Too Much of a Good Thing: Regulatory Discretion and Monitoring Intensity of Building Inspector Performance

Manuel Barron[1], Guadalupe Bedoya[2], Diego García Montúfar[3], Paul Gertler[4], and Ana Goicoechea[5]

[1]Department of Economics, Universidad del Pacífico
[2]Development Impact Evaluation, The World Bank
[3]Finance, Competitiveness and Innovation, The World Bank
[4]Haas School of Business, UC Berkeley and NBER
[5]Finance, Competitiveness and Innovation, The World Bank

February 10, 2026

Abstract

We use a field experiment on third-party monitoring in the context of business safety inspections and show that intense monitoring can backfire, undermining regulatory efficiency. In our setting, all firms are required to pass a business safety inspection to obtain an operating license. In collaboration with the municipality of Lima in Peru, business safety inspectors were randomly allocated to a treatment or a control group. Treatment inspectors were exposed to high-intensity third-party random backchecks for around 70 percent of their inspections, while those in the control group were exposed to low-intensity backchecks for around 20 percent of their inspections, all conducted 24 to 48 hours after the inspection. Periodic feedback on inspector performance, measured as the rate of discrepancy between the auditor and the inspector, was sent to inspectors, and their managers received aggregate performance indicators. There was no consequence for poor performance. After four months, the proportion of businesses passing the inspection fell to 61 percent in the high-intensity monitoring group, compared to 80 percent in the low-intensity group. The reduction in firms passing the inspection in the high-intensity group is mostly driven by a reduction in inspection accuracy, with an increase of compliant items incorrectly assessed as noncompliant. This leads to increases in post-inspection investments by firms, a small increase on an index of corruption, and large decreases in firm satisfaction with inspectors, the inspection process, and the regulation. We hypothesize these results can be due to loss- or risk-aversion with inspectors being harsher for fearing being found making mistakes by passing firms that did not comply with the regulation.

1 Introduction

Third-party monitoring has become one of the main policy tools employed to ensure regulatory and contract compliance in a wide array of settings such as agrifood (Hatanaka et al., 2005; Potoski and Prakash, 2005), credit rating (Bolton et al., 2012; Strobl and Xia, 2012; White, 2010; Ronen, 2010), environmental and labor law standards (Duflo et al., 2013, 2018; Johnson et al., 2023), public procurement (Celhay et al., 2024; Zamboni and Litschig, 2018) and even teacher absences (Duflo et al., 2012). Although there is substantial evidence that monitoring improves compliance, little is known about how much monitoring is optimal. Intensive monitoring may backfire if increased intensity leads to defensive behaviors that increase costs without improving outcomes. For example, the threat of monitoring health outcomes combined with the threat of medical malpractice litigation leads to expensive defense medicine – e.g., ordering unnecessary diagnostic tests and prescribing non-efficacious medicines (Kakemam et al., 2022; Nahed et al., 2012; Eftekhari et al., 2023).

We examine the effect of the intensity of building safety inspection monitoring conducted as part of the business licensing process, one of the most significant bureaucratic barriers to doing business in lower and middle-income countries (Molfetas, 2019). Building safety regulations typically allow substantial inspector discretion. For example, in Peru, a regulation requires evacuation routes (entrances, exits, hallways, stairs, etc.) to be at least 1.20m wide "and/or allow evacuation in a safe manner". Although the regulation provides some guidance as to the minimum allowed width allowed, ultimately it is up to each inspector to decide whether the evacuation route allows for "safe" evacuation. Regulatory discretion also applies when there are multiple items of the same type. How many electrical outlets need to be inspected and how are they chosen? An inspector who examines all power sockets is more likely to find faulty sockets than an inspector who examines only a sample. Because inspector performance is not monitored and contracts are not performance based, regulatory discretion may lead to poor performance either through sub-optimal effort or corruption, i.e. seeking side payments from firms for favorable inspection reports.

One potential solution is to introduce inspector accountability through external monitoring of inspector performance that provides feedback to inspectors and their managers. We use a field experiment to examine the impact of such a monitoring scheme in which anonymous external auditors collect information on compliance with safety regulations in a random sample of recently inspected establishments shortly after the official inspection. The quality of inspection is the discrepancy between the assessments of inspectors and third-party auditors. We ask whether more intensive monitoring improves inspection quality or leads to a defensive response, thus reducing quality and increasing costs.

We conduct the study in Lima, Peru.[1] Inspectors were randomized into two groups: (1) a high-intensity group where a random sample of 70 percent of an inspector's inspections was monitored, and (2) a low-intensity group, where a random sample of 20 percent of an inspector's

[1]In Peru, the costs of inspection are high. Firms cannot open without passing inspection. More than 30% of firms report that they undergo inspection two or more times. More than 40% hire technical consultants to help prepare for inspection and often make substantial and unnecessary investments in infrastructure to pass inspection. The process can delay business opening by months.

inspections was monitored. Inspectors were informed ex ante about the share of inspections that would be monitored. The third-party audit visits to establishments were conducted between 24 and 48 hours after the original inspection was completed and were unannounced, meaning that the inspectors did not know which businesses would be monitored. Reports of inspection quality based on discrepancies between the assessments of inspectors and third-party auditors were provided to inspectors and their managers.

We have four main findings. First, high-intensity monitoring led to a significant decrease in the proportion of businesses that passed the inspection by 19 percentage points. Second, the pass rate reduction is explained mostly by an increase in items that were (wrongly) marked as non-compliant by inspectors but judged as in-compliance by the auditors. These results are consistent with risk adverse inspectors using their discretion to flag more potentially non-complaint items to protect themselves from a poor performance report, considering that erring on the side of caution is safer. Third, we find large increases in the regulatory compliance costs to firms and large decreases in satisfaction with the inspection process. Finally, we find evidence of increased corruption based on two indirect indicators: firms report that inspectors suggested that they hire a specific consultant, and firms report knowing other businesses that were asked for bribes.

These results provide an important caveat to a growing body of evidence suggesting that independent monitoring can improve public service delivery and reduce inefficiencies.[2] Duflo et al. (2013) find that independent private auditors and third-party verification of their audits decrease false reports of plants complying with regulations on pollution limits and reduce pollution among treatment plants in India. Avis et al. (2018) show that auditing the use of public resources in municipalities reduces corruption in Brazil, and the reduction is higher for those which had been previously audited. Duflo et al. (2012) show that monitoring teachers' attendance, combined with attendance-based incentives, reduces teacher absenteeism, and increases student learning outcomes. These settings have at least two important features in common. First, the monitoring is conducted on outcomes that have a low degree of discretion, such as whether the plant pollution reading on a monitoring device complies with the regulatory limits, the public works construction exists, or teachers are present teaching their classes. Second, the incentives—whether through the type of contracts with the agents subject to the monitoring, the type of performance-based financial payments, legal punishment, or reputation costs—were aligned to produce such impacts.

A central issue in third-party monitoring is discretion. At the regulatory agency level, discretion plays a positive role in selecting the focus of auditing resources. For example, regulating agencies have been effective in monitoring dirtier industries with higher intensity (Gupta et al., 2019; Duflo et al., 2018; Franco-Bedoya and Mani, 2020; Griffin and Tang, 2011). While Duflo et al. (2018) show that a regulating agency with some discretion can be more effective than random auditing in selecting which establishments to inspect, Johnson et al. (2023) show that this selection can be further improved with machine learning methods. In a similar vein, Dal Bó et al. (2021) show that supervisors can be effective in selecting which agents (in their setting, agricultural extension agents) would be more responsive to monitoring. However, at the inspector level, discretion can be undesirable as it can cause stark inconsistencies in bureaucratic decision-making (Ho and Sherman, 2017). Most of the specialized literature puts the blame on

[2] In a study of illegal fishing in Chile, Gonzalez-Lira and Mobarak (2021) also find that auditing is more effective when conducted at a lower frequency.

incentive misalignment. For instance, it has been shown that auditors are more lenient when paid by the firm or when they offer consulting services in parallel (Duflo et al., 2013; Short et al., 2016), despite some heterogeneity in inspector quality even when offering consulting services in parallel (Duflo et al., 2013). Bazerman et al. (2002) argue that behavioral biases also play a role in the heterogeneity of inspectors' assessments. This claim has been explored by Ibanez and Toffel (2020), who document behavioral biases like outcome effects and daily schedule effects among inspectors of the restaurant industry in the US.

In our setting, with flexible regulation that allows discretion and misaligned incentives for inspectors, higher monitoring seems to have backfired. In this case, low accuracy could stem from low effort, lack of knowledge, erroneous judgment, or corruption. We could consider risk-averse inspectors with a utility function depending positively on income and negatively on effort and the probability of being caught making a bad judgment call. Mistakenly reporting an item as compliant when it is not compliant is perceived by the inspector as a graver error than the opposite, due to its correlation with risk. In a setting like ours, excessive monitoring by a third-party auditor may increase a particular type of error: those resulting from compliant items marked as noncompliant by the inspector. Thus, inspectors may become too strict.

These results have important policy implications. As governments explore accountability mechanisms, such as increased monitoring of public officials, aimed at improving service delivery and reducing inefficiencies, careful attention should be paid to whether the incentives are aligned with the objectives of these measures. In addition, high levels of discretion and flexible implementation guidelines in a context with information asymmetries can potentially create additional sources of costs and inefficiencies for businesses and the government.

2 Institutional Context

Business safety inspections to assess establishments' risks of fire or collapse are a requirement for new firms seeking to start up and formally operate in Peru. Inspections are included as part of the process of obtaining an operating license, a procedure under the jurisdiction of local governments. Firms that pass the inspection are issued a safety certificate, which must be renewed every two years or whenever the firm modifies its existing operating license. During an inspection, a government inspector evaluates compliance with safety items related to risks of fire, collapse, electrocution, falls, and other risks.

Inspectors have a high degree of discretion in assessing regulatory compliance, which implies that two different inspectors assessing the same firm can reach different conclusions. For example, one checklist item requires evaluating if evacuation routes (entrances, exits, hallways, stairs, etc.) are at least 1.20m wide "and/or they allow the evacuation of persons in a safe manner". Even though the regulation provides some guidance as to the minimum permissible width (1.20m), ultimately it is up to each inspector to decide whether the evacuation route allows for a safe evacuation of occupants. Regulatory discretion also applies to what to inspect. The regulation does not provide guidelines on what process to follow when multiple items of the same type need to be checked. An inspector that examines all power sockets is more likely to find faulty sockets than an inspector who examines only a sample.

There are no systematic monitoring mechanisms to assess inspector performance (i.e., the

quality of their assessments). In general, assessing inspector performance based on safety outcomes is difficult because adverse safety outcomes that result from inadequate inspection (e.g., fire or building collapse) are not frequently observed, and those events can have other natural causes unrelated to the inspection. Once these low-probability events occur, however, they impose very large societal costs: in 2001, a fire in a commercial area of the district of Lima caused more than 400 deaths and the collapse of several buildings (Vadillo, 2021).

Municipalities have limited capacity to monitor inspectors' work. For example, it is common for municipalities to review inspection checklists and photographic evidence from the inspections after they have been submitted by inspectors, but this review focuses on whether the forms are complete, not on indicators of quality. Finally, most inspectors work under temporary contracts, and many work in multiple municipalities, receiving a fixed wage per day, which provides little incentive to exert effort in carrying out the inspections.

The diligence and thoroughness of inspections is important because it can affect inspection outcomes, the firm's costs of compliance, granting business licenses and safety. However, inspectors have few incentives to be thorough or fair. Their remuneration is not performance based nor are there consequences for poor performance. The resulting inefficiency in the inspection process can create high compliance costs to firms (e.g., having to apply for a certificate multiple times) and opportunities for corruption.

Inspection Types, Inspection Checklist, and Inspection Results

Before the on-site inspection, firms are classified by the Municipality into high, medium, or low risk, based on a few self-reported characteristics such as business sector and establishment size. High-risk firms are required to pass inspection before they can obtain a business and begin to operate. The inspection comes before the license is issued, i.e., it is called an ex ante inspection. Low- and medium-risk businesses are issued an operating license upon application and are allowed to operate immediately but must pass the business safety inspection within a specified amount of time afterward, i.e., this is an ex post inspection. Ex-ante inspections are conducted by teams of two or three inspectors, whereas ex post inspections are conducted by just one. In the municipality of Lima, 92 percent of the inspections conducted in 2021 were for low- and medium-risk businesses. We study ex-post inspections only.

Ex-post inspections are conducted using a checklist (formally called Annex 6) that contains 67 items, of which 51 are labeled major ("relevant") and 16 minor ("irrelevant") depending on the risk they carry. Major risk items are those that pose serious safety threats (e.g., having a faulty gas tank), whereas minor risk items pose a small security risk that can be fixed with relative ease (e.g., relocating a fire extinguisher). An inspection visit can have five different potential outcomes: (a) passed — the firm complies with all applicable items; (b) failed —the firm does not comply with at least one major-risk item; (c) corrections needed — the firm complies with all major items, but does not comply with at least one minor item; (d) suspended - the inspection could not be conducted or was interrupted because the establishment was closed or there were physical impediments; and (e) not conducted — the firm is not adequately set up or the wrong risk level was declared.

When corrections are needed, the firm has up to three business days to submit documen-

tation (certificates or photos) demonstrating that the noncompliant items have been corrected. Inspectors then evaluate these documents to determine if the firm ultimately passes or fails the inspection. Not submitting the documentation on time results in failing the inspection.

Firms that fail the inspection or whose process terminates before they pass the inspection have their operating licenses revoked and must cease operating. These firms must pay US$ 50 to reapply for a business license and start the process again. They are not allowed to restart operations until they pass the building safety inspection and are issued their safety certificate. In Lima, 23 percent of firms do not pass the first inspection and must go through this process.

3 Intervention

The intervention used a team of qualified third-party auditors to conduct backchecks of the inspection reports between 24 and 48 hours after the inspectors finished their assessments. Reports detailing the performance of the inspectors, measured by the rate of discrepancies between the inspector and the auditor were sent to the inspectors and their managers.

Auditor recruitment and training

Auditors consisted of a team of eight professionals who were duly accredited as business safety inspectors by the Ministry of Housing (MVCS) with valid registration in the National Registry of Business Safety Inspectors. The auditors were selected from a pool of 14 applicants who took a written exam. They had at least 5 years of experience as business safety inspectors in local governments or advising businesses preparing for inspections. Selected candidates underwent an onboarding process that included training sessions on how to use the verification electronic checklist app, two mock verifications, and a training session to ensure alignment of evaluation criteria. The auditors also participated in field-training sessions in which they conducted practice verifications in partner businesses that were further reviewed by members of a specialized firm hired to oversee the technical and logistical aspects of verifications. After the first field-training session, auditors received feedback based on the discrepancies identified between their results and those of their trainers. A second field-training session followed, in which groups of auditors conducted verifications of the same businesses along with their trainers to assess and improve inter-rater reliability, that is the degree to which different auditors consistently agreed on the assessments of the same firm.

Firm Assignments and Visit Protocol

Auditors received a daily list of establishments they had to visit. This list was selected at random based on the daily inspection roster shared by the municipality with the research team. Inspections that had been suspended were excluded from the audit list.

To avoid collusion between inspectors and auditors, none of the auditors were from Lima or known to the Lima inspectors. Inspector names and which firms they inspected were not revealed to the auditors at any time. Auditors were rotated and were not attached to specific inspectors. Inspectors were not informed as to which items were being audited. Finally, supervisors conducted random quality checks of the auditors' work.

Auditors were provided with the phone number of firms to call in advance to announce their visit. They introduced themselves with a letter from the Municipality of Lima indicating that the business had been selected for a random check of the municipality inspectors work, with the aim of improving the inspection process. The letter had the phone numbers of the municipality and a member of the research team, so firms could contact the research team or the municipality to inquire about the auditor and ensure that their visit was legitimate. Auditors informed that their visit was not a follow-up inspection, that it did not entail an evaluation of the firm's security condition, and that it would have no effect on the official inspection result - that is, results would not be overturned if auditors found mistakes in the inspector's evaluations. Auditors requested consent from the firm before starting the audit, which, including the introduction, lasted on average 10 to 15 minutes.

Assessment Protocol

As in a regular inspection, the auditors had to assess whether each item was compliant or noncompliant. Audits covered 34 of 67 items in the checklist, 21 of which were major and 13 minor (see appendix A, table A.1). Items were selected using three criteria: (a) how frequently the items were evaluated in past inspections – items that were inspected regularly were prioritized), (b) how easy it was to evaluate the item – items whose evaluation required opening electrical boxes or obtaining certificates that the firm might not have at hand were excluded, and (c) the level of discretion of the item – items that relied heavily on inspector discretion were excluded.

Auditors entered the results in a data entry app on tablets, and were instructed to take photographs of all items whenever possible. If an audit could not be conducted, auditors had to register the reason why.

3.1 Discrepancy Reports

Discrepancies between auditor and inspector assessment were classified as described in appendix table A.2. There are two types of discrepancies:

1. Inspector type I error. This happens when an inspector evaluates an item as noncompliant, but the auditor evaluates the same item as compliant. We call it type I error because the inspector rejected compliance mistakenly. An illustration of this case is presented on the left-hand side of Figure 1.

2. Inspector type II error. This happens when an inspector evaluates an item as compliant, but the auditor evaluates the same item as non compliant. An illustration is presented in the right-hand side of Figure 1.

Figure 2 presents a matrix with examples of the two types of discrepancies incorporated in the analysis (Type I and II errors) and their interaction with item relevance (major- or minor-risk), which affects the inspection outcome.

Figure 2: Types of Discrepancy by Item Relevance - Examples Notes: Major risk items in the regulation are those for which noncompliance poses a serious safety risk; minor risk items are those for which noncompliance poses a small security risk. Item relevance is determined by the



The glass cabinets are laminated, as required by the regulation, but the inspector marked the corresponding item as noncompliant. This is a "compliant item evaluated as noncompliant" (Type I error).



The space in front of this electrical board (in the red circle) and the evacuation route (marked by the yellow arrow) are blocked by merchandise; however, the inspector marked both items as compliant. These are two "noncompliant items evaluated as compliant" (Type II error).

Figure 1: Examples of Type I and Type II Errors

Source: Barron et al. (2022)

regulation. Failure to comply with at least one major item results in failing the inspection. If all items are compliant but there are noncompliant, minor items, the firm has up to three business days to fix the items and present evidence, through photographs or certificates.

Inspectors received personalized discrepancy reports twice a month (examples presented in appendix B). These reports included the inspector's discrepancy rate in the current period and in previous periods so that inspectors could keep track of their progress. They also included the average discrepancy in the municipality so they can compare their individual result to their peers. The individual reports did not include details on the type of discrepancy or on specific discrepancies that had been found. Inspectors were informed that these reports were individual and would not be shared with the municipality or their managers, who would only receive an aggregate report.

Municipality reports included the average discrepancy rate and the name and mean performance of the three inspectors with the lowest discrepancy rates (the best performers). They also included case studies with the most common type of discrepancies found in the current period. These case studies included photographs and descriptions explaining why the case constituted a discrepancy. For example, a photograph of a plastic electric box indicating that the item "the electric box made of metal is connected to the earth system" had been marked as noncompliant when it should have been marked as nonapplicable because the electric box was plastic, not metal.

Inspectors received their first personalized report on January 12, 2022. The inspector manager received their first aggregate report on February 17, 2022. The aggregate report for the municipality included the average discrepancy rate in the current period, and in previous periods, for all inspectors. By June 30, 2022, inspectors had received 12 personalized reports, and the municipality had received 8 aggregate reports. Appendix B provides examples of the inspector and municipal reports. The intervention started in mid-December but due to a slow implementation in the first two months due to the holiday session, COVID-19, and delays in the delivery of

the audit reports, we focus the impact analyses on the last four months of the intervention when the implementation was working in full. In addition, we present in the appendix the impacts for the whole period (table A.7) and the differences in impact between months 1-2, 3-4, and 5-6 (table A.8).

Compliance

The audit success rate was 89 percent (meaning 89 percent of inspected businesses assigned to receive a verification were verified). Unsuccessful audits were due to the business being closed (66 percent), the manager or representative being absent (8 percent), or other reasons (26 percent), such as the auditor not finding the address or running out of time to complete the audits assigned for the day. Table A.6. in the appendix presents statistics for compliance with treatment—that is, the percentage of inspections that were verified per treatment group and the percentage of inspectors that received the reports with feedback.[3]

4 Experimental Design

Inspectors were randomly assigned to a high-intensity monitoring groups with 70 percent of each inspector's inspections reviewed by the auditors or a low-intensity monitoring group with 20 percent of each inspector's inspections reviewed by the auditors. Before the intervention began, each inspector was informed as to the share of their inspections that would be reviewed by the auditors.

We used a stratified randomized experimental design to estimate the causal impact of high-intensity monitoring on inspection results, inspection quality, and corruption. Inspectors in Lima were randomized into one of two groups stratified by inspector specialization. The 21 inspectors were randomized on December 19, 2021, with 10 assigned to the low intensity group and 11 to the high intensity group. Table A.3 in the appendix presents the number of inspectors in treatment and control groups at randomization and at the end of the intervention.

Overall, the study sample represented 79 percent of the ex post inspections conducted in Lima during the intervention period. The missing inspections were conducted by inspectors hired after the study began and some by inspectors who were assigned to other activities (supervisory visits, administrative activities) at the time of randomization but were reassigned to conduct business safety inspections after randomization.

5 Data Sources

The first source includes administrative inspection outcome data from inspections conducted with ICT tools, as well as administrative records at the file level[4], which form a high-frequency

[3]Inspectors did not receive reports when they did not conduct inspections in the previous reporting period. This could be due to vacation/absence or to inspectors being assigned to other activities. In addition, inspectors did not receive a report when they conducted inspections using paper forms and the data was not processed (scanned and digitized) in time for the report to go out. In these cases, inspectors received the information in the next reporting period.

[4]Administrative records at the file level include all the milestones of the safety certificate process from business inspection request to certificate issuance or file closure.

panel dataset on inspection results. Administrative records at the file level include all the milestones of the safety certificate process from business inspection request to certificate issuance or file closure.

Inspectors recorded inspection outcome data electronically in real time using an application for tablets. They filled out an electronic version of the inspection checklist uploaded to a web-based platform where inspectors and municipal officials can access inspection reports and photographic evidence. The uploaded checklists are used to generate inspection reports sent to firms by email shortly after the inspection visit. Ninety-three percent of inspections were conducted with ICT tools. The remaining 7 percent of inspections were conducted using paper forms that were later scanned and digitized.

Inspectors underwent an intensive training program to ensure a smooth transition to using electronic checklist on the tablet. Before the start of the study, inspectors participated in a three-week training in the use of the electronic checklist via classroom and field training sessions, mock inspections, and shadowing sessions. Inspectors used the electronic checklist for one month before the study began allowing inspectors to get familiar with the electronic checklist and to generate baseline data, and had access to on-demand support over the phone for general troubleshooting issues and overall support.[5]

The second source consists of administrative verification outcome data. Auditors entered the results of the audit in modified electronic checklist on a tablet and received separate training in the use of the tablet like the inspector training.

The third source is an endline survey of business representatives. The survey covered indicators of corruption, business' compliance costs, and business perception and rating of inspector's performance, the inspection process, and the regulation. Fourth, we have baseline data on inspector characteristics (age, experience, education, gender, specialization) and data on personality characteristics (Big Five) collected in a survey conducted in June 2021.

6 Timeline

The timeline of intervention and data collection is presented in Figure 1. The complete period of the intervention was from December 15, 2021, to June 30, 2022. Baseline data on inspections and audits was compiled or collected between November 12 and December 14, 2021. Baseline data includes 2,091 inspections and 149 audits. The success rate for baseline audits was 83 percent. Intervention data on inspections and audits was collected between February 15, 2021, and June 24, 2022. We have data for 3,622 inspections and 1,935 audits with 35,907 checklist items assessed. The success rate for intervention audits was 89 percent. Finally, in June 2022 we conducted an endline survey of 1,001 business managers whose businesses were inspected between March and June 2022. Tables A.3 and A.9 in appendix A detail the sample size and consent rates for the different data sources.

[5]See Barron et al. (2020) for more details on the electronic checklist functionality and performance.

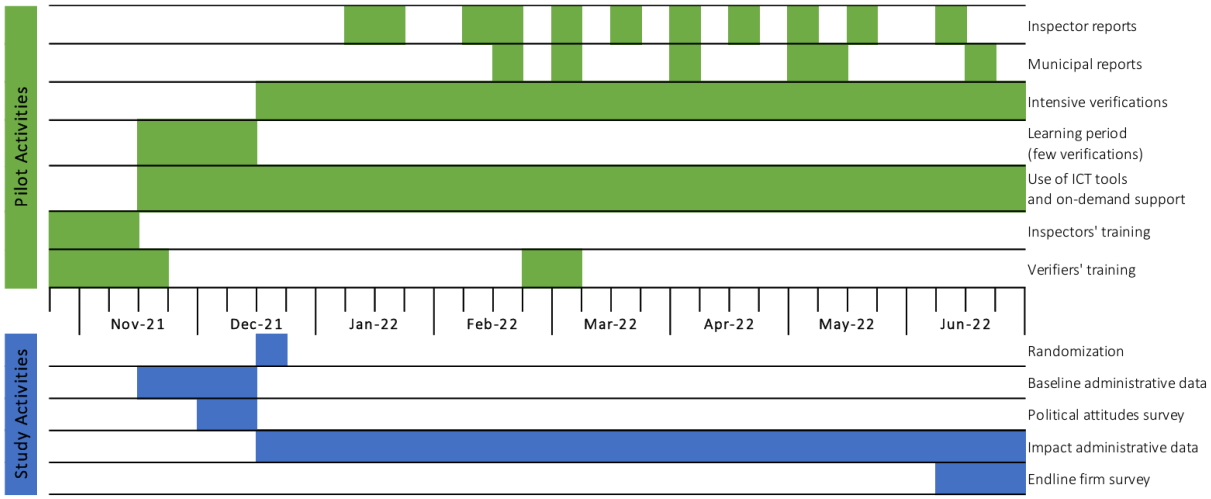


Figure 2: Figure 1: Timeline of the Intervention and Evaluation Activities

Note: ICT = information and communication technology.

7 Outcomes

The outcomes consist of indicators of whether the firm passed the inspection, inspection quality, corruption, investments by the business after the inspection, and indicators of firm satisfaction with inspector's performance, inspection process, and regulation (See A.4 for further details). The inspection outcomes indicators include:

- Passed inspection. Whether the business passed the inspection, and the establishment was issued a safety certificate.
- Number of Evaluated items. Natural log of the number of items evaluated during the inspection (of 67 possible).
- Inspector type I error. Whether an item evaluated as noncompliant by the inspector is evaluated as compliant by the auditor.
- Inspector type II error. Whether an item evaluated as compliant by the inspector is evaluated as noncompliant by the auditor.
- Corruption Index. A standardized weighted average of two measures following Anderson (2008): i) inspector suggested during the inspection to hire consultant, and ii) business owner knows any businesses that were asked for bribes. This index is standardized using the low-intensity group's mean and standard deviation.
- Investment after the inspection (USD): Self-reported amount (in USD) invested in implementing improvements to the establishment after the inspection.
- Investment after the inspection (days): Self-reported time (days) taken to implement improvements to the establishment after the inspection.
- Satisfaction with the inspector's performance. Whether the business representative reports being "satisfied" or "very satisfied" with the inspector's performance during the inspection.

- Satisfaction with the inspector attributes, the inspection process, and related regulations. Whether the business representative reports each of the following indicators as “good” or “very good” from a 5-category scale ranging from "very bad" to "very good.”
- Inspector’s professionalism
- Inspector’s impartiality
- Clarity of the regulation
- Possibility of reporting problems

8 Empirical Methods

We estimate the impact of the intervention as the mean difference in the outcomes of interest between the treatment and the control groups. Equation 1 presents the basic specification for inspection-level data:

$$Y_{ijt} = \alpha + \delta T_i + \emptyset V + \gamma X_{ij} + \varepsilon_{ijt} \tag{1}$$

Y_{ijt} indicates any of the outcomes of interest (for example, passing an inspection) for inspector i in business j at time t; T_i is a treatment indicator that equals one for inspectors belonging to the treatment group and zero for inspectors in the control group; δ is the parameter of interest and represents the impact estimate; X_{ij} represents inspector baseline characteristics and businesses fixed characteristics, including inspector’s age, experience, and the establishment’s area and economic function; and ϵ_{ijt} is the disturbance term. Since we stratified by inspectors’ specialization, we follow Bruhn and McKenzie (2009) and include V, which is a dummy variable equal to one for one of the two randomization strata.

For our quality outcomes at the inspection item level, comparing inspector’s and auditor’s assessments, we use the following specification:

$$Y_{iljt} = \alpha + \theta_l \delta T_i + \emptyset V + \gamma X_{ij} + \varepsilon_{ijt} \tag{2}$$

Y_{ijlt} is an indicator equal to one if there is a discrepancy in any of the items in the regulatory checklist between inspector i and auditor l in business j at time t; θ_l are auditor’s fixed effects; and all other notations are similar to those in equation 1. We additionally estimate how impacts vary by pre-treatment characteristics, particularly, whether the item verified is major-risk, using the following specification:

$$Y_{iljt} = \alpha + \theta_l + \delta T_i + \gamma T_i W + \delta W + \emptyset V + \pi X_{ij} + \varepsilon_{iljt} \tag{3}$$

W is a binary variable equal to one if the item analyzed is major-risk. All other notations are similar to equation 1.

Standard errors are clustered at the inspector level, unless otherwise stated. In addition, p-values are adjusted to control for potential variation stemming from the treatment assignment itself, using randomization inference, with corresponding p-values in brackets, due to the small

number of inspectors or "clusters" in our design (Bruhn and McKenzie, 2009; Young, 2019).[6] Furthermore, to account for multiple hypothesis testing, we also report sharpened q-values in curly brackets following Anderson (2008).

While the intervention lasted 6 months, we only use the last four months of data for analysis to allow time for the auditors to learn their job and get up to speed. We present the full results and heterogeneity analyses by period in the Appendix.

9 Results

Baseline Descriptive Statistics

Table 2 (panel A) shows that in our data, the average inspector in the study is 48 years old and has 11 years of experience as inspector; most inspectors are men (71 percent); have a bachelor's degree as the highest education level (80 percent); and a hold a relevant specialty (57 percent). At the inspection level (panel B), most inspections were conducted in commerce businesses (94 percent), and only a small portion were related to office and administrative services or other types of businesses. These businesses are relatively small, with an area of 41.4 square meters on average. Around 77 percent of firms passed the inspection at either the first or second inspection visit.

Barron et al. (2022) report that even businesses that fail the inspection and need to repeat the process comply with a high proportion of the items in the checklist, with the highest compliance reported for items related to risk of collapse (91 percent of items), followed by risk of fire (81 percent), and other risks (75 percent). As panel C in table 2 shows, most of the inspection quality issues are type II errors, that is, noncompliant items that are evaluated by the inspector as compliant (19 percent of all audited items). Just 1 percent of discrepancies are type I error, that is, compliant items evaluated as noncompliant by the inspector. This suggests inspectors tend to be lenient in their assessment during the inspections, erring on the benefit of the businesses and rarely making mistakes that penalize them[7].

Balance and Attrition

At baseline, we find no systematic differences across our main outcome variables for which data are available (table A.4 in appendix A). We find some differences in covariates such as age and we include these as well as business characteristics at baseline (area and economic function) as controls in all our estimates. As Table A.5 in the Appendix shows, inspector attrition is balanced across treatment arms[8].

[6]To implement the randomization inference procedure, we use the "ritest" command in Stata with 500 replications(Heß, 2017).

[7]Other discrepancies were identified when comparing the assessments of the auditors and the inspectors, but they were inconsequential. For instance, inspectors marked as compliant items that did not apply to the firm under assessment, that is, there was no risk to be assessed, and since it was marked as compliant, it did not change the result of the inspection. Therefore, we focus on the two most important cases.

[8]By the end of the intervention, 16 out of the 21 inspectors remained active. Inspectors are considered inactive when they have conducted fewer than two inspections over the previous month. This is usually a result of them being reassigned to other activities.

Inspection Outcomes

Table 3 presents one of the most important highlights of the intervention: a significant decrease in the proportion of firms that pass the inspection. As column 1 shows, the high-intensity monitoring decreases the proportion of firms that pass the inspection by 19.5 percentage points (p-value < 0.001, q-value $= 0.001$), or 24 percent, compared to 80 percent in the low-intensity group. Businesses either fail because they have at least one item that is noncompliant, are asked to make corrections related to minor items and fail to implement them correctly, or their inspections are suspended on repeated attempts.

Column 2 shows that the intervention had no significant impact on the number of items evaluated, a measure of inspector thoroughness.

Inspection Quality

As column 3 (table 3) shows, the intervention has a negative impact on inspector quality, measured by type I errors, which increase by 2.3 percentage points (p-value $= 0.006$, q-value $= 0.013$) or 147 percent compared to their control counterparts. These errors correspond to items evaluated by the auditor as compliant, but evaluated as noncompliant by the inspector. Column 4 shows a decrease in type II errors (noncompliant items evaluated as compliant by the inspector), which means that the treatment leads to more agreement in terms of noncompliant items between the inspector and the auditor, but these impacts are not statistically significant. All in all, although the impacts of the intervention on those two types of errors go in the opposite direction, one reducing inspection quality (increases on type I error), and the other increasing it (reductions on type II error), both imply that inspectors are marking more items as noncompliant. This will in turn affect the result of the inspection, particularly if these errors are more likely to happen on major-risk items, since a single noncompliance in these items is enough for a business to fail the inspection.

Table 4 (column 1) shows that high-intensity inspectors increase significantly type I errors for major- and minor-risk items by 2 percentage points (p-value $= 0.004$), and 2.5 percentage points (p-value $= 0.000$), accordingly, compared to their counterparts in the low-intensity group. These imply increases in type I errors of 198 percent for major-risk items and of 125 percent for minor-risk items versus inspectors in the low-intensity group. As discussed previously, failing to comply with only one item leads to failing an inspection, as well as failing to provide appropriate evidence of corrections in minor items within 3 days. Therefore, type I errors in both types of items help shed light on the mechanisms explaining the large drop in firms passing the inspection, which is largely explained by a decrease in inspection quality. Column 2 shows that the intervention decreases type II error for both major- and minor-risk items, increasing items marked as noncompliant during the inspection, and therefore more firms failing the inspection, but these impacts are measured with less precision.

It is worth highlighting that these results might represent lower-bound impacts on quality due to the fact that we selected among the least discretionary items to maximize the interpretation of the discrepancies between the inspectors and auditors as errors, as opposed to discretion allowed by the regulation. These results link directly to our hypotheses that inspectors' loss- or risk-aversion to being found marking items compliant when they are not, could explain these results,

given that this could be perceived as a graver fault than the opposite.

Impacts on Corruption

Colum 5 in table 3 shows the intervention has a small impact on an index of corruption, which increases by 0.13 standard deviations (q value = 0.053) relative to the low-intensity group[9]. Table A.12 in the appendix shows that this index is constructed using two indicators: i) an indicator of whether the inspector suggested the business owner to hire a consultant during the inspection (this measure reports an increase of 0.9 percentage points (p-value = 0.516), compared to 4.7 percent in control), and ii) an indicator of whether the business owners report knowing any other businesses that were asked for bribes (this measure increases by 3.0 percentage points in treatment (p-value = 0.05), compared to 4.4 percent in the control group). Even though our proxies for corruption have limitations, we find a low frequency of reports of corruption from different sources we use to triangulate these data[10].

(INSERT TABLE4 HERE)

Heterogeneity

Given the literature documenting the role of behavioral biases on inspector assessment (e.g., Bazerman et al. (2002), 1997; Ibanez and Toffel (2020)), we explored if the number of faulty items detected in an inspection could affect the inspector's evaluation in the next inspection (during the same day). Table A.10 shows that this is indeed the case, and that the effect is considerable: an additional noncompliant item in an inspection is associated with a reduction of 1.2 percentage points in the probability of passing the next inspection, and with an additional 0.12 faulty items in the next inspection (recall that one faulty item is enough for the business to fail the inspection). To account for the fact that some businesses are geographically clustered and share some characteristics, we control for location fixed effects and cluster the standard errors at the level. However, table A.11 shows that treatment had no effect on this relation, with small and not significant point estimates.

Impacts on Post-inspection Investment

Table 5 (column 1) shows that the intervention increased the share of businesses conducting investments to comply with the regulation by 8 percentage points (p-value = 0.03). Column 2 shows that treatment increased investment on average by US$16.7 (p-value = 0.026), or 51 percent compared to the control group. Conditional on positive investment, average investment is US$125. Similarly, the intervention increased by 6 percentage points the share of firms that invested any time in doing improvements (p-value = 0.048). On average, the intervention increased

[9]Our index of corruption is standardized with respect to the control so that the mean of the index in the control group equals zero and its standard deviation equals one. Therefore, the impacts could be interpreted in standard deviations of the control group.

[10]Another measure of corruption from our endline survey, whether the business representative was asked for a bribe by the inspector, has almost no variation (2 out of 1,001 responses were affirmative) and was therefore excluded from the analysis. In addition, both the verification and endline surveys included open-ended questions about how to improve the inspection process. In these, we coded very few instances of mentions of corruption: 1 out of 2523 respondents of the verification survey, and 4 out of 1,001 respondents of the endline survey, mention corruption.

the amount of time invested to implement these improvements by 0.32 days (p-value= 0.008), compared to 0.81 days in the control group (column 4). The intervention did not affect the share of businesses hiring a consultant or the average pay for consultants (columns 5 and 6, respectively).

Impacts on Business Satisfaction with the Inspection Process and Regulation

We assess how the intervention impacted on the satisfaction of business representatives across multiple dimensions, including inspector's overall performance, inspector's professionalism, inspector's impartiality, clarity of the regulation, and the possibility of reporting problems encountered during the inspection process. Overall, we find that the intervention reduced business satisfaction across all measured outcomes, reflecting business negative perceptions and experiences due to the intervention.

Table 6 reports that treatment reduced the share of businesses that report being satisfied or very satisfied with inspector's performance by 12.8 percentage points (p-value = 0.004, q-value = 0.009), compared to 91.9 percent of respondents in the control group. Similarly, the intervention reduced businesses rating inspector professionalism as "good" or "very good" by 10.2 percentage points (p-value = 0.016, q-value = 0.011), inspector's impartiality by 13.4 percentage points (p-value $<$ 0.001, q-value = 0.001), the clarity of the regulation by 12.2 percentage points (p-value = 0.006, q-value = 0.009), and the possibility to report problems encountered during the inspection by 6.2 percentage points (p-value = 0.208, q-value = 0.044)[11]. It comes as no surprise that, if higher audit intensity increased the proportion of businesses that fail the inspection by increasing the number of items incorrectly evaluated as noncompliant, businesses had a poorer valuation of the inspector and the regulation process.

(INSERT TABLE6 HERE)

Additional Analyses

Dynamic Treatment Effects

We explore whether impacts are dynamic estimating differences by bimonthly periods, including the first two months excluded from previous analyses. As Figure 3 below and Table A.8 in the Appendix show, the intervention took some time to start producing impacts: during the first two months, impacts were smaller (in absolute value), and less precisely estimated although in the same direction than impacts for the next two periods. In particular, the proportion of firms passing an inspection decreases by 4.9 percentage points (p-value = 0.084) versus control, compared to decreases by 15.4 percentage points (p-value = 0.001) and 17.2 percentage points (p-value $<$ 0.001) for months 3-4 and 5-6, respectively. Similarly, impacts on Type I error are increasing and only statistically significant for the two last periods, although we lack the power to detect small differences such as this due to the sample split in three periods. Impacts on Type II errors are always negative and decreased in the last period, but they are not precisely estimated. These dynamic treatment effects could be partly explained by a slow start of the intervention during

[11] A sharpened q-value may be smaller than its corresponding p-value if there are several other significant q-values (Anderson (2008)).

the first two months, due to multiple exogenous factors including COVID and end-of-the-year holidays, as well as to delays in the initial feedback reports. However, this reflects issues likely to occur in a scaled-up version of the program. These results are consistent with dynamics in the response to accountability regimes documented recently in the literature (**?**).

Robustness Checks

One concern about the stability of our results stems from the small number of clusters (inspectors) in our sample. As Figures A.1 to A.3 in the Appendix show, our results are stable when each of the clusters (inspector) is excluded from the estimates, that is, using a leave-one-out estimator, following Young (2019). The impact on the proportion of firms that passed the inspection is negative, statistically significant, and very stable around 19 percentage points (Figure A1). Similarly, the impact on Type I Error discrepancies is positive, statistically significant, and relatively stable around 2 percentage points in all cases (Figure A2), and the impact on Type II error is negative in all cases, but not precisely estimated, as when all clusters are included (Figure A3). These robustness checks provide evidence that our results are not driven by one or a few inspectors.

We provide additional robustness check estimates including excluding an auditor with suspected quality issues (Table A13), and including the last week when tablets were not used by inspectors anymore (Table A14) and all results remain qualitatively the same, and so do their patterns of statistical significance.

Challenges of and Lessons from Implementing the Intervention

In this section, we provide a brief overview of the main challenges and lessons learned in the process of implementing the monitoring scheme evaluated, aiming to contribute to the discussion on how to leverage accountability mechanisms and the potential of digital government in low- and middle-income countries. Challenges were of four broad types: (a) challenges working with public sector partners from different levels of government, (b) challenges implementing an intervention with an ICT component, (c) general methodological and logistical challenges, and (d) logistical and technical challenges implementing the monitoring scheme. One of the main difficulties of conducting a study on business safety inspections in this setting was the complex institutional arrangement surrounding them, with multiple and revolving state actors across different levels of government (national and subnational). When the intervention was in its planning stage, the inspection regulation was in the process of being changed. This not only delayed the implementation's start date but also affected the composition of the project's institutional partners, as oversight of business safety inspections was changed from the National Center for Estimation, Prevention, and Reduction of Disaster Risk to the Ministry of Housing, Construction and Sanitation (MVCS for its acronym in Spanish). In addition, once enacted, the new regulation had a transition period which coincided with municipal elections and the entry of new municipal administrations in all local governments across the country. Once installed, municipal teams faced great instability; one municipality had a total of five managers responsible for business safety inspections during the planning and intervention period, which made it impossible to conduct the study with them despite the existence of a memorandum of understanding with the institu-

tion. Another municipality faced severe logistical and staffing limitations that prevented them from implementing the procedures for participating in the study. For example, they were unable to maintain a daily inspection plan with basic information on requests and visit plans (which were needed to coordinate the verifications). Overall, the difficulty of working with multiple partners across different levels of government, each with different roles in the inspection process and different incentives, introduced uncertainty to municipalities' participation and continued commitment to the study.

A second difficulty stemmed from using an ICT component as part of the intervention. The research team initially underestimated both the complexity of the technological infrastructure required to incorporate ICT tools in certain day-to-day activities of inspectors and the challenges for municipalities and inspectors to make the transition from paper to ICT-based forms (see Barron et al. (2022)). One challenge was related to authorizations to use ICT tools. Although the majority of businesses authorized the municipality to share the inspection report electronically, some businesses preferred paper inspections for a variety of reasons (for example, unfamiliarity with electronic processes, lack of an email account to receive electronic reports). Another challenge was that the ICT tools presented small issues and bugs, such as signatures not properly uploading to reports, although this happened in less than 5 percent of reports. Ultimately, the research team had to adapt and collect inspection data both on paper and electronically, which involved implementing additional processes to scan and digitize paper inspection forms. These difficulties, combined with the institutional complexities described in the previous paragraph, meant that the study could only be implemented in one municipality out of seven that had been initially recruited for the intervention. Methodological and logistical challenges were also important.

The main methodological challenge was related to the inclusion of ex-ante inspections, the type of inspection conducted for high- and very-high risk businesses, which require teams of multiple inspectors. In addition, given their complexity, these inspections are arguably a more significant barrier for businesses obtaining their operating license or renewing their safety certificate, although they represent only 8 percent of inspections in the district of Lima. Including ex-ante inspections was challenging from a technological perspective as well, as it required merging electronic forms coming from different inspectors, a process that was affected by connectivity issues and the technical skills of inspectors. However, an even larger barrier was methodological, as working with ex-ante inspections required using a cluster design, where inspectors were assigned to treatment and control groups based on fixed clusters of three inspectors. The problem was that it was difficult for municipalities to keep these clusters fixed because of staff rotation and the municipality's internal programming requirements. This issue was further compounded by the onset of the COVID-19 pandemic. Peru was hit extremely hard (it is still the country with the highest COVID-19 death rate in the world), and the government implemented multiple lockdowns that interrupted inspections for about a year. Once inspections resumed their activities, COVID-19 continued to affect the project in multiple ways, with inspectors getting sick, or being suspended from the municipality for refusing to get vaccinated. Overall, however, the resilience of inspectors and municipalities to continue their work despite these challenging conditions has been one of the greatest lessons of this intervention.

A fourth challenge was related to the implementation of the monitoring scheme. Some inspectors were opposed to their inspections being monitored due to multiple factors, including methodological disagreements (inspectors wanted to see evidence of all their discrepancies with the independent auditors) and fear that their reports would be part of their personal files and be shared with the municipality and MVCS. Some inspectors went as far as asking firms to prohibit the entry of auditors to their premises. These issues were ultimately resolved by requesting the intervention of the municipality and stressing the message that inspector reports were personal and confidential and that individual results were therefore not shared with the municipality or the MVCS.

On the side of the third-party auditors, a major challenge involved one auditor not following the study protocols and possibly submitting low-quality data, which led to his removal from the team and the research team strengthening quality controls. Another challenge was the difficulty in obtaining consent from firms to participate in the verification process. Because auditors were hired by a third-party firm, they did not have municipal credentials, and some firms were wary of the purpose of their visit. To address this issue, auditors were given an introductory letter with contact numbers from the municipality that firms could call. While 11 percent of firms refused to participate, this led to a high verification success rate.

10 Conclusions

We present results of a field experiment assessing the impact of intensive independent monitoring of inspectors on business inspection results, inspection quality, corruption, and business outcomes including investments and satisfaction.

In our experiment, we randomly assigned inspectors of low- and medium-risk businesses into either a high-intensity monitoring group where we independently verified a random sample of around 70 percent of their inspections, or a low-intensity group, where we verified a random sample of around 20 percent of their inspections. The intervention includes periodic reports with feedback on the accuracy of the inspections to inspectors and the municipality, but the municipality only receives average aggregate performance and the names of the three inspectors with the lowest discrepancy rates measured as the differences between the inspectors' evaluation and a third-party auditor across a select group of items verified within two days of the inspection. There were no consequences for poor performance (high error rates) for inspectors. The experiment sheds light on mechanisms to improve the performance of the regulator. It does not assess whether performance is linked to improvements in welfare (in this case, a reduction in incidents of fire, collapse, or related safety hazards). This is an interesting topic for further research.

All in all, our results suggest that high levels of monitoring of inspectors in a highly discretionary environment with misaligned incentives can create perverse incentives that led to undesired outcomes. The intervention decreases considerably the proportion of businesses that pass the inspection (61 percent in treatment, compared to 80 percent in control that passes the inspection). However, this decrease is mostly due to an increase in compliant items that the inspector wrongly marks as noncompliant (a reduction in inspection quality), which we argue is related to risk-aversion from inspectors compounded by a discretionary regulation. It is worth highlighting that our impacts on quality might represent lower-bound impacts since the items

subject to verification were selected among the least discretionary to maximize the interpretation of the discrepancies between the inspectors and auditors as errors, as opposed to discretion allowed by the regulation. We also find a small increase in our index of corruption, a significant increase in post-inspection investments, and reductions in business satisfaction with inspector's performance, the inspection process, and the regulation.

There were multiple challenges of implementing this type of intervention, which can shed light on important lessons for governments looking into implementing similar mechanisms to improve accountability. First, ICTs can be a powerful tool to assist in data recording for monitoring, but their implementation must be carefully planned to include adequate training and sufficient resources to ensure their strength and sustainability. Second, monitoring mechanisms themselves must be carefully planned to ensure that adequate incentives are in place to obtain high-quality and reliable records systematically. Third, working in regulatory contexts that involve stakeholders with different roles and from different levels of government requires careful planning and an adequate intervention time frame to provide enough flexibility to adapt to changes.

A key policy implication is that in a context of high discretion allowed by the regulation and misaligned incentives for inspectors, intense monitoring seems to have backfired, likely making inspectors excessively cautious of failing to detect faults in an inspection. In addition, inadequate regulation, with high levels of discretion and inadequate implementation guidelines can potentially compound misaligned incentives, creating additional costs and inefficiencies for businesses and the government.

References

Anderson, M. 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.

Avis, E., Ferraz, C., and Finan, F. (2018). Do government audits reduce corruption? estimating the impacts of exposing corrupt politicians. *Journal of Political Economy*, 126(5):1912–1964.

Barron, M., Bedoya, G., García Montúfar, D., Gertler, P., and Goicoechea, A. (2022). The impact of electronic recording of business inspections on compliance with regulatory standards and inspectors' performance.

Barron, M., Bedoya Arguelles, G., Garcia Montufar Garcia, D., and Goicoechea, A. (2020). Electronic system for business safety inspections in peru.

Bazerman, M. H., Loewenstein, G., Moore, D. A., et al. (2002). Why good accountants do bad audits. *Harvard business review*, 80(11):96–103.

Bolton, P., Freixas, X., and Shapiro, J. (2012). The credit ratings game. *The Journal of Finance*, 67(1):85–111.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232.

Celhay, P. A., Gertler, P., Olivares, M., and Undurraga, R. (2024). How managers can use purchaser performance information to improve procurement efficiency. Technical report, National Bureau of Economic Research.

Dal Bó, E., Finan, F., Li, N. Y., and Schechter, L. (2021). Information technology and government decentralization: Experimental evidence from paraguay. *Econometrica*, 89(2):677–701.

Duflo, E., Greenstone, M., Pande, R., and Ryan, N. (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.

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

Duflo, E., Hanna, R., and Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American economic review*, 102(4):1241–1278.

Eftekhari, M. H., Parsapoor, A., Ahmadi, A., Yavari, N., Larijani, B., and Gooshki, E. S. (2023). Exploring defensive medicine: examples, underlying and contextual factors, and potential strategies-a qualitative study. *BMC Medical Ethics*, 24(1):82.

Franco-Bedoya, S. and Mani, M. (2020). The drivers of firms' compliance to environmental regulations: The case of india.

Gonzalez-Lira, A. and Mobarak, A. M. (2021). Slippery fish: Enforcing regulation when agents learn and adapt. Technical report, National Bureau of Economic Research.

Griffin, J. M. and Tang, D. Y. (2011). Did credit rating agencies make unbiased assumptions on cdos? *American Economic Review*, 101(3):125–130.

Gupta, S., Saksena, S., and Baris, O. F. (2019). Environmental enforcement and compliance in developing countries: Evidence from india. *World Development*, 117:313–327.

Hatanaka, M., Bain, C., and Busch, L. (2005). Third-party certification in the global agrifood system. *Food Policy*, 30(3):354–369.

Heß, S. (2017). Randomization inference with stata: A guide and software. *The Stata Journal*, 17(3):630–651.

Ho, D. E. and Sherman, S. (2017). Managing street-level arbitrariness: The evidence base for public sector quality improvement. *Annual Review of Law and Social Science*, 13:251–272.

Ibanez, M. R. and Toffel, M. W. (2020). How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management science*, 66(6):2396–2416.

Johnson, M. S., Levine, D. I., and Toffel, M. W. (2023). Improving regulatory effectiveness through better targeting: Evidence from osha. *American Economic Journal: Applied Economics*, 15(4):30–67.

Kakemam, E., Arab-Zozani, M., Raeissi, P., and Albelbeisi, A. H. (2022). The occurrence, types, reasons, and mitigation strategies of defensive medicine among physicians: a scoping review. *BMC Health Services Research*, 22(1):800.

Molfetas, A. (2019). Business licensing reforms: Insights from selected country experiences.

Nahed, B. V., Babu, M. A., Smith, T. R., and Heary, R. F. (2012). Malpractice liability and defensive medicine: a national survey of neurosurgeons. *PloS one*, 7(6):e39237.

Potoski, M. and Prakash, A. (2005). Green clubs and voluntary governance: Iso 14001 and firms' regulatory compliance. *American journal of political science*, 49(2):235–248.

Ronen, J. (2010). Corporate audits and how to fix them. *Journal of Economic Perspectives*, 24(2):189–210.

Short, J. L., Toffel, M. W., and Hugill, A. R. (2016). Monitoring global supply chains. *Strategic Management Journal*, 37(9):1878–1897.

Strobl, G. and Xia, H. (2012). The issuer-pays rating model and ratings inflation: Evidence from corporate credit ratings. *Unpublished working paper*.

Vadillo, J. (2021). Mesa redonda: A 20 años del incendio que cobró la vida de más de 400 personas. *Andina*.

White, L. J. (2010). Markets: The credit rating agencies. *Journal of Economic Perspectives*, 24(2):211–226.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, 134(2):557–598.

Zamboni, Y. and Litschig, S. (2018). Audit risk and rent extraction: Evidence from a randomized evaluation in brazil. *Journal of Development Economics*, 134:133–149.

Table 1: Baseline Characteristics

	Mean (SD)
A. Inspector level	
Age (years)	47.857
	(7.371)
Experience as inspector (years)	10.571
	(6.516)
Inspector is a woman	0.286
	(0.463)
Inspector's highest education level	
Bachelor's degree	0.800
	(0.414)
Master's degree and above	0.200
	(0.414)
Inspector has relevant specialization	0.571
	(0.507)
Inspector has a smartphone	0.933
	(0.258)
Number of inspectors	21
B. Establishment level	
Proportion passed	0.766
	(0.424)
Proportion of establishments by function	
Commerce	0.942
	(0.234)
Office/admin service	0.039
	(0.194)
Area (m^2)	41.431
	(127.459)
Number of establishments	2,091
Number of inspectors	21
C. Verification level	
Type I error (compliant item evaluated as non-compliant by inspector)	0.010 (0.097)
Type II error (noncompliant item evaluated as compliant by inspector)	0.189 (0.392)
Number of items verified	2,200
Number of inspectors	19
Number of establishments	149

Notes. This table reports baseline means for inspector, establishment, and verification characteristics, and (standard deviations) in parenthesis below the means. In Panel B, we used data from electronic administrative records provided by the Municipality of Lima covering the period from 11/12/2021 to 12/14/2021. For two inspectors that were sick during this period, we used data from the same source covering the period from 10/15/2021 to 11/11/2021. In Panel C, between 1 and 18 establishments were verified per inspector (mean=7.7). For 3 inspectors, no establishments were verified during baseline due to sickness, absence, or impediments during verifications. SD = standard deviation.

Table 2: Balance Checks

	At randomization			At intervention end		
	Control mean (SD)	Treatment mean (SD)	Adjusted difference (p-value)	Control mean (SD)	Treatment mean (SD)	Adjusted difference (p-value)
	(1)	(2)	(3)	(4)	(5)	(6)
A. Inspector level						
Age (years)	44.900 (6.064)	50.545 (7.673)	5.649* [0.088]	44.000 (5.679)	49.600 (7.382)	5.577* [0.095]
Experience as inspector (years)	9.400 (7.090)	11.636 (6.087)	2.213 [0.464]	8.222 (6.399)	12.300 (5.982)	4.183 [0.165]
Inspector is a woman	0.200 (0.422)	0.364 (0.505)	0.160 [0.456]	0.222 (0.441)	0.300 (0.483)	0.077 [0.731]
Master's degree and above	0.143 (0.378)	0.250 (0.463)	0.090 [0.677]	0.143 (0.378)	0.286 (0.488)	0.097 [0.678]
Inspector has relevant specialization	0.600 (0.516)	0.545 (0.522)	0.000 [.]	0.556 (0.527)	0.600 (0.516)	−0.000 [.]
Inspector has a smartphone	1.000 (0.000)	0.875 (0.354)	−0.119 [0.400]	1.000 (0.000)	0.857 (0.378)	−0.130 [0.401]
Personality indicators (Big Five)						
Extraversion (0-30)	23.429 (3.735)	21.750 (4.132)	−1.851 [0.356]	23.429 (3.735)	22.143 (4.298)	−1.721 [0.429]
Agreeableness (0-30)	26.286 (3.450)	23.750 (3.991)	−2.716 [0.154]	26.286 (3.450)	24.286 (3.988)	−2.429 [0.232]
Conscientiousness (0-30)	25.571 (3.359)	23.375 (5.655)	−2.463 [0.270]	25.571 (3.359)	23.857 (5.928)	−2.416 [0.320]
Negative emotionality (0-30)	12.000 (2.828)	13.125 (3.227)	1.254 [0.419]	12.000 (2.828)	12.714 (3.251)	1.013 [0.542]
Open-mindedness (0-30)	19.571 (2.299)	19.500 (1.773)	−0.090 [0.936]	19.571 (2.299)	19.571 (1.902)	−0.039 [0.974]
Number of inspectors		21			19	
F-test statistic		0.77			0.69	
p-value		0.665			0.714	
B. Establishment level						
Proportion passed	0.742 (0.438)	0.793 (0.405)	0.058 0.289	0.718 (0.450)	0.791 (0.407)	0.070 [0.220]
Proportion of establishments by function						
Commerce	0.947 (0.224)	0.935 (0.246)	−0.011 0.489	0.948 (0.223)	0.932 (0.252)	−0.017 [0.306]
Office/admin service	0.030 (0.172)	0.049 (0.216)	0.017* 0.083	0.032 (0.176)	0.049 (0.216)	0.018* [0.095]
Area (m^2)	33.313 (122.837)	50.715 (131.998)	17.630 0.184	34.243 (129.072)	54.376 (140.293)	20.105 [0.164]
Number of establishments		2,091			1,834	
Number of inspectors		21			19	
F-test statistic		2.06			1.89	
p-value		0.124			0.156	
C. Verification level						
Type I error (compliant item evaluated as noncompliant by inspector) Type I error (compliant item evaluated as noncompliant by inspector)	0.005 (0.068)	0.014 (0.118)	0.009 [0.121]	0.005 (0.071)	0.015 (0.122)	0.010 [0.130]
Type II error (noncompliant item evaluated as compliant by inspector) Type II error (noncompliant item evaluated as compliant by inspector)	0.203 (0.402)	0.176 (0.381)	−0.028 [0.178]	0.201 (0.401)	0.175 (0.380)	−0.025 [0.292]
Number of items verified		2,200			1,931	
Number of establishments		149			131	
Number of inspectors		19			17	

Table 3: Impacts on Inspection Results, Inspection Quality and Corruption

	Inspection results		Inspection quality		Corruption
	Passed	Natural log of number of items evaluated	Type I error (compliant item evaluated as noncompliant by inspector)	Type II error (noncompliant item evaluated as compliant by inspector)	Corruption index
	(1)	(2)	(3)	(4)	(5)
Treatment	−0.195***	0.021	0.023**	−0.011	0.134*
SE	(0.024)	(0.013)	(0.007)	(0.010)	(0.056)
p-value	[0.000]	[0.230]	[0.006]	[0.460]	[0.050]
q-value	{0.001}	{0.130}	{0.013}	{0.226}	{0.053}
Number of:					
Establishments	3,617	3,019	1,958	1,958	1,001
Inspectors	21	21	21	21	20
Items verified	n.a.	n.a.	36,338	36,338	n.a.
Control Mean	0.798	3.337	0.016	0.110	0.000
Impact (%, SD)	(−24%; −0.485)	(1%; 0.193)	(147%; 0.185)	(−10%; −0.034)	(n.a.; 0.134)

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level based on the FDR-adjusted q-values including the 5 main outcomes in this table. This table reports Ordinary Least Squares estimates of treatment effects. For each outcome variable, we report the estimated coefficient of interest and its (robust standard error) in parentheses, clustered at the inspector level. We report randomization inference [p-values] in brackets, and FDR-adjusted {q-values} for the main outcomes in braces. All regressions include randomization strata, inspectors' age and experience, area, and the function of the inspected establishment as controls. Regressions of inspection results and inspection quality outcomes also control for the proportion of passing firms at last visit at baseline by inspector (based on administrative records). Regressions of inspection quality outcomes include verifier fixed effects. For Type I and II Error, we assume the null hypothesis is whether the item is noncompliant. Errors are estimated as the discrepancies in assessment between an independent reviewer conducting the verification and the inspector. The outcomes are described in detail in the appendix. Excludes data from the first two months of the intervention. Number of establishments corresponds to the number of businesses visited by an inspector (column 1), visited and inspected (column 2), verified (columns 3 and 4), and surveyed at endline (column 5). The difference in the number of observations between columns 1 and 2 are suspended or cancelled inspections, which imply that an inspector visited the premises but did not conduct an inspection as part of the process. n.a. = not applicable. SE = standard error.

Table 4: Impacts on Inspection Quality: Interacted With Item Relevance

	Type I error (compliant item evaluated as noncompliant by inspector) (1)	Type II error (noncompliant item evaluated as compliant by inspector) (2)
Treatment	0.025***	−0.013
SE	(0.010)	(0.016)
p-value	[0.000]	[0.282]
Relevant item × Treatment	−0.005	0.007
SE	(0.008)	(0.018)
p-value	[0.998]	[0.156]
Relevant item	−0.010	−0.140**
SE	(0.007)	(0.014)
p-value	[1.000]	[0.012]
Number of:		
Establishments	1,958	1,958
Inspectors	21	21
Items verified	36,338	36,338
Control mean		
Nonelevant item	0.020	0.176
Relevant item	0.010	0.036
Impact Nonelevant item		
%	125%	-7%
SD	0.18	-0.03
Impact Relevant item		
%	198%	-17%
SD	0.20	-0.03
Test Treatment + Treatment × Relevant item = 0 p-value	0.004	0.465

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level based on the randomization inference p-values. This table reports Ordinary Least Squares estimates of treatment effects. For each outcome variable, we report the estimated coefficient of interest and its (robust standard error) in parentheses, clustered at the inspector level. We report randomization inference [p-values] in brackets. p-value for the test of equality is the naive p-value. All regressions include randomization strata, inspectors' age and experience, area, the function of the inspected establishment, and the proportion of passing firms at last visit at baseline by inspector (based on administrative records) as controls, as well as verifier fixed effects. For Type I and II Error, we assume the null hypothesis is whether the item is noncompliant. Errors are estimated as the discrepancies in assessment between an independent reviewer conducting the verification and the inspector. The outcomes are described in detail in the appendix. Excludes data from the first two months of the intervention. SE = standard error; SD = standard deviation.

Table 5: Impacts on Post-Inspection Investment

	Any investment (1)	Investment in USD (2)	Any time (3)	Time in days (4)	Consultant hired (5)	Consultant payment in USD (6)
Treatment	0.076*	16.7*	0.062*	0.321*	0.009	39.2
SE	(0.029)	(6.2)	(0.029)	(0.127)	(0.013)	(38.3)
p-value	[0.030]	[0.026]	[0.048]	[0.008]	[0.516]	[0.458]
q-value	{0.053}	{0.053}	{0.064}	{0.051}	{0.208}	{0.208}
Number of:						
Establishments	974	974	980	980	1,001	949
Inspectors	20	20	20	20	20	20
Impact (%, SD)	(29%; 0.173)	(51%; 0.189)	(23%; 0.139)	(39%; 0.174)	(18%; 0.040)	(19%; 0.108)
Control mean						
Overall	0.263	32.8	0.273	0.814	0.047	211
Overall conditional on investment > 0	1.0	124.9	1.0	3.0	1.0	291.6

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level based on the FDR-adjusted q-values for the outcomes in this table. This table reports Ordinary Least Squares estimates of treatment effects. For each outcome variable, we report the estimated coefficient of interest and its (robust standard error) in parentheses, clustered at the inspector level. We report randomization inference [p-values] in brackets, and FDR-adjusted {q-values} calculated for the outcomes within the table in braces. All regressions include randomization strata, inspectors' age and experience, area, and the function of the inspected establishment as controls. The outcomes are described in detail in the appendix. Excludes data from the first two months of the intervention. SE = standard error; SD = standard deviation.

Table 6: Impacts on Satisfaction

	Proportion of respondents rating indicator as:				
	Satisfied or very satisfied	Good or very good			
	Overall inspector rating (1)	Inspector's professionalism (2)	Inspector's impartiality (3)	Clarity of regulation (4)	Possibility to report problems (5)
Treatment	−0.128***	−0.102**	−0.134***	−0.122***	−0.062**
SE	(0.031)	(0.033)	(0.032)	(0.030)	(0.033)
p-value	[0.004]	[0.016]	[0.000]	[0.006]	[0.208]
q-value	{0.009}	{0.011}	{0.001}	{0.009}	{0.044}
Number of:					
Establishments	997	1,001	1,001	1,001	1,001
Inspectors	20	20	20	20	20
Control mean	0.919	0.826	0.818	0.768	0.604
Impact (%, SD)	(−14%; −0.466)	(−12%; −0.269)	(−16%; −0.346)	(−16%; −0.290)	(−10%; −0.127)

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level based on the FDR-adjusted q-values for the outcomes in this table. This table reports Ordinary Least Squares estimates of treatment effects. For each outcome variable, we report the estimated coefficient of interest and its (robust standard error) in parentheses, clustered at the inspector level. We report randomization inference [p-values] in brackets, and FDR-adjusted {q-values} calculated for the outcomes within the table in braces. All regressions include randomization strata, inspectors' age and experience, area, and the function of the inspected establishment as controls. The outcomes are described in detail in the appendix. Excludes data from the first two months of the intervention. SE = standard error; SD = standard deviation.