rises by at most about 0.2% relative to the realized schedule.

The dashed red and blue lines show the deterrence effects for firms below the top turnover decile. For these firms, the risk-score schedule always outperforms both the realized and inspector schedules because it assigns them higher audit probabilities. However, these gains are more than offset by lower reported turnover among the largest firms.

Using our estimated semi-elasticity of -2.25 as a benchmark and applying it uniformly across firms, we estimate that the inspector schedule would increase aggregate reported turnover by 2% relative to the realized schedule through stronger deterrence.

Relation to the Literature What is a plausible range for the (semi-)elasticity of evasion with respect to audit probability? Studies of direct deterrence generally find moderate, significant, and persistent effects of audits on compliance (Advani et al., 2023; Alm, 2019). Summarizing this evidence, Alm (2019) reports income–audit-rate elasticities of 0.1–0.2 and notes that deterrence is small and diminishing at higher audit rates. Work studying changes in communicated audit probabilities provides mixed evidence on indirect deterrence. Bergolo et al. (2023) find no effect of varying audit probabilities in letters, which they attribute to probability neglect. They note that even studies finding positive reporting effects of communicated audit probabilities (Dwenger et al., 2016) imply small deterrence elasticities.[37] Evidence on firm size heterogeneity is mixed: some studies find larger deterrence effects among smaller firms (Pomeranz, 2015; Naritomi, 2019), while others find no heterogeneity (Bergolo et al., 2023; Best et al., 2021).

These results suggest that the deterrence elasticity estimated here lies toward the upper end of the literature, and that the strong size dependence we estimate may not be general. Yet even then, the risk-score selection does not outperform inspector selection in terms of deterrence.

References

Alm, James (2019) "What motivates tax compliance?" *Journal of economic surveys*, 33 (2), 353–388. OA20

D'Agosto, Elena, Marco Manzo, Stefano Pisani, and Filippo Maria D'Arcangelo (2018) "The Effect of Audit Activity on Tax Declaration: Evidence on Small Businesses in Italy," *Public Finance Review*, 46 (1), 29–57, 10.1177/1091142117698035. OA20

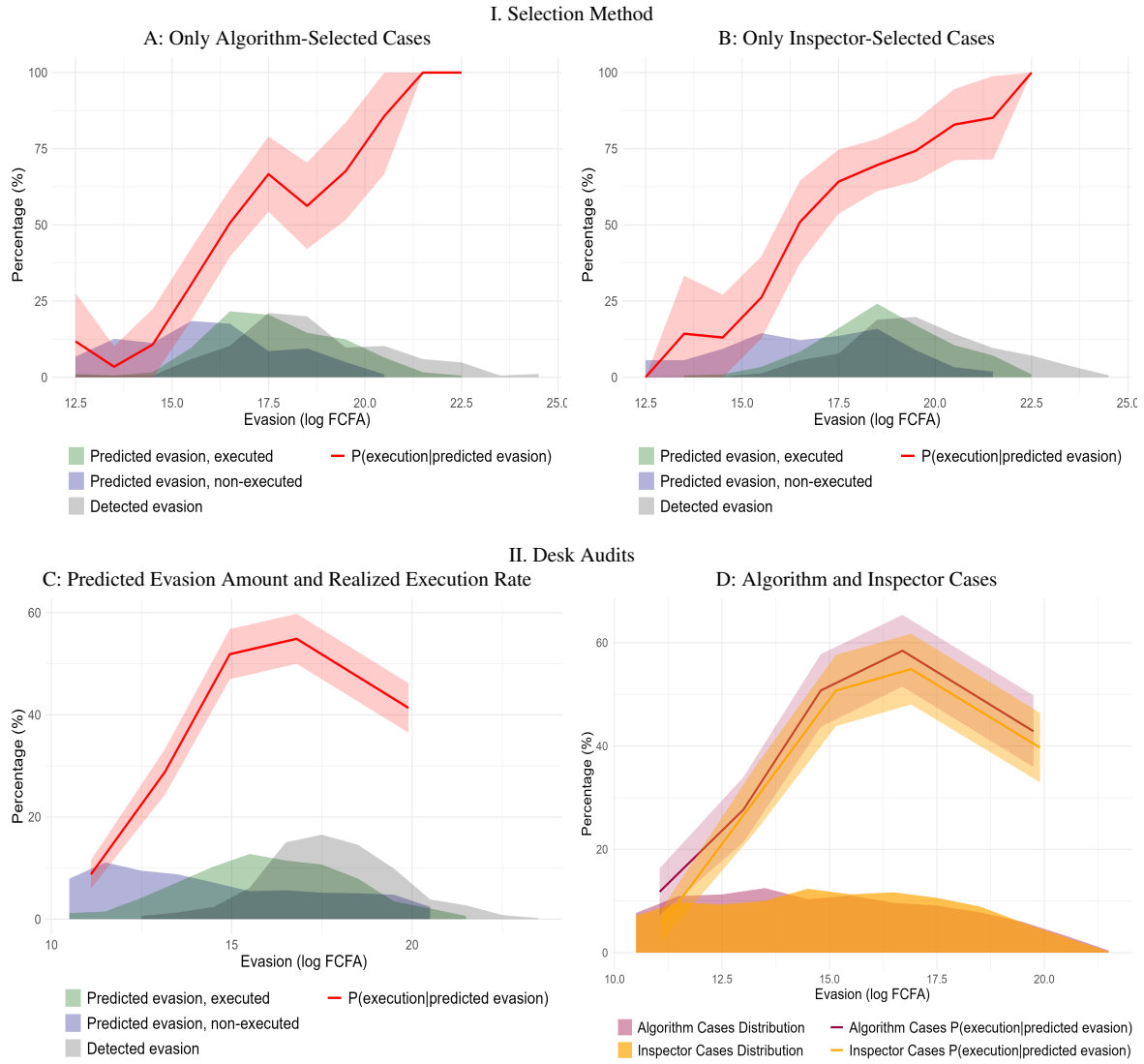
Dwenger, Nadja, Henrik Kleven, Imran Rasul, and Johannes Rincke (2016) "Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany," *American Economic Journal: Economic Policy*, 8 (3), 203–232. OA20

Paradisi, Matteo and Elia Sartori (2024) "Using Prediction Models to Design Tax Enforcement: Incentives vs Targeting," *CEPR Working Paper*. OA15

[37] One exception is D'Agosto et al. (2018), who find that a 1% increase in the provincial audit rate raises compliance by 1.4%. This may reflect differences between macro- and micro-level approaches.

E Additional Results on Mechanisms

Figure E.1: Predicted Evasion, Detected Evasion, and Audit Execution



Notes: This figure supplements Figure 4. Panels in part I replicate Panel A of Figure 4 separately for audits selected by the algorithm (Panel A) and audits selected by inspectors (Panel B). Panels in part II present the results for desk audits, for the whole subsample (Panel C) and comparing algorithm and inspector-selected cases (Panel D). This figure is discussed in Section 7.2.

Table E.1: Observable Characteristics Cannot Explain the Execution or Detection Gap

A: Outcome P(Execution)					
	(1)	(2)	(3)	(4)	(5)
Algorithm	-0.172***	-0.123***	-0.121***	-0.117***	-0.115***
	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)
N	944	944	944	944	944
R2	0.262	0.332	0.334	0.338	0.344
Mean outcome	0.53	0.53	0.53	0.53	0.53

B: Outcome P(Detection|Execution)					
	(1)	(2)	(3)	(4)	(5)
Algorithm	0.019	0.024	0.027	0.027	0.027
	(0.029)	(0.030)	(0.030)	(0.029)	(0.030)
N	508	508	508	508	508
R2	0.110	0.122	0.130	0.130	0.137
Mean outcome	0.89	0.89	0.89	0.89	0.89

C: Outcome log(evasion)|Detection					
	(1)	(2)	(3)	(4)	(5)
Algorithm	-0.651***	-0.528***	-0.556***	-0.533***	-0.532***
	(0.183)	(0.186)	(0.187)	(0.186)	(0.186)
Turnover	No	Yes	Yes	Yes	Yes
Profit rate	No	No	Yes	Yes	Yes
Productivity	No	No	No	Yes	Yes
Firm char.	No	No	No	No	Yes
N	457	457	457	457	457
R2	0.323	0.368	0.386	0.404	0.406
Mean outcome	19.26	19.26	19.26	19.26	19.26

Notes: Column 1 in this table reproduces the main results from Table 6 for full audits. Columns 2-5 successively add firm characteristics as control variables. Column 5 adds controls for firm location and sector. Robust standard errors (Huber-White) are shown in parentheses. This table is discussed in Section 7.2, footnote 29.

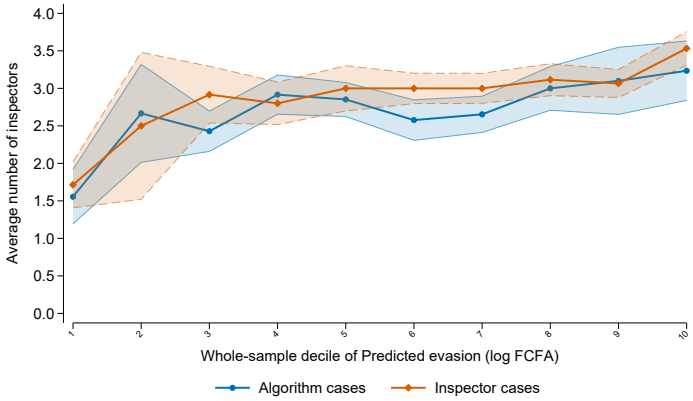
Table E.2: Prediction Model Using Ranking of Turnover Only

	Full Audits			Desk Audits		
	All executed	Algorithm	Inspectors	All executed	Algorithm	Inspectors
DGE	68	58	73	45	64	47
CME1	86	84	86	64	65	69
CME2	81	79	80	75	74	81
CPR	73	57	88	44	44	51
DSF	53	50	63	61	68	62

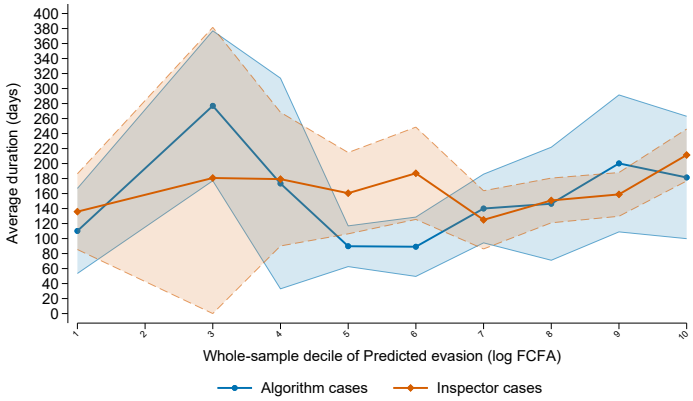
Notes: This table shows the percentage of audited cases that are correctly predicted by a simple turnover ranking within tax office and year, using declared turnover of the year prior to selection. The columns show the percentages of the N executed cases within each office that are also among the N largest cases in the selection list. For the algorithm and inspectors' columns, N is defined as the number of executed cases among algorithm-selected and inspector-selected cases. This table is discussed in Section 7.2.

Figure E.2: Team Size and Audit Duration by Predicted Evasion and Case Type

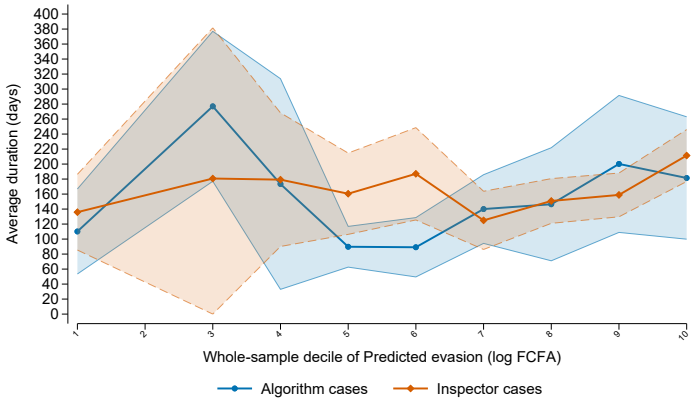
A: Number of Inspectors



B: Duration in Days (Survey)



C: Days from Start to Conf. (Admin. Data)



Notes: This figure is similar to Figure 4 but considers how investment in audits, as captured by the number of inspectors on the audit team (Panel A), and the duration of audits (Panels B and C) varies with predicted evasion and audit selection mechanism, conditional on execution. This figure is discussed in Section 7.2.

Table E.3: Learning from High-Potential Algorithm Cases, Full Audits

	No FE			Selection year FE			Bureau FE		
Outcome: Execution Rate of Algorithm-Selected Cases	Top 2	First 10% seq	First 20% seq	Top 2	First 10% seq	First 20% seq	Top 2	First 10% seq	First 20% seq
A: Discrete measures									
High potential evasion case in list =1	0.218*	0.289**	0.207*	0.176	0.224**	0.199**	0.144	0.245	0.165
	(0.124)	(0.123)	(0.108)	(0.105)	(0.102)	(0.087)	(0.185)	(0.173)	(0.153)
Low potential evasion case in list =1	-0.079	0.044	-0.108	-0.079	-0.007	-0.072	-0.067	0.085	-0.049
	(0.108)	(0.112)	(0.116)	(0.105)	(0.110)	(0.112)	(0.122)	(0.127)	(0.171)
Selection Year FE	No	No	No	Yes	Yes	Yes	No	No	No
Bureau FE	No	No	No	No	No	No	Yes	Yes	Yes
Observations	29	29	29	29	29	29	29	29	29
R-squared	0.129	0.118	0.208	0.366	0.350	0.420	0.424	0.431	0.430
B: Continuous measure									
Share of high predicted evasion case at the top	0.532**	0.530**	0.556***	0.457**	0.403	0.481**	0.398	0.492	0.444
	(0.225)	(0.240)	(0.187)	(0.190)	(0.238)	(0.216)	(0.325)	(0.333)	(0.310)
Selection Year FE	No	No	No	Yes	Yes	Yes	No	No	No
Bureau FE	No	No	No	No	No	No	Yes	Yes	Yes
Observations	29	29	29	29	29	29	29	29	29
R-squared	0.112	0.076	0.097	0.350	0.313	0.343	0.416	0.422	0.417

Notes: This table reports regression estimates of how the presence of high- or low-predicted-evasion (high- or low-potential) cases at the top of an algorithm list is associated with the execution rate of algorithm-assigned audits. The unit of observation is a bureau-by-selection-year list (b, t). The dependent variable is y_{bt}, the execution rate of algorithm-selected cases in bureau b and selection year t, defined as the number of executed algorithm-selected cases divided by the number of assigned algorithm-selected cases. To classify audit-case potential, we use estimated predicted evasion from the Random Forest model and partition it into quintiles within tax office $\times$ selection year. High predicted evasion corresponds to the top quintile, and low predicted evasion corresponds to the bottom quintile. Panel A presents results from the following discrete OLS specification: $y_{bt} = \alpha + \beta_H \text{HighTop}_{bt} + \beta_L \text{LowTop}_{bt} + \delta_t + \mu_b + \varepsilon_{bt}$, where $\text{HighTop}_{bt} = 1$ if the top of the ranking contains a high-predicted-evasion case under the relevant top-of-list definition (Top 2, first 10% of the sequence, or first 20% of the sequence), and 0 otherwise. LowTop_{bt} is defined analogously for low-predicted-evasion cases. Panel B presents results from an alternative specification using a continuous measure of high-potential cases: $y_{bt} = \alpha + \theta \text{ShareHighTop}_{bt} + \delta_t + \mu_b + \varepsilon_{bt}$, where ShareHighTop_{bt} is the share of high-predicted-evasion cases at the top of the list under the same top-of-list definitions used in Panel A. Columns labeled No FE omit fixed effects, Selection Year FE include δ_t only, and Bureau FE include μ_b only. Standard errors are robust. Stars denote significance at $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$. This table is discussed in Section 7.3.

Table E.4: Association of Risk Score Components with Audit Outcomes

	P(Execution)				P(Detect)|Execution				log(Evasion)|Detection			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Full audits	Full audits	Desk audits	Desk audits	Full audits	Full audits	Desk audits	Desk audits	Full audits	Full audits	Desk audits	Desk audits
Algorithm	-0.27***	-0.27***	-0.07***	-0.08**	-0.02	-0.02	-0.05	-0.02	-1.06***	-1.26***	0.07	-0.03
	(0.05)	(0.06)	(0.03)	(0.03)	(0.04)	(0.05)	(0.04)	(0.05)	(0.35)	(0.38)	(0.24)	(0.29)
VAT anomalies risk score		0.01		0.00		0.01		-0.00		0.34		0.22
		(0.03)		(0.02)		(0.02)		(0.02)		(0.21)		(0.17)
CIT anomalies risk score		-0.04		0.02		-0.01		0.00		0.13		-0.07
		(0.03)		(0.02)		(0.03)		(0.02)		(0.20)		(0.17)
Inconsistencies risk score	0.03	0.04	0.01	0.01	0.02	0.02	0.03	0.03	0.27	0.24	-0.14	-0.15
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.20)	(0.20)	(0.14)	(0.15)
Anomalies risk score	0.04*	0.06	0.02	0.00	0.03*	0.03	-0.02	-0.02	0.17	-0.09	-0.02	-0.11
	(0.02)	(0.04)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.12)	(0.21)	(0.13)	(0.22)
Algorithm x Inconsistencies	-0.00	-0.01	0.01	0.01	-0.03	-0.03	-0.02	-0.01	-0.43*	-0.41	0.19	0.22
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.25)	(0.25)	(0.18)	(0.18)
Algorithm x Anomalies	-0.05	-0.05	-0.02	-0.01	-0.06*	-0.06	0.02	-0.03	-0.05	0.27	0.05	0.20
	(0.03)	(0.04)	(0.02)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.18)	(0.26)	(0.15)	(0.27)
Algorithm x VAT anomalies		-0.05		-0.01		0.00		0.08*		-0.49*		-0.40*
		(0.04)		(0.02)		(0.03)		(0.04)		(0.27)		(0.23)
Algorithm x CIT anomalies		0.06*		-0.00		0.00		-0.01		-0.02		0.23
		(0.03)		(0.02)		(0.03)		(0.03)		(0.24)		(0.20)
N	944	944	2731	2731	508	508	992	992	457	457	728	728
R2	0.27	0.27	0.32	0.32	0.12	0.12	0.43	0.44	0.34	0.34	0.42	0.43
Mean outcome	0.53	0.53	0.37	0.37	0.89	0.89	0.73	0.73	19.26	19.26	17.73	17.73

Notes: This table is similar to Table 9, but shows the association between audit outcomes and components of the risk score. This table is discussed in Section 7.4.1.

F Additional Results on Machine-Learning Algorithm

Table F.1: Machine Learning Applications to Enhance the Work of Government Officials

Authors, Year	Paper	Journal	Policy, Country	Tool	Result
Andini et al. (2018)	Targeting with Machine Learning: Tax Rebates	JEBO	Tax rebates to boost individual consumption, Italy	Decision Tree, k-NN, Random Forest	ML-based targeting of the tax rebate to consumption-constrainted households, compared to means-tested targeting, could have increased food consumption by 41.8% or, alternatively, 29.5% of program funds could have been saved, holding food consumption constant.
Andini et al. (2022)	Machine Learning for Public Credit Guarantees	JEBO	Credit guarantees for firms, Italy	Decision Tree, Random Forest, Lasso	ML-based targeting of guarantees to firms that are both credit-constrained and credit-worthy could increase firms' likelihood of obtaining a bank loan by approximately 100%, compared to a simple rule-based targeting.
Battaglini et al. (2025)	Refining Public Policies with Machine Learning: The Case of Tax Auditing	JoEctrics	Tax audit selection, Italy	Random Forest	The ML algorithms can rank audits based on expected tax evasion and recovered evasion. Replacing the 10% worst-performing audits with audit cases selected by the algorithm can increase detected evasion by 38% and recovered evasion by 29%.
Chalfin et al. (2016)	Productivity and Selection of Human Capital	AER P&P	Police hiring, US	Stochastic Gradient Boosting	ML-based police hiring decisions could reduce officer miscondoct (involvement in shootings or verbal abuse) by 4.8%.
Chalfin et al. (2016)	Productivity and Selection of Human Capital	AER P&P	Teacher tenure decisions, US	Regressions with lasso penalty for model complexity	ML-based tenure decisions for teachers, compared to relying on principal ratings, could increase student test score gains by 75% for maths and 105% for English. Using ML for teacher tenure decisions is 2-3 times as effective as reducing elementary school classroom size by one third.
Glaeser et al. (2016)	Crowdsourcing City Government	AER P&P	Restaurant inspections, US	Random Forest, Grad Boosted Trees	The study uses a prediction tournament to explore how tournaments can improve upon traditional consultancy. Implementing the top-performing algorithm from the tournament could increase restaurant inspection productivity by 30-50%.
Hino et al. (2018)	Machine Learning for Environmental Monitoring	Nature	Water pollution inspections, US	Regression Forest	ML-based water quality inspections can detect 600% more violations of pollution limits compared to discretionary targeting. When accounting for local budgetary constraints and maintaining a minimum probability of inspection for all facilities, the algorithm could still increase the number of violations detected by 100%.
Johnson et al. (2023)	Improving Regulatory Effectiveness through Better Targeting	AEJ App	Workplace safety inspections, US	Super Learner	ML-based targeting of inspections could avert 120% more injuries than the current random assignment. This approach could generate up to 876 million in social value by optimizing inspection allocations.
Kleinberg et al. (2018)	Human Decisions and Machine Predictions	QJE	Bail decision, US	Gradient Boosted Trees	ML-based bail decisions can reduce crime by up to 24.7% without increasing jail rates, or alternatively, decrease the number of people jailed by up to 41.9% without raising crime rates.

Notes: This table summarizes the literature on how machine learning algorithms can enhance or substitute for the work of government officials. This table is discussed in Sections 1.1 and 8.

Table F.2: Performance of Random Forest Prediction

A: Full Audits				
	A1: Predicting Detection | Execution		A2: Predicting Evasion Amount | Detection	
	Precision	Recall	MSPE	R^2
	(1)	(2)	(3)	(4)
Overall OOS	0.88	1	2.18	0.35
Overall OOB	0.90	1	3.33	0.25
2018	0.92	1	2.47	0.31
2019	0.81	1	1.93	0.43
2020	0.94	1	2.27	0.09

B: Desk Audits				
	B1: Predicting Detection | Execution		B2: Predicting Evasion Amount | Detection	
	Precision	Recall	MSPE	R^2
	(1)	(2)	(3)	(4)
Overall OOS	0.80	0.93	2.38	0.24
Overall OOB	0.79	0.94	2.51	0.24
2018	0.91	0.95	2.94	0.31
2019	0.66	0.89	2.10	0.18
2020	1.00	0.98	2.07	0.05

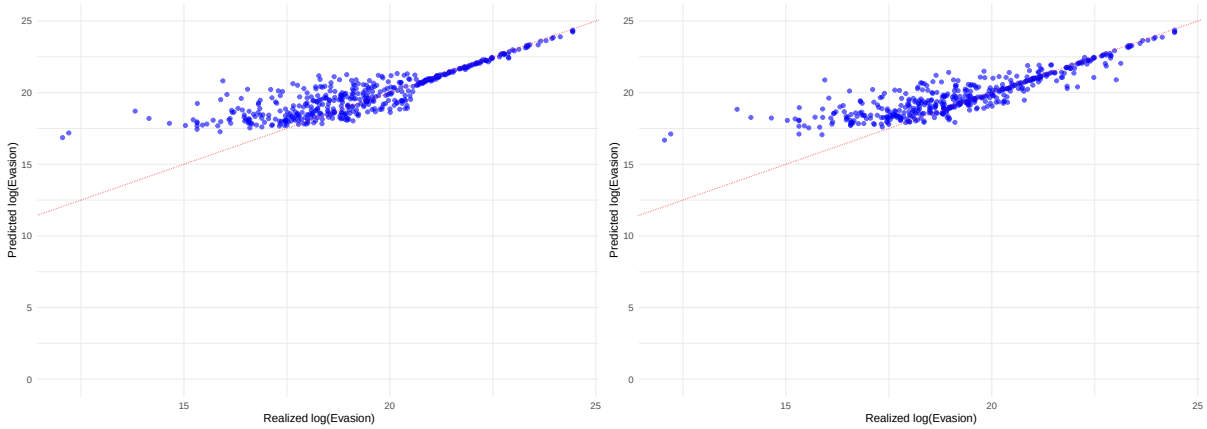
Notes: This table documents the performance of our random forest algorithm for predicting detection of evasion conditional on audit execution (Panels A1 and B1) and predicting the amount of evasion conditional on detection (Panels A2 and B2). The performance metrics are either calculated out-of-sample (i.e. in the hold-out/testing sample, rows 1 and 3-5 in each panel), or out-of-bag (row 2). The precision rate (column 1) is the number of true positives over the sum of true positives and false positives. The recall rate (column 2) is the number of true positives over the sum of true positives and false negatives. A true positive is a case with detected evasion that is predicted to detect evasion. A false positive is a case with predicted evasion in which no evasion was detected in reality. The mean squared prediction error (MSPE, column 3) is the average squared difference between the predicted and observed values. The out-of-sample R2 (column 4) is the ratio of the variance explained to the total variance in the outcome. Figure F.1 further evaluates the accuracy of predicted evasion compared to realized evasion. This table is discussed in Sections 7.1 and 8.

Figure F.1: Within-Sample Model Fit: Comparison of Predicted and Realized Evasion

A: Full Audits

A1: Top Quartile Weights

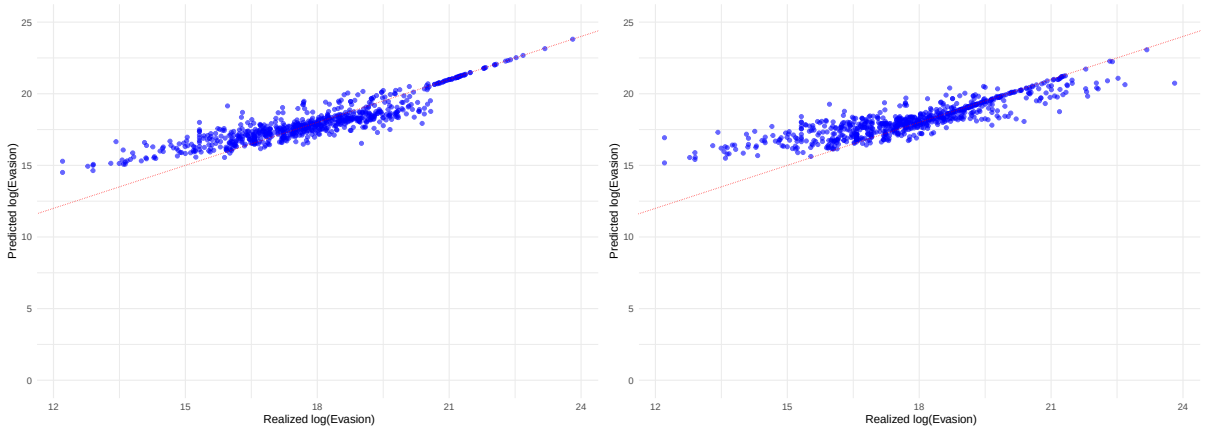
A2: Top Quartile Weights Within Office



B: Desk Audits

B1: Top Quartile Weights

B2: Top Quartile Weights Within Office

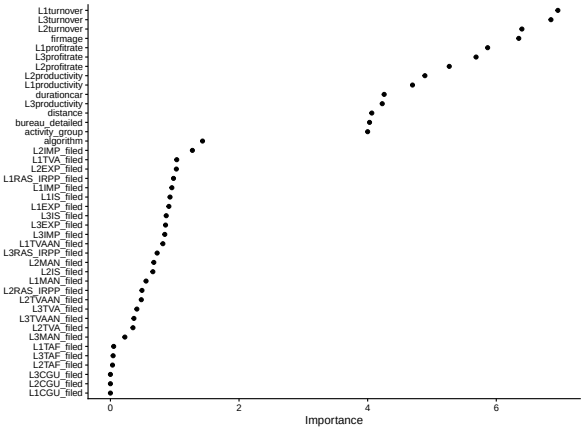


Notes: This figure shows scatterplots of the predicted and realized evasion, for full audits (Panel A) and for desk audits (Panel B), conditional on predicted evasion being non-zero. In each panel, we show results for a model where we increase the weight ten-fold on all cases in the top quartile of realized evasion (left column), or on all cases in the top quartile of realized evasion within each tax office (right column). Table F.2 shows other performance metrics of the prediction. This figure is discussed in Sections 7.1 and 8.

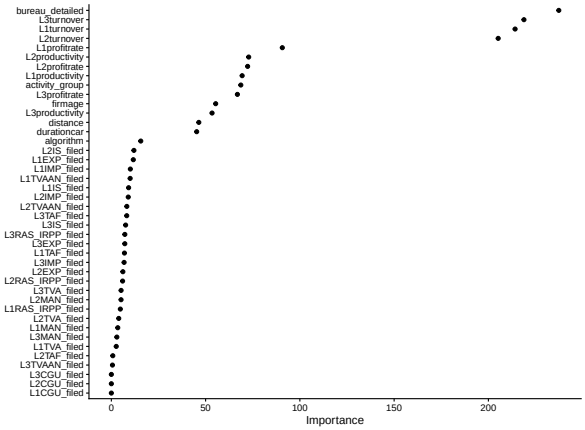
Figure F.2: Importance Ranking of Predictors in Random Forest Algorithm

A: Full Audits

A1: Predicting Detection

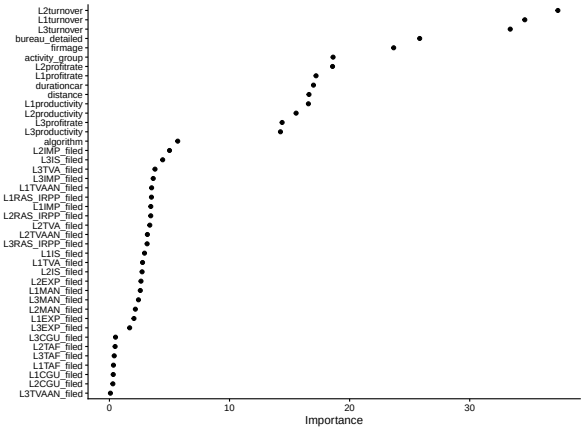


A2: Predicting Evasion Amount | Detection

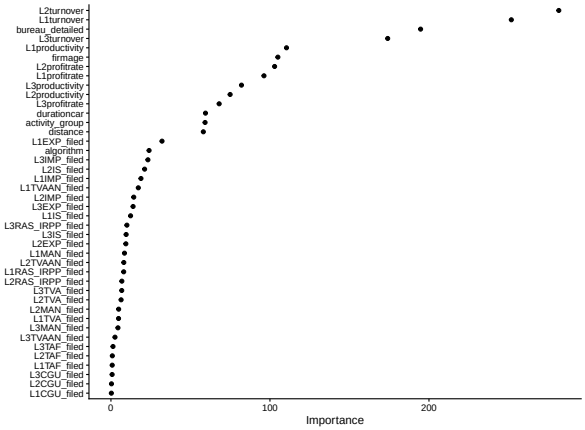


B: Desk Audits

B1: Predicting Detection



B2: Predicting Evasion Amount | Detection



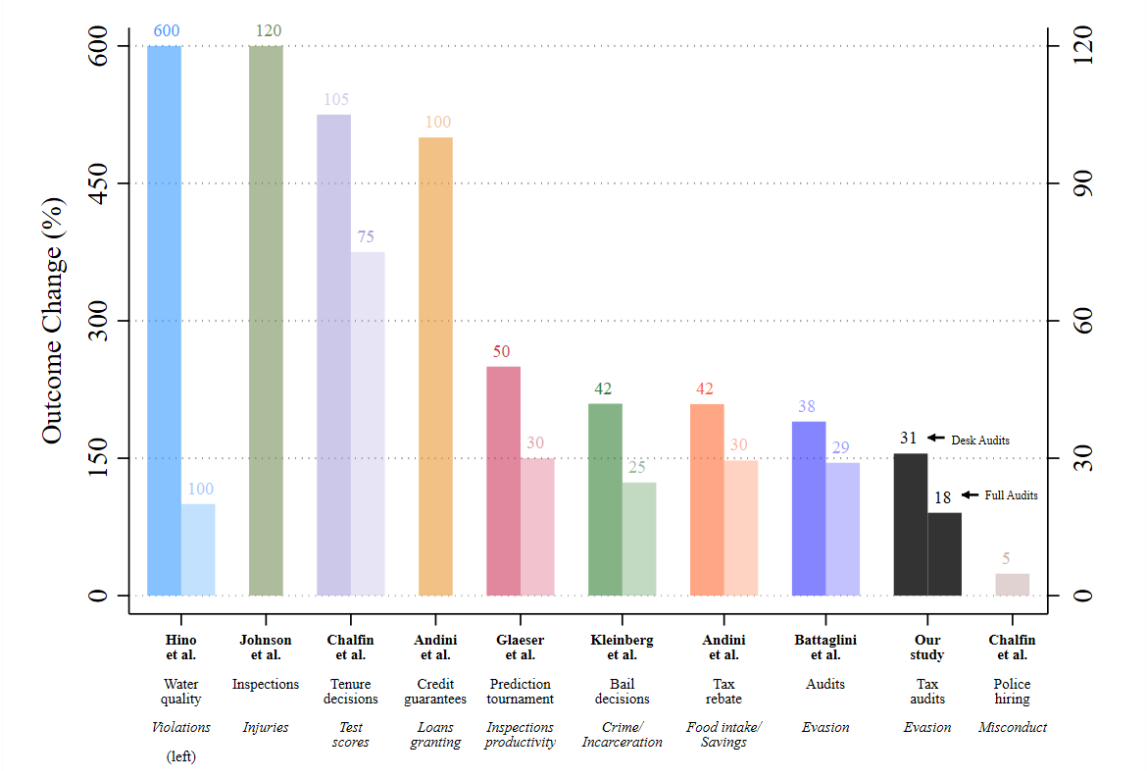
Notes: This figure shows the importance ranking of predictors in the random forest algorithm used to predict whether or not an audit case detects any evasion (column 1), and for predicting the amount of evasion among the cases with non-zero evasion (column 2), for full audits (Panel A) and desk audits (Panel B). This figure is discussed in Sections 7.1 and 8.

Table F.3: Optimization Gains from Machine-Learning Algorithm Trained on Audit Outcome Data, Results Disaggregated by Tax Office, for Full Audits

Realized Revenue Log(mean)	Predicted Revenue Log(mean)	Δ Revenue vs Predicted w/ RF Selection Among Program Cases	Overlap Between Optimized and Realized Audit Program
(1)	(2)	(3)	(4)
A: Large Taxpayer Office			
21.92	21.82	+ 19.17%	60%
B: Medium Taxpayer Office			
19.87	19.87	+ 10.75%	77%
C: Liberal Professions Office			
20.66	20.46	+ 10.3%	62%
D: Small Taxpayer Office			
18.38	18.30	+ 91.15%	30%

Notes: This table is similar to Table 11, Panel A, but shows the results disaggregated by tax office. This table is discussed in Section 8.

Figure F.3: Comparison of Optimization Results With Literature



Notes: This figure depicts the estimated improvements in outcomes in several papers that use machine learning to automate decisions in several settings. The outcomes and study designs of each of these other papers are explained in Appendix Table F.1. The results of our study are shown in black bars, and refer to the estimated increases in detected evasion according to the optimized execution of cases for full and desk audits, as discussed in Table 11. This figure is discussed in Section 8.