Government Oversight and Inter-Institutional Legibility: Evidence from Colombia *

Natalia Garbiras-Díaz† Tara Slough‡

October 7, 2025

Abstract

Effective governance requires reliable information. Many types of administrative data are produced from decentralized entities' reports to the central government. The use of these data for policymaking or oversight creates incentives for decentralized bureaucrats to strategically (mis)report, reducing the quality of the resultant data. We conceptualize this challenge as one of *inter-institutional legibility*, shaped by principal-agent dynamics between central and decentralized governments. We study these dynamics experimentally in the production of Colombia's National Transparency Index by partnering with a national watchdog agency to assess the effect of increasing the salience of oversight on decentralized entities' reporting behavior. More salient oversight changes reported scores and may marginally increase reporting rates. Comparing self-reports to an independent audit, we show that selection, distortion, and noise in reports limit the legibility of these entities to the central government. Our findings underscore the challenge of administrative data collection and the use of these data.

*Thanks to Carolina Torreblanca for excellent research assistance. Special thanks to Carolina Bernal, Marco Castradori, Kirill Chmel, Benjamin Gelman, Anna Houk, Kyle Van Rensselaer, and Hanying Wei for replicating and extending this work. We are grateful to Dan Berliner, Saad Gulzar, Macartan Humphreys, Ian Turner, and audiences at EuroWEPS, the LSE-NYU Political Science and Political Economy Conference, ITAM, Universidad de los Andes, Columbia, Georgetown, Princeton and APSA for helpful comments. We thank Innovations for Poverty Action for their incredible work managing this project.

†Assistant Professor, Harvard Business School, ngarbirasdiaz@hbs.edu.

‡Assistant Professor, New York University, tara.slough@nyu.edu. Corresponding author.

Governance requires reliable information (Scott, 1998).[1] Governments use information to manage populations, allocate resources, and design policies (Garfias and Sellars, 2021; Callen et al., 2020). While government data sourced from citizens and households (e.g., censuses or vital records) are increasingly studied (Grajalez et al., 2013; Bowles, 2020; Lee and Zhang, 2016), another important source of government information is reports from decentralized entities—local governments or subsidiary national government agencies—to the central government.

We argue that this form of administrative data is produced within a principal-agent relationship between central and decentralized government entities. Specifically, central governments use these data to monitor the performance of decentralized agents. This reliance on self-reported data for policymaking and oversight introduces a fundamental challenge: the same agents responsible for generating the data may also have incentives to strategically misreport, ultimately undermining data quality. In decentralized settings, the quality of information provided by decentralized entities—what we term *inter-institutional legibility*—is therefore a central, yet understudied issue. We examine these dynamics through the production of Colombia's National Transparency Index (ITA), a measure of transparency practices across public entities. We provide empirical evidence consistent with our theoretical framework, showing that decentralized entities misreport and that their reporting behavior is sensitive to the degree of oversight exercised by the central government.

Our analysis proceeds in three steps. First, we develop a framework that links the reporting behavior of decentralized entities to inter-institutional legibility through three known pathologies of measurement: missingness, systematic measurement error, and non-systematic measurement error. Missingness arises when bureaucrats fail to submit data; systematic measurement error occurs when agents intentionally distort true (latent) measures; and non-systematic measurement error, or noise, results from a lack of effort. When decentralized agents anticipate that their responses may draw oversight attention from the central government with the possibility of enforcement, they may alter their reporting behavior to avoid this attention or potential punishment. Optimal report-

[1]The pre-analysis plan for the experiment in this manuscript is available at `https://osf.io/am3qu`.

ing behavior from the perspective of decentralized entities can thus introduce measurement error, ultimately limiting the quality of data and legibility of decentralized governments to the center.

Second, we test the idea that decentralized entities' reporting decisions are shaped by their perceptions of the oversight. We partner with the Office of the Attorney-Inspector General (PGN), a national-level watchdog agency that collects and compiles the ITA annually using self-reports submitted by all public sector entities in Colombia. To understand entities' sensitivity to oversight, the PGN randomly varied whether these entities received direct communication about their reporting obligation. We contrast this direct communication treatment condition with a status quo condition in which the PGN delegated all communication to other national agencies, none of which have watchdog mandates or use ITA data for enforcement functions. This variation allows us to assess how oversight salience affects both data submission and reported scores.

Our findings suggest that principal-agent dynamics may limit inter-institutional legibility: entities report lower (less desirable) levels of compliance with transparency practices when oversight is more salient. Using a novel decomposition that builds upon Lee (2009)'s bounding approach, we separate this effect into changes in the reporting behavior of: (1) entities that always report, regardless of their perceptions of PGN oversight, and (2) entities that report only because of direct oversight. On average, treated always-reporting entities report lower scores than their control counterparts. Entities that report due to direct communication report lower average scores than those that always report. While oversight salience increases the probability of reporting by 2.9 percentage points, this effect is not statistically significant at conventional thresholds.

Finally, after the intervention, and outside the scope of our partnership with the PGN, we also conduct an independent audit of a subset of components in the index to approximate a true latent measure of transparency practices at the entity level. We compare reported practices to the audit-based measure to assess the extent and nature of misreporting. This design allows us to learn how decentralized bureaucrats' anticipation of oversight conditions the data they submit, and to describe how these reports relate to true levels of transparency practices.

Our audit reveals substantial variation in bureaucratic reporting behavior. True transparency practices correlate with three key measurement pathologies in the data: (1) high-performing entities disproportionately select into reporting, (2) low- and middle-performing entities systematically overreport their scores, and (3) low-performing entities exhibit greater variance in their reported scores. We further show that, contrary to common explanations of poor data quality, these patterns persist across all levels of state capacity, suggesting that strategic misreporting is a broader challenge of bureaucratic incentives. Together, our findings provide empirical evidence of strategic reporting by bureaucrats and suggest that central government reliance on data produced by decentralized entities can undermine the accuracy and observability of that very data.

Related literature. Our findings speak to several strands in the literature. First, we contribute to work on intergovernmental relations by focusing on the principal-agent problems in decentralized governance. While theories about the decision to decentralize governments or firms describe moral hazard problems associated with decentralized arrangements (Baliga and Sjöström, 1998; De Groot, 1988; Bardhan and Mookherjee, 2000), these problems are still present but less discussed when governments have chosen to decentralize authority. We analyze one strategy by which central government principals engage in the monitoring of decentralized entities: data collection on entity performance. As we show, this strategy is scalable (and widespread), but monitoring on the basis of self-reported data undermines the quality of information available to principals.

By conceptualizing national-subnational relations within a principal-agent framework, we prior work that models the decision to decentralize (De Groot, 1988; Baliga and Sjöström, 1998; Mookherjee, 2006; Tommasi and Weinschelbaum, 2007) to consider how principals monitor agents after decentralization has occurred. Given the wave of decentralization reforms in the 1990s and 2000s (Willis, da C. B. Garman, and Haggard, 1999; Faletti, 2005), we draw attention to the use of data collection to oversee decentralized governments under contemporary institutional arrangements.

Second, we contribute to the literature on state legibility. Existing accounts emphasize how governments collect data on individuals, households, and economic activity to enhance taxa-

tion, resource allocation, and administrative control (Scott, 1998; Lee and Zhang, 2017; Sánchez-Talanquer, 2020). We extend this logic to intergovernmental relations by introducing the concept of inter-institutional legibility and providing a framework to study the quality or legibility of data reported by bureaucrats in decentralized entities.

Third, we contribute to research on strategic misreporting. Most existing work focuses on autocratic regimes, where bureaucrats manipulate economic data to align with political incentives, often for career advancement or regime survival (Martínez, 2022; Trinh, 2021; Lorentzen, 2014; Wallace, 2016; Edmond, 2013). Recent studies document data manipulation in democratic settings, particularly in politically sensitive areas like crime reporting by US police departments (Eckhouse, 2022; Cook and Fortunato, 2022). We expand these scope conditions by showing that even routine government reporting is subject to manipulation. Our findings challenge the assumption that high-quality data is an inherent feature of democracies and demonstrate that incentives to misreport extend beyond politically sensitive measures.

Finally, we make a methodological contribution to the study of selection. In our experimental design, we derive new bounds (building on Lee 2009) to decompose changes in reported data into two sources: selection into reporting and changes in reporting behavior among "always reporters." Given the prevalence of post-treatment selection in experimental and quasi-experimental research designs (Slough, 2023), our approach offers broad applications for scholars studying data quality, administrative records, and governance oversight mechanisms.

1 Theoretical Framework

1.1 National-Subnational Government Relations as a Principal-Agent Problem

We posit that the relationship between a (unitary) central government and a decentralized public entity—such as a municipal government—can be seen through the lens of a principal-agent

problem.[2] Following canonical characterizations of principal-agent problems (e.g., Jensen and Meckling, 1976), the principal—here, the central government—delegates the provision of public goods and services to agents—here, decentralized entities. The interests of the central government and decentralized entities often diverge: local needs may differ from national priorities (e.g., Oates, 1999), and in democracies, leaders at each level of government are typically elected independently by distinct electorates (Tommasi and Weinschelbaum, 2007). Local officials may prefer to deliver different service bundles or follow governance standards that deviate from those set by the center. Indeed, a large literature theorizes or documents corruption or collusion by local governments as a possible cost of decentralization (Bardhan and Mookherjee, 2000; Fan, Lin, and Treisman, 2009).

Within this principal-agent relationship, the central government seeks information—here, administrative data—about local government outputs and performance in order to inform policy decisions and, crucially, to monitor agent behavior. There are different ways to collect such information. The central government might send its own personnel to gather data, for example through audits of local accounts (e.g., Ferraz and Finan, 2008). Alternatively, it can request information directly from the agent—i.e., the local government. For example, Cook and Fortunato (2022) document the use of police-produced data by state legislatures (the principal) to oversee local police departments (agents). Relying on the agent to self-report performance data is generally cheaper and more scalable, but when such data is used for monitoring and enforcement, it creates incentives for distortion. We refer to the resulting quality of this information as the *inter-institutional legibility* of decentralized entities to the central government.

1.2 Monitoring on the basis of self-reported information

Consider a setting in which the central government (principal) wants to assess a decentralized entity's (agent's) performance with respect to a policy or regulation. This could be the use of intergovernmental transfers, the provisions of public services, or compliance with transparency

[2]Whether the relationship between a federal government and federal entities is as a principal-agent relationship remains an open question, as reflected in U.S. case law (e.g., *South Dakota v. Dole*, 483 U.S. 203 [1987]). We do not take a position on this question, but argue the strategic dynamics we describe are present in many federal systems.

regulations etc. Represent the agent's performance or compliance as $\theta \in \mathbb{R}$, which is known to the decentralized entity (or a bureaucrat within the entity). We will assume that θ is sticky, at least in the short run. This assumption reflects the idea that changing budgetary execution, the provision or allocation of services, or cleaning up local government is costly and affecting such changes requires the sustained effort of multiple actors within the local government.

When the central government requests information, a bureaucrat within the entity determines (a) whether and (b) what to report. We denote their report as $r \in \{\emptyset, \mathbb{R}\}$, defined by two choices: how much effort to exert and whether to intentionally misreport θ. Effort, $e \geq 0$, first determines whether any information is reported: if a bureaucrat exerts no effort, no report is made and $r = \emptyset$. In this case, the central government observes that no information was submitted. Second, if the bureaucrat exerts effort, they make a report, $r \in \mathbb{R}$. The more effort they exert, the more accurate the report. Formally, greater effort reduces the extent of idiosyncratic error, ε which is drawn from a mean-zero random variable with variance $\sigma^2(e)$ where $\sigma^2(e)$ is strictly decreasing in e.

In addition to exerting effort, the bureaucrat can choose whether to systematically distort θ by reporting $\theta + d$, where $d \in \mathbb{R}$. We assume that when the bureaucrat is indifferent about reporting with or without distortion, they choose not to distort their scores ($d = 0$).[3] While we term these distortions "intentional" or "systematic," we make no normative claims about them. A bureaucrat might distort ($d \neq 0$) either to limit pesky interventions from the central government that disrupt more salient (and benevolent) goals of their entity (e.g., service provision) or to reduce the likelihood that corruption is detected through further monitoring. Our focus is to explain and describe this reporting behavior and its consequences—not to make a normative judgment about it.

Given a bureaucrat's choice of effort (e) and systematic distortion (d), the report that the central

[3]This assumption can be interpreted either as a behavioral assumption that the bureaucrat prefers to report honestly when distorting scores carries no benefit or as the outcome of imposing an arbitrarily small cost of misreporting.

government observes is given by:

$$r = \begin{cases} \theta + d + \varepsilon & \text{if the report is made} \\ \emptyset & \text{otherwise} \end{cases} \tag{1}$$

This expression follows directly from conventional expositions of measurement error in statistics (Cochran, 1968; Rubin, 1976). The terms d and ε capture systematic and non-systematic measurement error, respectively, and the possibility that $r = \emptyset$ reflects missing data.

How does a bureaucrat in a decentralized entity decide what to report? We propose a decision-theoretic framework to represent their incentives within the broader principal-agent relationship between central and decentralized governments. The central government can use the reported data, r, to target some form of enforcement (e.g., sanctioning noncompliance with a law) or initiate a data validation exercise. We denote the probability that the central government targets an entity for further investigation or validation as $\rho(r) \in [0, 1]$. We do not make assumptions about the functional form of $\rho(r)$ or whether non-reports draw more or less scrutiny than reports.

Second, the central government can impose some penalty on entities in the course of targeted audits on the basis of the information that is uncovered. Audits provide additional information about the true quality or state, θ, that the central government seeks to measure through reports. The magnitude of the penalty imposed, $P(r; \theta) > 0$, includes the costs of undergoing an investigation and any resultant punishment. It may vary with true performance (θ), the reported data (r), and/or the discrepancy between these measures. For example, a penalty strategy that seeks to promote accurate reporting will generally impose penalties that increase with the gap between reported and actual performance, i.e., $\partial P(r; \theta)/\partial |r - \theta| > 0$ (and convex in $|r - \theta|$). While we do not specify the precise functional form of P, in the context we study, the penalty is perceived to punish poor performance (i.e., low θ) or distortions in the reported data (i.e., a larger difference between r and θ).

Third, we assume that collecting, collating, entering, and reporting data requires the bureaucrat

to exert costly effort. We parameterize the cost as $c(e) > 0$, where $c'(e) > 0$ and $c(0) = 0$. Empirically, the cost of effort likely varies across bureaucracies as a function of administrative capacity, which includes the human capital of bureaucrats, the resources available for data collection and reporting, and access to technology.[4] These terms enter the bureaucrat's utility function in (2).

The expected punishment from the central government is given by $\rho(r)P(r;\theta)$, which we refer to collectively as "oversight." In this formulation, we assume that the bureaucrat internalizes (to some degree) the oversight of their organization more broadly. Bureaucrats may be directly punished for submitting faulty data or failing to report, and oversight activities—even for high-performing entities—often impose burdensome administrative work. It is important to note that bureaucrats may not know precisely $\rho(r)$ or $P(r;\theta)$; in these cases, what matters is their beliefs about these policies. Our experimental design targets these beliefs of decentralized bureaucrats.

$$E[U_B(d,e;\theta)] = \begin{cases} -\underbrace{E[\rho(\theta+d+\varepsilon)P(\theta+d+\varepsilon;\theta)]}_{\text{Oversight}} - c(e) & \text{if } r \in \mathbb{R} \\ -\underbrace{E[\rho(\emptyset)P(\emptyset;\theta)]}_{\text{Oversight}} & \text{if } r = \emptyset \end{cases} \tag{2}$$

Inter-institutional legibility. Inter-institutional legibility captures the degree to which the central government principal can "see" the performance of agents through the observed reports (r). Legibility would be perfect if all decentralized entities opted not to intentionally distort their performance ($d = 0$) and exerted sufficient effort such that $\sigma^2(e) \to 0$. In this case, $r \to \theta$. Legibility deteriorates when entities exert less (or no) effort, yielding noisy (or missing) reports, or when they engage in varying amounts of intentional distortion. We use this simple framework to understand what incentives drive departures from inter-institutional legibility.

Examining how central government oversight affects the quality of submitted reports suggests two countervailing tendencies, summarized in Remarks 1-2. First, since effort is costly, Remark

[4]To the extent that costs of effort proxy for administrative capacity, we follow Huber and McCarty (2004) in assuming that lower capacity increases noise in outcomes.

1 shows that the decision to report *any* information must be driven by the prospect of oversight. Specifically, the probability of investigation ($\rho(r)$) and/or penalty ($P(r;\theta)$) must be sufficiently large to outweigh the costs of effort incurred when reporting. In order to overcome this cost, it must be the case that the bureaucrat anticipates that oversight is costlier in the absence of a report.

Remark 1. *A decentralized entity can only be incentivized to report information to the central government, $r \in \mathbb{R}$, by the prospect of oversight (probability of investigation and/or penalty).*[5]

To file a report, the bureaucrat must exert some effort. In this case, the choice of effort depends on both the prospect of oversight and the cost of effort. Recall that the magnitude of unintentional errors decreases with bureaucrat's effort (e). This variance, $\sigma^2(e)$, enters the bureaucrat's expected utility if expected oversight, $\mathbb{E}[\rho(r)P(r;\theta)]$, is a non-linear function of the bureaucrat's report, r. This means that for many parameterizations of the monitoring rate and punishment, the bureaucrat considers how unintentional errors will shape their expected disutility from oversight, in addition to considerations of the cost of effort, $c(e)$. Importantly, we provide (reasonable) examples in Appendix A2.2 to demonstrate how the bureaucrat's anticipated prospect of oversight can *increase* or *decrease* in the variance of the bureaucrat's report. When the prospect of oversight is decreasing in this variance, the bureaucrat will exert less effort, yielding a noisier (or more ambiguous) report. This suggests that in some settings bureaucrats may prefer a higher degree of ambiguity.[6]

When is it optimal for a bureaucrat to intentionally distort their report? As can be seen from (2), distortion enters the bureaucrat's objective function only through the oversight term. The bureaucrat chooses to distort when she can shield herself from oversight by reporting a different score. Importantly, the national government may endeavor to avert this misrepresentation by punishing inaccurate reporting. Penalties of this form create a cost for misreporting, which enter the bureaucrat's utility through $E[\rho(r)P(r;\theta)]$, which depends on d (since $r = \theta + d + \varepsilon$).

[5] All proofs are available in the Appendix.

[6] This mechanism is proposed in other contexts including the choice of policy platforms by office-oriented politicians (Alesina and Cuckierman, 1990).

Remark 2. *Conditional on reporting, the bureaucrat's optimal effort, $e > 0$, is a function of the cost of effort and can be a function of the prospect of oversight. Intentional distortion, $d \neq 0$, can be optimal if and only if the prospect of oversight depends on reported scores, r.*

Together, Remarks 1 and 2 show that the anticipation of oversight has mixed effects on legibility. On one hand, the decentralized agent has no incentive to report in the absence of oversight. On the other, anticipating oversight can reduce report accuracy by attenuating the agent's effort and/or inducing the agent to misreport. Since oversight drives both systematic and non-systematic misreporting, the choices of d and e are likely to covary, but characterizing this covariance requires more concrete parameterization of the components of oversight. We examine this question empirically.

We note that the targeting of oversight and determination of penalties are ultimately policies set by the central government. Our primary goal is to understand how decentralized bureaucrats' beliefs about these policies shape their reporting behavior, allowing us to better characterize the incentives they face. As such, we analyze the bureaucrat's decision while treating government policies as exogenous. We revisit equilibrium considerations after presenting our empirical findings.

2 Case Context

Colombia is the most populous unitary state in the Americas. Following fiscal, political, and administrative decentralization in the 1980s and 1990s, it experienced a sharp increase in local-level data collection, consistent with our framework of a central government monitoring decentralized entities through data (World Bank, 2011). Today, the national government relies heavily on territorial governments' self-reported data to inform policy and guide oversight, as evidenced in discussions with our partner, the PGN, and in semi-structured interviews with bureaucrats who submit data to the national government.[7] As one secretary of planning in a small municipality noted: "*Data requests from [the national government] take so much time to complete. Some entities hire people just to fill out all such forms, but others that are smaller, are bound by law and cannot hire*

[7]See Appendix A5 for discussion of our sampling strategy for these interviews.

external contractors to do so, which means we have to do it with our own resources."

Despite some efforts to streamline reporting, data collection and submission remain central tasks for decentralized bureaucrats in Colombia. These tasks require non-trivial effort: in an original survey of bureaucrats in Colombian municipal governments (*alcaldías*), for example, Slough (2024) finds that 48% reported meetings or calls with national agencies in the past week, averaging two hours. National government data requests may overburden local bureaucrats, creating trade-offs that weaken policy or service delivery in other domains (Dasgupta and Kapur, 2020).

Interviews also suggest that the data reported by bureaucrats are consequential for the entities themselves. In an interview, a former National Planning Department official responsible for developing monitoring and evaluation systems in local governments remarked that requests for data "are not only perceived to be consequential, but they are in reality, as they are used to allocate national transfers, budget, evaluate entity performance, and even target oversight by national watchdogs through their local offices." We focus on the collection and use of data to target oversight.

2.1 The Transparency Law, the PGN, and the ITA matrix

We study the collection of the 2020 Transparency and Access to Information Index (ITA), an annual measure of institutional compliance with transparency practices, first implemented in 2018. The ITA was mandated by Colombia's Transparency and Open Data Law (*Ley 1712 de 2014*). The law aims to (i) guarantee citizens' right to access public information, (ii) promote proactive disclosure by public entities, and (iii) establish enforceable standards for what information must be disclosed and by which institutions. The law also provides a regulatory framework for central government oversight of transparency compliance across the public sector.

The Procuraduría General de la Nación (PGN), Colombia's principal watchdog agency, was tasked with implementing the ITA and plays a central role in this oversight function: it is the only institution mandated to systematically monitor transparency compliance across all public entities. Although the ITA matrix is publicly available, it is not designed for widespread citizen use; rather,

it serves as a key tool for the PGN to assess institutional risk and guide investigations.

The PGN: This central government entity investigates and sanctions irregularities or misconduct by elected officials, civil servants, and public entities. The PGN is widely recognized among Colombian bureaucrats, even at the local level. Multiple interviewees noted that all public servants must complete mandatory training—administered by the Administrative Department of Public Service (DAFP)—that explains the PGN's oversight functions. Furthermore, data on sanctioned public officials reveal that, outside the security services, the most frequently sanctioned type of official between 2013 and 2019 was a bureaucrat, not an elected politician (Table A3).[8] This is consistent with our claim that bureaucrats internalize some of the risk of oversight within their entity.

The PGN collects ITA data as part of its preventive mandate to monitor public officials and entities, aiming to reduce corruption and misconduct. This mandate relies on "police patrol" oversight (McCubbins and Schwartz, 1984), whereby ITA data are used to target investigations to specific public sector entities. If irregularities are detected, the PGN initiates disciplinary proceedings. ITA-directed investigations can lead to prosecution under the National Transparency Act or, more commonly, under other relevant public sector regulations. Among executive and legislative institutions (national, departmental, and municipal) with elected leaders ($n = 2,259$), 52.2% were investigated by the PGN in the year before our study (2019), and 24.6% of those investigations proceeded to a disciplinary investigation against an elected politician (see Table A2).[9]

The ITA matrix: The law mandates that more than 50,000 entities report data on transparency practices annually through the ITA, which classifies entities into three categories. First, *traditional subjects* consist of public sector entities, oversight bodies, and public companies that belong to the state. While these public sector entities include both central and territorial (decentralized) institutions, over 95% of these public-sector institutions are territorial entities, largely departmental

[8]Of course, there are fewer elected politicians than bureaucrats, so the rate of sanction for politicians is likely higher, but we do not have denominators to accurately construct these rates.

[9]These institutions comprise 34.5% of our main sample. We rely on data from this subsample because more information on investigations is available for these entities.

and municipal government institutions. The remaining organizations fall into two additional categories: (2) private firms or individuals contracting with the state and (3) political parties and social movements. Discussion of the latter two categories is relegated to the Appendix.

The ITA questionnaire asks agents of all entities to self-report their entity's compliance with transparency practices related to public contracting, oversight, regulation, and budgeting, among other aspects of management or governance. The survey consists of approximately 200 yes/no responses, which are weighted according to a predetermined formula to generate the final ITA score, ranging from 0 to 100. A score of 100 indicates full compliance with the relevant transparency practices, while 0 reflects non-compliance with these regulations.[10] The PGN publishes these measures in a consolidated report, which compiles ITA scores across entities. Each year, the PGN delegates the request to complete the ITA to various central government agencies, referred to as "heads of sector." In practice, this means that entities receive the ITA submission request from a different (non-watchdog) entity. Most public sector entities receive the request from the DAFP.

The PGN sought a collaboration with researchers on the 2020 ITA data collection due to concerns about high rates of non-response. In 2019, just 52.2% of public sector entities completed ITA. Per our framework, non-completion could stem from (a) prohibitively high costs of effort; (b) a belief that oversight does not depend (sufficiently) on what is reported; or (c) a combination of effort costs and sufficiently low anticipated rate of PGN investigation in a given year. While the PGN states that these data inform preventative anti-corruption efforts, low response rates and unknown accuracy render reliance on ITA to target police patrol investigations problematic. These reporting pathologies may also create perverse incentives: entities that honestly disclose imperfect transparency practices may be penalized, while those that fail to report or falsely claim compliance may be able to skirt oversight. Beyond these broad contours, the exact use of ITA data by the PGN (i.e., monitoring rates) remains undisclosed. Interviews with bureaucrats who submitted ITA data

[10] All items in the questionnaire refer to observable, objective practices (e.g., whether a document is posted online) and do not involve the central government's subjective evaluation of the quality of implementation or the adequacy of the disclosed information.

revealed varying beliefs about the use of the ITA matrix in monitoring.

3 Research Design

We conduct a pre-registered field experiment in collaboration with the PGN, which sought to determine whether low-cost strategies could increase rates of complete data submission. We also seek to understand the quantities reported and their fidelity to actual transparency practices. To do so, we conduct an independent audit to describe the quality of the ITA data.

3.1 Sampling

Our unit of assignment is the entity or organization. The experimental sample includes the near-universe of public sector entities in Colombia (99%). For the audit, we draw a stratified random sample, oversampling national (central government) entities given their relatively small number. Table 1 provides details on the population of entities, and the experimental and audit samples.

	All Public-Sector Entities*		**Experimental Entities**		**Audited Entities**	
Category	Count (n)	%	Count (n)	%	Count (n)	%
National	237	3.6%	237	3.6%	200	8.3%
Territorial	5,928	90.4%	5,928	90.4%	2,200	91.7%
Undesignated	391	6.0%	391	6.0%	0	0%
TOTAL	6,556	(100%)	6,556	(100%)	2,400	(100%)

Table 1: Sampling of public-sector entities in experiment and audit outcome measurement. *Total omits 62 randomly sampled entities that were used in a pre-test of intervention implementation.

3.2 Intervention and Assignment

The experiment has two levels of treatment. The first tests the effects of increased oversight salience by the PGN on bureaucrats' reporting behavior. Our primary manipulation involves direct communication from the PGN to public entities. In the status quo (control condition), the PGN followed its past practice (from 2018 and 2019) of delegating data requests to sector heads, all central government entities. Indirect communication from sector heads typically consists of social media posts and other online messaging. To increase the salience of the PGN's role in data collection,

we randomly assign some entities to receive a direct email from the PGN requesting the data. This first-level treatment compares delegation to sector heads with a combination of delegated communication *and* direct communication from the PGN. Direct communication is expected to heighten perceptions that responses will be scrutinized and that non-compliance may lead to sanctions.

Interviews with bureaucrats who submitted the 2020 ITA data suggest that the link between communication source and oversight salience aligns with their perceptions. For example, an official at a public university stated: "*There can exist sanctions, as this is one of the PGN's core functions: to monitor what we do. But, to be honest, I don't know the types of sanctions that can be imposed for those who either do not complete the form or fill it out inaccurately.*" Direct communication clarifies the PGN's role in collecting and using the data, aiming to reduce such uncertainty. These accounts support interpreting direct communication as a shock to perceived oversight.

Within entities randomly assigned to receive direct communication, we subtly vary the content and frequency of the messages using a $2 \times 2 \times 2 \times 2$ factorial design. Table 2 summarizes this variation in content (Table A6 provides the full text of the messages). We refer to these second-level treatments as "nudges," following Thaler and Sunstein's (2008) definition of nudges as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (p. 6). Unlike the direct communication treatment, which highlights the PGN's authority and use of the data and thereby shifts bureaucrats' perceived incentives, the nudges consist of small in message framing.

While theory-driven, these nudges represent much weaker interventions than the top-level randomization to direct communication. The nudges serve both practical purposes for the PGN and analytical benefits for our study. Direct messaging is costly for the PGN, requiring staff time and expertise to tailor communications and respond to an increased volume of inquiries. In contrast, modifying message content is costless once emails are being sent. How to optimize these communications at no additional cost was an important consideration for our partners. The specific nudges were informed by PGN officials' hypotheses about the sources of non-compliance, as detailed in

Table 2. From a design perspective, one might worry that "direct communication" is too compound a treatment. By varying the content of these messages, we can partially isolate the effect of communicating oversight from potential artifacts introduced by the message text itself.

Nudge	Levels	Motivation
Past (retrospective) oversight	0 = No mention of past compliance with collection of ITA data. 1 = Acknowledgement of compliance/non-compliance with 2019 ITA data collection.	Highlight the PGN's observation of past data outputs. Note that the content of the message varies according to past compliance (two versions of the text).
Future (prospective) oversight	0 = No mention of possible audits to 2020 ITA submissions 1 = Mention of possible audits of 2020 ITA submissions.	Increase perceptions of the likelihood of sanction or enforcement for non-completion of ITA.
Training	0 = No information on training resources for filling out ITA. 1 = Link to PGN resources (including videos) on how to fill out ITA.	Increase the capabilities of agents with respect to ITA data submission.
Reminder	0 = Single direct communication from PGN to entity. 1 = Direct communication + a reminder from PGN to the entity.	Reinforce perception of PGN oversight over ITA completion.

Table 2: Nudge treatments randomized within the direct contact communications between the PGN and the entities. These treatments were implemented as a $2 \times 2 \times 2 \times 2$ factorial design.

We block-randomized treatment across the 6,556 public entities in our experimental sample. First, we stratified entities based on ITA completion in 2019, generating two subgroups. Within each subgroup, we created blocks of 18 entities by minimizing Mahalanobis distance on covariates, using (1) PGN's classification of entity type and (2) department indicators. This distance minimization ensures, for example, that local governments in the department of Antioquia are most likely to be in the same block as other local governments in Antioquia. This process yielded 190 complete blocks among entities that completed the ITA in 2019 (3,420 entities), and 174 complete blocks plus one partial block among those that did not (3,136 entities). Within each block, we randomly

assigned 2 entities to a pure control condition and the remaining 16 to the $2 \times 2 \times 2 \times 2$ factorial design, bringing the number of entities per block to $n = 18$. As such, $\frac{8}{9}$ of entities received some form of direct communication. Figure 1 summarizes the experimental design. We report balance on pre-treatment covariates, including recent exposure to PGN oversight, in Figure A2.

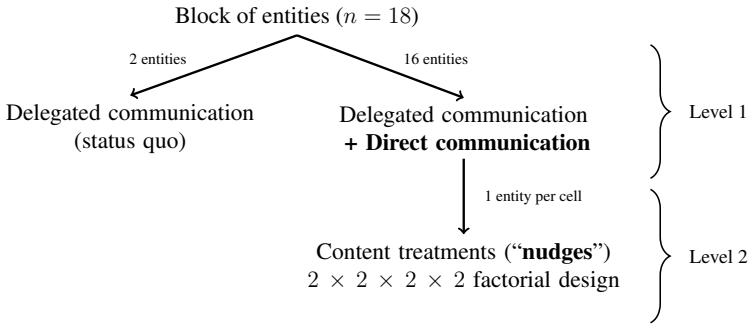

Figure 1: Treatment assignment scheme within each block of 18 entities.

3.3 Independent Audit of Data Quality

Outside our collaboration with the PGN, we independently audited ten components (each with at least one item) comprising 32.4 points of the 100-point ITA scale, focusing on some of the most prominent transparency concerns. The audit was conducted by an independent firm hired by the researchers in June-July 2021. Auditors were trained to search for selected ITA index items using a standardized protocol. They recorded compliance with each item. We describe the audited components in Appendix A4. While the selection of components was not random, the sample disproportionately includes components that are both highly relevant to citizens and weighted more heavily in the total ITA score, underscoring their importance to the central government (Table A9). These components vary meaningfully in how difficult they are to fulfill.[11] Our measure of latent quality is constructed from indicators of compliance with each constituent ITA item.[12]

[11] For instance, one audited component requires entities to host an online form for public information requests—a relatively low-effort action signaling passive transparency. In contrast, another component requires entities to disclose all public contracts and upload procurement data to the national contracting platform (SECOP)—a task that demands continuous coordination, recordkeeping, and interaction with a third-party system.

[12] Our main findings are robust to dropping any single component in leave-one-out analyses reported in Figure A11.

Crucially, we conduct the audit in parallel for entities that reported and those that failed to report in the 2020 ITA data collection. Given the large number of entities in our study and the time requirements of the audit, we restricted the audit to a stratified random sample of 2,400 public sector entities. Since the audit sampling oversamples national entities, we include indicators for national versus decentralized entities throughout the analysis of the audit data. In Tables A7-A8, we show that, conditional on this indicator, assignment to the independent audit is balanced across past (2018 and 2019) ITA submissions and scores, as well as the experimental treatments.

One potential concern is that, given complications in identifying, contracting, and training the firm for this non-standard audit, too much time may have elapsed between ITA submission and the audit six to seven months later. Although we assume θ is sticky in the short term, it is possible that six months was enough time for entities to change their performance. These reforms should not bias our results unless (1) entities became more transparent due to the treatment only after submitting their data to the PGN, or (2) changes in transparency practices between submission and audit vary with the true level of transparency. In Figure A3, we show that experimental treatments do not affect the underlying quality measure, which addresses the first concern. As for the second, our interviews suggest that entities are more likely to tailor their transparency practices before submitting reports than after, and revisions rarely occur before the next official request. The PGN did not begin using the 2020 ITA data for oversight until the second half of 2021, after our audit.

The audit affords us a measure of "true quality," or latent transparency practices within entities. While θ is undoubtedly measured with some error, our primary goal was to ensure that measurement error in θ is independent of the measurement error in the data submission process: systematic distortion of transparency practices (d) or random error (ε). By hiring auditors outside the confines of our collaboration with the PGN, we eliminate the specific incentives for misrepresentation that are potentially present in the relationship between the PGN and reporting entities.

3.4 Measures

We measure the theoretical parameters r, entities' reports of transparency practices, and θ, the true level of transparency practices. Our primary measure of r comes from PGN's internal record of scores, which we transform to create two outcome measures. The first is a binary indicator for data submission to the PGN, coded "1" if an entity submitted ITA data. The second is the ITA index score, which ranges from 0 to 100. Naturally, scores are only observed when data is submitted.

Our measure of θ comes from the audit. To maximize comparability to the overall score and maintain the weighting used in indexing, we reconstruct an ITA-like index for the audited items, yielding a score between 0 and 32.4 by weighting binary indicators of compliance with each item following the weights in the index. We contrast the outcomes of these calculations to measure the divergence between reported and actual transparency practices. To facilitate this comparison, we construct an analogous index for audited items from the microdata, also ranging from 0 to 32.4.[13]

3.5 Identification and Estimation

The first level is a two-arm design that identifies the average treatment effect (ATE) of direct communication from the PGN. The second level introduces a factorial design that allows us to estimate the average marginal component effects (AMCEs) of four nudges embedded in the message content. We estimate all effects using OLS, as specified in Equation 3. The ATE of direct communication is captured by β_1, while the AMCEs of the message components are captured by β_2 through β_5:

$$\begin{aligned} Y_{ib} =& \beta_1 \text{Direct Communication}_i + \beta_2 \text{Reminder}_i + \beta_3 \text{Training}_i + \beta_4 \text{Retrospective Oversight}_i \\ &+ \beta_5 \text{Prospective Oversight}_i + \boldsymbol{\psi}_b + \epsilon_{ib} \end{aligned} \tag{3}$$

Each of the treatments is a binary indicator of assignment to the treatment condition. $\boldsymbol{\psi}_b$ represents a vector of block fixed effects. The block indicators subsume past completion of ITA given our

[13] We discuss the quality of the microdata in greater detail in Appendix A4.

exact blocking strategy. We also report estimates of the ATE of direct communication that pools over the message treatments in (3) by omitting the indicators for the nudge treatments.[14]

We also estimate the effects of the experimental treatments on reported ITA scores using the same specification. Because the sample for this outcome is conditioned on submission, the $\widehat{\beta}$'s are not, in general, estimators of well-defined causal effects. However, as we show in Appendix A8, the post-treatment estimand can be decomposed into a convex combination of the conditional average treatment effects (CATEs) of direct communication among entities that would always report and the average reported score among entities that report *because* of the direct communication treatment. The latter quantity is not a causal effect. However, both quantities correspond to behavior we discuss in Section 1.1. To decompose these two effects, we invoke a monotonicity assumption and then use Lee (2009) trimming bounds to bound CATEs among always reporters. This allows us to algebraically back out an interval estimate of the average reported scores of if-treated reporters. This decomposition is a novel contribution of this paper that permits us to study both selection into reporting and changes in reporting behavior.

Our framework also emphasized the importance of description of the relationships between "true" latent levels of transparency practices and reporting behavior. In our non-experimental analysis, we examine the relationship between our audit measure of θ, denoted Audit_i and reporting outcome Y_i. The basic form of these OLS regressions is:

$$Y_i = \gamma_0 + \gamma_1 \text{Audit}_i + \boldsymbol{\kappa}\mathbf{X}_i + \epsilon_i \tag{4}$$

Our goal in these analyses is to describe the association between the latent and reported data. We also use more flexible specifications to characterize potential non-linearities in the associations between the these variables.

[14]Power calculation simulations reported in our pre-analysis plan suggest that the design is powered to detect a 4.5 percentage point increase in response rate due to direct communication and a 2.75 percentage point increase in response rate due to the nudge treatments (for $\beta = 0.8$).

3.6 Linking ITA to the Theoretical Framework

Two aspects of the ITA index and our analytic strategy depart from the abstract formulation of reporting in our theoretical framework. First, in our model of a single decentralized entity, variance in reported scores comes from non-systematic distortion in reported scores (ε), which is a function of the bureaucrat's effort. However, as in any quantitative study, a second source of variance in observed scores comes from examining multiple decentralized entities. This admits variation in beliefs about how ITA reports will be used by the PGN and in effort costs. In light of these sources of variance, we discipline our interpretation of what generates variation by (1) stratifying on our measures of θ in the audit analysis,[15] and (2) reasoning through which observed patterns could be driven by non-systematic measurement error or variation in beliefs about oversight in isolation.

Second, in our framework, performance (θ) is continuous and unbounded. However, the ITA index is a 100-point index constructed from the weighted average of binary responses. This means that non-systematic errors come from clicking the wrong "yes" or "no" response. But if an entity has perfect performance ($\theta = 100$), inadvertent errors can only be "no" responses, which would lead to a *downward* bias in reported score. For an entity with the worst possible performance ($\theta = 0$), inadvertent errors can only be "yes" responses, which would lead to an *upward* bias in reported scores. This means that $E[\varepsilon|\theta]$ varies in θ and is generally non-zero. The overall bias attributable to these unintentional errors therefore depends in the distribution of θ in the sample. Since we measure θ directly in the audit, we use simulations of different error rates to inform our interpretation of the incidence and consequences of non-systematic measurement error in sample.

4 Results

4.1 Direct Communication and Bureaucrats' Reports

How does increasing the salience of oversight change reporting behavior? Panel A of Table 3 (columns 1–2) reports estimates of the ATE of direct communication and, Panel B reports the

[15] This is important because optimal reporting strategies generally depend on θ.

AMCEs of the nudge treatments on the probability of submitting ITA data.[16] We find that direct communication increases the probability of reporting by 3 percentage points, though this increase is only marginally statistically significant ($p < 0.1$) in the fixed-effects specification that pools over the nudges (Panel A, column 2). Repeated direct communication in the form of a reminder increases the probability of reporting by an additional 1.2 percentage points, which is similarly not significant. Combined, however, these estimates suggest that a higher dosage of direct communication from the PGN increases submission rates by 4.2 percentage points ($p < 0.037$ in a two-tailed test). The estimated AMCEs of the other nudge treatments are near zero and are not significant.

	Completed ITA $\mathbb{I}(r \neq \emptyset)$		Score r	
	(1)	(2)	(3)	(4)
PANEL A: EFFECTS OF DIRECT COMMUNICATION				
Direct communication	0.029	0.029*	-7.972***	-7.817***
	(0.019)	(0.015)	(1.162)	(1.076)
PANEL B: EFFECTS OF DIRECT COMMUNICATION, NUDGE TREATMENTS				
Direct communication	0.029	0.030	-6.066***	-6.030***
	(0.022)	(0.018)	(1.477)	(1.356)
Oversight of past completion	0.000	0.000	0.587	0.911
	(0.012)	(0.010)	(0.950)	(0.856)
Possible future audit	-0.005	-0.006	-0.453	-0.925
	(0.012)	(0.010)	(0.950)	(0.859)
Direct reminder	0.013	0.013	-2.836***	-2.418***
	(0.012)	(0.010)	(0.949)	(0.856)
Training	-0.008	-0.009	-1.094	-1.127
	(0.012)	(0.010)	(0.950)	(0.858)
Num. Obs.	6556	6556	4446	4446
Block FE		yes		yes
Control mean (std. dev.)	0.65 (0.48)	0.65 (0.48)	80.49 (23.12)	80.49 (23.12)
DV range	{0,1}	{0,1}	[0,100]	[0,100]

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 3: ATE and AMCE estimates of the messages and message content on ITA data submission (columns 1-2) and the association between treatments and transparency index scores, conditional on submission (columns 3-4). Heteroskedasticity-robust standard errors in parentheses.

The estimated effects of direct communication on report submission suggest that while communication from an oversight body increases reporting, the effects are small in magnitude. While our framework emphasizes the importance of oversight in motivating the submission of a report (Remark 1), it is possible that direct communication instead served as a pesky reminder to bureau-

[16] Across public sector entities assigned to direct communication, 94.8% of emails from the PGN were delivered as reported in Table A4.

crats. If this were the case, we might expect direct communication and reminders to be especially effective for entities that had not filled out the ITA in the past. To test this, we examine whether direct communication and the nudges have differential effects on reporters versus non-reporters (from 2019) in Appendix Figure A5. While past non-reporters were 48% less likely to report in 2020, we do not detect different effects between these two subgroups. These differences are all near-zero and statistically indistinguishable from zero.

In Columns (3)-(4) of Table 3, we regress scores on the experimental treatments, restricting the sample to entities that submitted a report. Because these specifications condition on a post-treatment outcome, estimates cannot be interpreted as causal effects. Our estimates suggest that direct communication is associated with *reductions* in reported scores. Recall that lower scores indicate less transparency and suggest worse performance to the PGN. Reminders are associated with an additional (additive) reduction in scores.

As we explain in Appendix A8, the post-treatment estimand can be decomposed into a weighted sum of two components: the conditional average treatment effect (CATE) among "always reporters" and the average reported score among "if-treated reporters" (see Section 3.5). For direct communication, for example, a non-zero CATE implies that some always-reporting entities report different scores when contacted directly by the PGN than they would have if not contacted. The selection term captures the expected score among entities that report only when contacted directly by the PGN but would not report otherwise.

Before presenting the results of this decomposition, we assess the assumption of *monotonicity* in selection into reporting. In this context, monotonicity requires that no entities report *because* they were not assigned to direct communication, and none fail to report *because* they were assigned to it. To evaluate this assumption, we use all pre-treatment covariates provided by the PGN to estimate heterogeneous treatment effects on reporting using generalized random forests (Athey, Tibshirani, and Wager, 2019). This analysis yields predicted CATEs for all entities in our sample. In Figure A7, we show that no units exhibit a negative treatment effect, while 886 out of 6,556

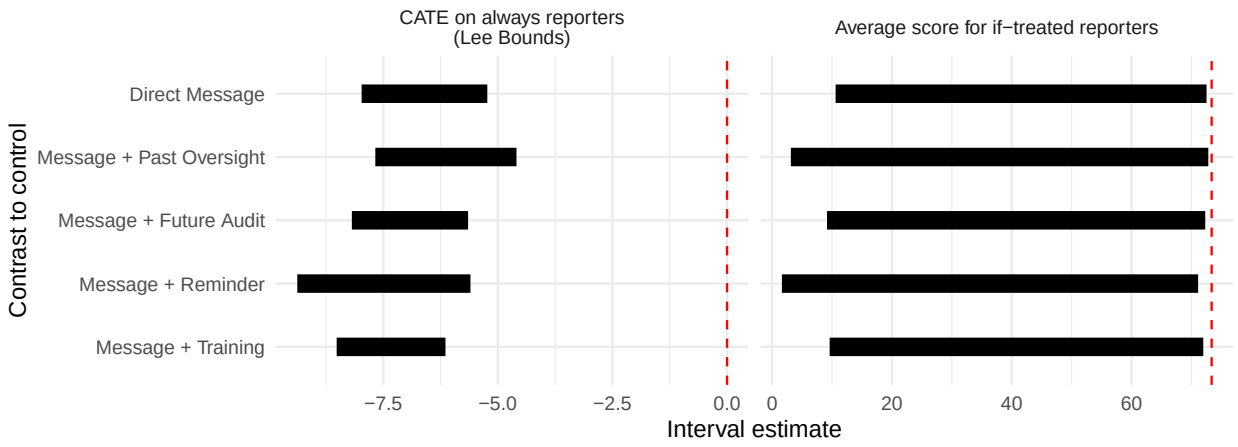

Figure 2: Decomposition of post-treatment estimands into a CATE on always reporters (left) and the average score among if-treated reporters (right). See Appendix A8 for the derivation of the decomposition. Dashed lines indicates a CATE of 0 (left) and the average reported score (right).

entities show a positive and statistically significant effect of direct communication on reporting. These findings support the assumption of monotonic selection into reporting.

In Figure 2, we report interval estimates of the CATE among "always reporters" and the average scores among "if-treated reporters." The top interval estimate defines treatment as the "direct message" alone (as in our previous discussion). The CATE estimates are clearly negative. This suggests that, on average, "always reporter" entities report *lower* average scores when exposed to oversight through direct communication. Our interval estimates on the average scores of if-treated reporters are wide across all operationalizations of treatment. Nevertheless, in all cases, these average scores are *lower* than the average scores of all reporters. This indicates that if-treated reporters must report *lower* average scores than always reporters. These findings suggest that exposure to oversight does measurably change the reporting behavior of bureaucrats in entities both through changes in the scores reported by bureaucrats and changes in sample selection. We provide bootstrapping-based uncertainty estimates in Table A11. The remaining intervals in Figure 2 redefine treatment as a direct message *and* one of the nudges versus pure control. We see that our inferences are robust to redefining the content of treatment in this way.

Collectively, Table 3 and Figure 2 provide compelling evidence that reporting behavior is sensitive to oversight by the PGN. Although we do not find detectable effects on ITA submission at standard thresholds, we show that when exposed to oversight, some entities report lower scores than they would otherwise report. This effect among always reporters suggests that direct communication is more than a mere reminder—it changes the scores reported by (some) entities that would have submitted the form to the central government regardless. While the experimental data allow us to show that entities adjust their reporting behavior in response to increased oversight salience, they do not allow us to assess the accuracy of those reported scores, as we lack a direct measure of actual transparency practices (θ). To this end, we turn to the audit data.

4.2 Assessing Inter-Institutional Legibility

1. Selection into reporting: We examine whether entities' propensity to report varies with their underlying transparency practices. The left panel of Figure 3 plots the probability of submitting a report as a function of the audit score (formally, $\Pr(r \neq \emptyset|\theta)$). We find strong evidence of positive selection: entities with higher levels of transparency are significantly more likely to report. For instance, entities scoring zero on the audit metric report with probability 0.43 (95% CI: [0.30, 0.48]), while those with perfect scores report with probability 0.82 (95% CI: [0.80, 0.85]). This suggests that higher-performing entities are substantially more likely to invest effort in reporting than lower-performing ones.[17] Table A12 further shows that, to the extent that marginal reporting costs decrease with administrative capacity, the probability of reporting increases with two measures of capacity—consistent with expectations about how effort costs influence reporting.

From the perspective of the PGN or another data analyst, the pattern of selective reporting depicted in Figure 3 yields scores on audited items distributed according to the purple conditional density in the right panel, while unreported scores are distributed according to the orange conditional density. The vertical lines denote the means of each distribution. The difference between these means (10.16 points) is equivalent to 0.73 standard deviations of the audit-based measure of

[17] In our framework, reporting requires investment of effort, whereas non-reporting does not.

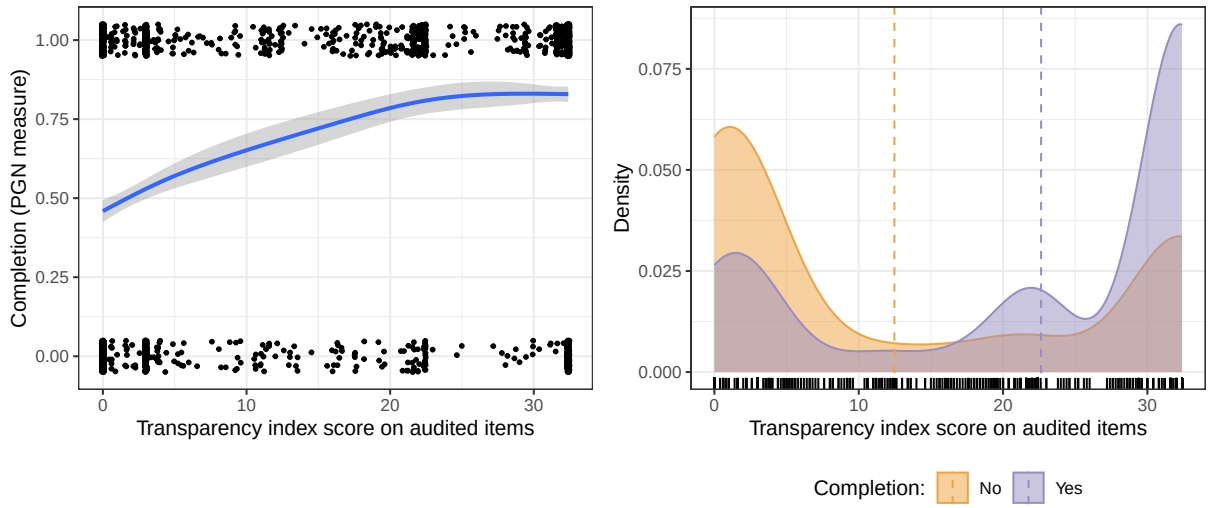

Figure 3: The association between the audit-measured transparency index and the probability of ITA data submission (left). The distribution of the audited-measured transparency index among entities that completed and failed to submit ITA data (right).

transparency practices. Finally, we note that the modal reporting entity is high performing on the audit-measured transparency index. This matters for our interpretation of misreporting because, within a sample of (disproportionately) high-performing entities, non-systematic measurement error should lead to downward bias in reported scores (on average), as illustrated in Figure A8.

2. Misreporting of transparency practices: We now turn to comparing the results of the audit to the data submitted by the entities directly to measure the accuracy of entities' reports. This analysis necessarily conditions on submission of ITA data. We first show that our audit-based measure of compliance with transparency practices (θ) correlates strongly with self-reported measures of compliance (r). We consider two different self-reported outcomes. First, we work from the available public reports to construct the reported compliance with the same subset index components that we audit. This subset of the transparency index constitutes 32.4 of the 100 points. Second, we use the PGN's official scores on the full transparency index. Table 4 reveals a positive correlation between scores and each of the self-reported outcomes.

How should the coefficient on the audit score (β_{Audit}) be interpreted? On one hand, $\beta_{\text{Audit}} = 0$

would indicate that reported scores were completely uninformative of actual transparency practices. On the other hand, because the transparency index is additive, in the absence of distortions in reporting behavior or measurement error in the audited data, we would expect that $\beta_{\text{Audit}} = 1$ for both outcomes. We can soundly reject both null hypotheses ($p < 0.001$ in all tests), suggesting that scores are somewhat informative but distortions are present.

	(1)	(2)	(3)	(4)	(5)	(6)
Audit score	0.565***	0.566***	0.557***	0.619***	0.620***	0.618***
	(0.025)	(0.025)	(0.026)	(0.063)	(0.063)	(0.064)
Intercept	8.466***	9.291***	9.069***	59.32***	64.287***	64.180***
	(0.708)	(1.009)	(1.016)	(1.791)	(2.465)	(2.487)
Num. Obs.	1307	1307	1307	1696	1696	1696
National entity idnicator	yes	yes	yes	yes	yes	yes
Experimental treatment indicators		yes	yes		yes	yes
Elected entity head indicator			yes			yes

$^{+}p < 0.1,^{*}\ p < 0.05,^{**}\ p < 0.01,^{***}\ p < 0.001$

Table 4: The association between audited and reported scores. Heteroskedasticity robust standard errors in parentheses.

We examine these distortions as a function of audit-measured transparency practices (θ) in Figure 4. In the left panel, we plot our audit-based measure θ against the difference between self-reported and audit-measured transparency practices on the same subset of items ($r - \theta$). The generalized additive models suggest that low- and middle-performing entities tend to overreport their compliance with transparency practices, as the curves are greater than—and statistically distinguishable from—zero. Importantly, given the bounded ITA index and the prevalence of high θ's among reporting entities, this negative slope could be consistent with non-systematic measurement error in isolation. The right panel examines the association between audit-measured transparency practices and the magnitude of distortions. We show that distortions are slightly decreasing in true levels of transparency. Collectively, these plots suggest that, on average, bureaucrats over-report compliance with transparency practices, at low and middling levels of transparency. These deviations from θ should be interpreted as a combination of systematic and non-systematic distortions.

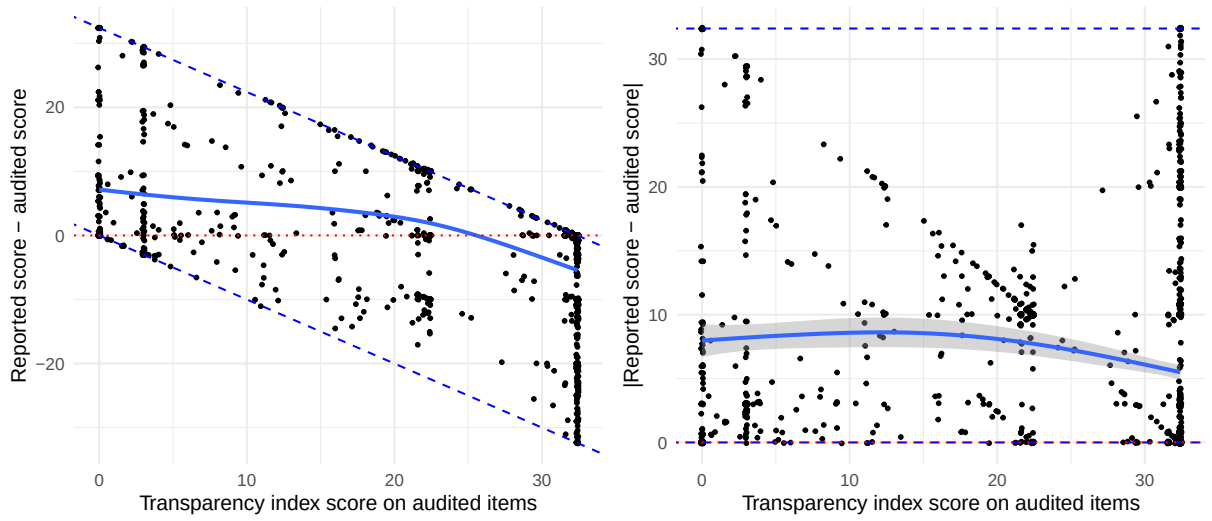

Figure 4: Discrepancies between the reported transparency practices and those detected in the audit. The left panel reports our measure of $r - \theta$ and the right panel reports $\mid r - \theta \mid$.

3. Noise in reporting: We next consider the variance in scores as a function of underlying transparency practices. At the individual entity level, our framework suggests that greater variance in reported scores is indicative of lower effort. However, when aggregating across entities, variance could also emerge from variation across entities in (1) beliefs about oversight or (2) in the cost of effort. In Figure 5, we estimate the standard deviation in scores—both for the subset of audited items from the microdata and for overall scores—as a function of audit-detected quality. We find that the standard deviation (and thus variance) in scores is higher among entities with lower transparency practices. This pattern appears both when examining the standard deviation within bins of audit-measured transparency practices (left panel) and when using a triangular kernel to estimate the conditional standard deviation across the support of the audit measure (right panel).

Why are reports noisier among less transparent entities? One possibility is that lower-performing entities exert less effort when submitting ITA data. Our simulation in Appendix Figure A8 shows that even with the bounded ITA measure, the pattern that we observe in Figure 5 cannot be produced by equal rates of non-systematic measurement error in isolation. The patterns we observe must be driven by variation in (i) effort—and thus non-systematic measurement error—at different

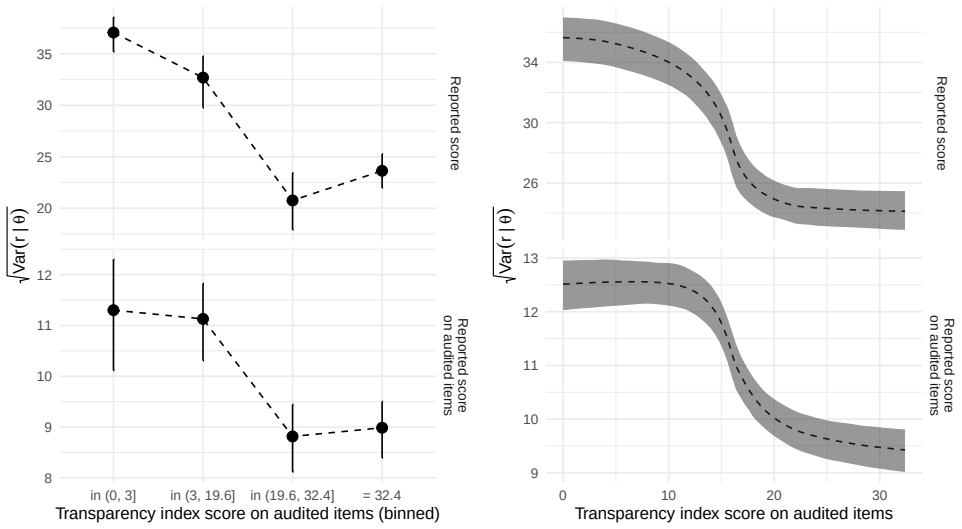

Figure 5: This plot shows how noise in reported ITA scores relates to the audit-measured transparency index. The left panels bins entities by level of audit-measured transparency. The right panels employs a triangular kernel to estimate the conditional standard deviation.

levels of θ or (ii) variation in costs of effort or beliefs about monitoring that correlate with θ. Using administrative capacity as a proxy for effort costs, Table A12 shows that lower-capacity entities report with more noise than higher-capacity counterparts, consistent with our expectations. However, Figure A9 reveals that the conditional standard deviation decreases in θ even within levels of capacity, suggesting that effort costs alone cannot explain the pattern. We cannot eliminate the possibility that low-θ entities hold more heterogeneous beliefs about oversight, which may also contribute to greater variation in reports.

4. Oversight: Our experiment shows that increasing the salience of oversight shapes reporting behavior, by changing *which* entities report and *what* they report. We now bring all the pieces together, aided by audit-based measures, to examine whether responses to oversight vary with entities' actual transparency practices. The left panel of Figure 6 shows that the effect of oversight on selection into reporting is strongest among entities with moderate-to-high levels of latent transparency. Entities with transparency scores around the mean of reporting entities (19.65 out of 32.4) are more likely to submit reports when oversight is made salient (see Figure A10). This is con-

sistent with our bounds in the right panel of Figure 2, which suggest that if-treated reporters have slightly lower average scores than the overall sample mean.[18]

In turn, the right panel of Figure 6 shows that direct communication is most effective at reducing distortions precisely among the entities in which upward distortions are most prevalent: at middling levels of transparency practices (see Figure 4). We interpret this result cautiously, since it conditions on reporting, which is itself an outcome of the oversight treatment. Overall, these findings reinforce the main messages from both the experiment and the audit, suggesting that our research design effectively captures underlying reporting behavior by bureaucrats.

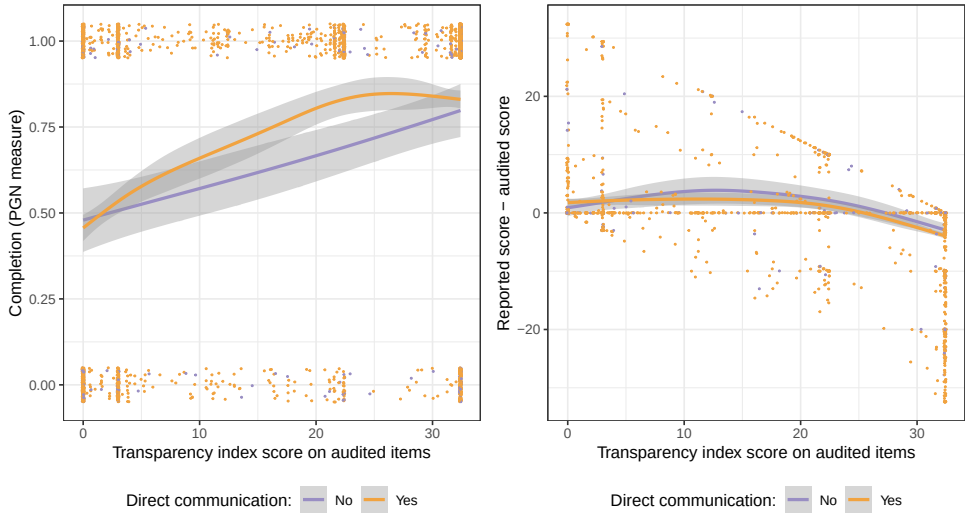

Figure 6: Conditional average treatment effects of oversight on the probability of ITA data submission at different levels of audit-measured transparency practices (left). Discrepancies between the reported transparency practices and those detected in the audit by treatment condition at different levels of audit-measured transparency practices (right).

5 Discussion

We have shown two central findings in the context of Colombia's ITA data collection. First, decentralized entities' reporting behavior responds to direct communication from the PGN about its role in the data collection. Second, both non-response and distortions in ITA data vary with entities'

[18]The overall sample mean is driven by high-performing always reporters, given the small effect on reporting rates.

true (latent) level of transparency practices, the quantity the PGN seeks to measure. These findings underscore the challenge for the central government—here, the PGN—in designing data collection schemes and using the resultant data.

The fact that the PGN invested in this collaboration with researchers suggests that they value better data quality. Their willingness to randomize indicates some uncertainty about how to best pursue this goal. The current interventions represent an effort to improve upon status quo communication of the ITA data collection that was previously delegated to other central government agencies. The intervention and the modest experimental results should be interpreted relative to this status quo control, rather than an evaluation of a communication strategy that should be effective in all settings. Specifically, our experimental design fixes the central government's behavior, allowing us to isolate the response of decentralized bureaucrats. Hence, our estimates capture partial equilibrium responses to changes in communication.

Still, our results suggest that central government choices about how to use data, or how to communicate their use, can affect the quality of data submitted by decentralized entities. We explore these dynamics further by considering the central government's strategic decisions. In our model, the center can adjust two policy levers: the targeting of oversight ($\rho(r)$) and the penalties associated with audit outcomes ($P(r;\theta)$)—in addition to how these policies are communicated. These tools give the central government leverage to shape reporting behavior, and hence the quality of the data it ultimately observes. As our findings show, decentralized entities respond strategically to perceived oversight, at least to the extent that they understand how the data will be used.

To build toward these considerations, we use the observed reports and our audit data to explore what the national government might "see" under different oversight strategies. Specifically, we focus on $\rho(r)$, the targeting of oversight. Our simulation takes the data inputs from the "direct communication" arm of the experiment (to measure r) and the audit (to measure θ) and we choose several functional forms for $\rho(r)$—the national government's method for targeting oversight—to

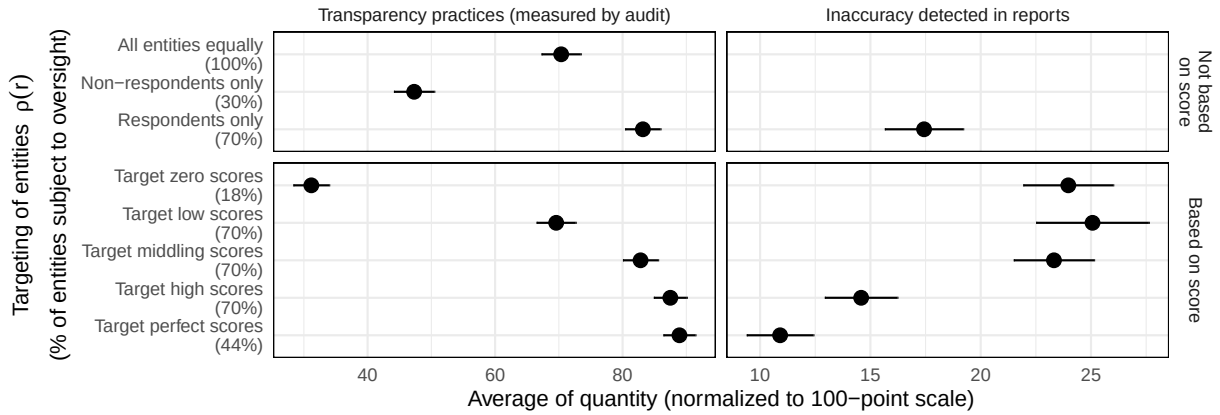

Figure 7: Average levels of θ (left) and $|r - \theta|$ (right) that the national government would observe under various monitoring strategies. Measures come from observed reports (r) and our audit (θ), whereas the monitoring rate, $\rho(r)$ is simulated according to the functional forms in Table A10. The percentages in the labels represent the share of public sector entities eligible for monitoring under a given monitoring strategy. 95% confidence intervals correspond to a sample of 1,000 entities.

simulate what such oversight would uncover.[19] The results are depicted in Figure 7.

Consistent with the results in Figure 3 and Table 4, strategies that audit (i) entities reporting zero scores or (ii) non-respondent entities are best able to target low-transparency entities (left panel). At the same time, these policies curtail substantially the set of entities that would ever be audited at just 18% and 30% of public sector entities, respectively. In contrast, to maximize data quality, an oversight strategy that audits low—but non-zero—scores with higher probability is apt to detect larger distortions in the data (right panel), consistent with Figure 4. In this case, all reporting entities (70% of public sector entities) are eligible but the probabilities are higher for entities that report lower scores.

By fixing the data inputs from our experiment, this simulation does not account for the strategic response of local bureaucrats to different targeting strategies. Our theory suggests that when designing and communicating an oversight strategy, the national government is likely to face a tradeoff between the quantity of reporting entities and the accuracy of the resultant data. For instance, if the national government simply sought to maximize completion—perhaps to maximize

[19] See Appendix A10 for more details about the input used in these simulations.

contact with decentralized entities—it would presumably focus on investigating non-respondents. Anticipating this, however, entities should report at higher rates, but not necessarily invest substantial effort in preparing their responses, thereby generating noisy reports. Conversely, if the national government sought to use the ITA to identify low-transparency institutions (as we interpret its present goal), entities with low θ would presumably abstain from reporting or overreport θ in an effort to avoid scrutiny, consistent with our empirical findings.

Different data collection strategies by central governments have different goals, and principals must weigh the countervailing effects of oversight on completion and accuracy in designing such policies. Characterizing *equilibrium* data production in this environment therefore requires considering both the central government principal's goals and the strategic response of central and decentralized governments. These considerations are important for understanding the quality of administrative data as well as the implications of using these data.

6 Conclusion

In decentralized settings, national governments routinely collect data from decentralized entities for policymaking and oversight.[20] Our results show that decentralized entities report (or fail to report) data in anticipation of how the data will be used. Strategic reporting, in turn, limits the quality of the data, and therefore the legibility of decentralized entities to the central government. This suggests that the way central governments design data collection systems and communicates the use of these data to the bureaucrats who report on behalf of their entities can have important consequences for the quality of administrative data.

We study these dynamics in the case of Colombia's central government efforts to monitor compliance with transparency regulations using self-reported data compiled in the ITA matrix. While our empirical setting is specific, we believe the underlying mechanisms travel beyond the

[20]For example, in the United States, frequent federal requests for local data led to the Paperwork Reduction Act of 2005, which limits burdensome data requests on decentralized governments. Similarly, international organizations such as the World Bank's International Comparison Program rely on country-reported GDP figures to construct purchasing power parity (PPP) values.

Colombian case. The strategic behavior we theorize—where decentralized actors shape what they report based on expectations of oversight—is likely to arise in any context where agents know that their reports may be used to monitor or sanction them. These dynamics are common across decentralized governance systems and even outside the public sector.[21] Moreover, as we show in Appendix A1.1, Colombia represents a valuable—but not extreme—case for studying national-subnational oversight. It is neither a "hard case," where oversight is absent or entirely ineffective, nor an "easy case," where strong institutional capacity of the central government leads to exaggerated reactions in reporting behavior. Rather, it combines moderate decentralization, active oversight institutions, and persistent variation in transparency practices and corruption, making it a compelling setting to observe dynamics that may generalize to other decentralized democracies.

Our work opens two important avenues for future study. First, understanding how central governments design existing data collection processes and uses of data should be studied more systematically to broaden our findings from a single data collection process. Our qualitative findings suggest considerable uncertainty how to design these processes by national government officials and variation in the extent to which these processes are understood and interpreted by decentralized governments. These observations indicate a need for the study of more data collection processes to build upon our findings from Colombia. Second, we argue that the study of state legibility should be broadened to incorporate collection of information about a state's agents and organizations, not only its citizens. Our approximation of central and decentralized governments as unitary actors obscures some of the dynamics involved in state data production. However, agency problems likely exist *within* both national and local governments, shaping how data collectors and reporters internalize the incentives we describe. Theoretical advances will facilitate understanding of the relationship between these overlapping agency problems to clarify our understanding of the data generation processes that produce administrative data.

[21] Our findings thus speak to broader concerns about how multilevel oversight affects the quality of internally produced performance data.

References

Alesina, Alberto, and Alex Cuckierman. 1990. "The Politics of Ambiguity." *Quarterly Journal of Economics* 105 (4): 829–850.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–1178.

Baliga, Sandeep, and Tomas Sjöström. 1998. "Decentralization and Collusion." *Journal of Economic Theory* 83: 196–232.

Bardhan, Pranab, and Dilip Mookherjee. 2000. "Capture and Governance at Local and National Levels." *AEA Papers and Proceedings* 90: 135–139.

Bowles, Jeremy. 2020. "The Limits of Legibility: How Distributive Conflicts Constrain State-Building." Working paper, available at `https://static1.squarespace.com/static/5d2610dac406240001ee7541/t/621443111ec98b55d4eae905/1645495060907/draft_10.pdf`.

Callen, Michael, Saad Gulzar, Ali Hasanain, Muhammad Yasir Khan, and Arman Rezaee. 2020. "Data and policy decisions: Experimental evidence from Pakistan." *Journal of Development Economics* 146: 102523.

Cochran, William G. 1968. "Errors of measurement in statistics." *Technometrics* 10 (4): 637–666.

Cook, Scott J, and David Fortunato. 2022. "The Politics of Police Data: State Legislative Capacity and the Transparency of State and Substate Agencies." *American Political Science Review* pp. 1–16.

Dasgupta, Aditya, and Devesh Kapur. 2020. "The Political Economy of Bureaucratic Overload: Evidence from Rural Development Officials in India." *American Political Science Review* .

De Groot, Hans. 1988. "Decentralization Decisions in Bureaucracies as a Principal-Agent Problem." *Journal of Public Economics* 36: 323–337.

Eckhouse, Laurel. 2022. "Metrics Management and Bureaucratic Accountability: Evidence from Policing." *American Journal of Political Science* 66 (2): 385–401.

Edmond, Chris. 2013. "Information manipulation, coordination, and regime change." *Review of Economic studies* 80 (4): 1422–1458.

Faletti, Tulia G. 2005. "A Sequential Theory of Decentralization: Latin American Cases in Comparative Perspective." *American Political Science Review* 99 (3): 327–346.

Fan, C. Simon, Chen Lin, and Daniel Treisman. 2009. "Political Decentralization and Corruption: Evidence from around the World." *Journal of Public Economics* 93 (1-2): 14–34.

Ferraz, Claudio, and Frederico Finan. 2008. "Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes." *The Quarterly journal of economics* 123 (2): 703–745.

Garfias, Francisco, and Emily A. Sellars. 2021. "Fiscal Legibility and State Development: Theory and Evidence from Colonial Mexico." Available at `https://www.dropbox.com/s/g06yaf7ib7m6u2t/FiscalLegibilityStateDevelopment.pdf?dl=0`.

Grajalez, Carlos Gómez, Eileen Magnello, Robert Woods, and Julian Champkin. 2013. "Great moments in statistics." *Significance* 10 (6): 21–28.

Huber, John D., and Nolan McCarty. 2004. "Bureaucratic Capacity, Delegation, and Political Reform." *American Political Science Review* 98 (3): 481–494.

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the firm: Managerial behavior, agency costs and ownership structure." *Journal of Financial Economics* 3 (4): 305–360.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76: 1071–1102.

Lee, Melissa M., and Nan Zhang. 2016. "Legibility and the Informational Foundations of State Capacity." *Journal of Politics* Journal of Politics (79): 1.

Lee, Melissa M, and Nan Zhang. 2017. "Legibility and the informational foundations of state capacity." *The Journal of Politics* 79 (1): 118–132.

Lorentzen, Peter. 2014. "China's strategic censorship." *American Journal of political science* 58 (2): 402–414.

Martínez, Luis R. 2022. "How Much Should We Trust the Dictator's GDP Growth Estimates?" *Journal of Political Economy* 130 (10): 2731–2769.

McCubbins, Matthew D., and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28 (1): 165–179.

Mookherjee, Dilip. 2006. "Decentralization, Hierarchies, and Incentives: A Mechanism Design Perspective." *Journal of Economic Literature* (367-390).

Oates, Wallace E. 1999. "An Essay on Fiscal Federalism." *Journal of Economic Literature* 37 (3): 1120–1149.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63 (3): 581–592.

Sánchez-Talanquer, Mariano. 2020. "One-Eyed State: The Politics of Legibility and Property Taxation." *Latin American Politics and Society* 62 (3): 1–43.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

Slough, Tara. 2023. "Phantom Counterfactuals." *American Journal of Political Science* 67 (1): 131–153.

Slough, Tara. 2024. "Oversight, Inequality, and Capacity." Working paper, available at `https://taraslough.com/assets/pdf/oci.pdf`.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Tommasi, Mariano, and Federico Weinschelbaum. 2007. "Centralization vs. Decentralization: A Principal-Agent Analysis." *Journal of Public Economic Theory* 9 (2): 369–389.

Trinh, Minh. 2021. "Statistical Misreporting Debilitates Authoritarian Governance." *Working paper* .

Wallace, Jeremy L. 2016. "Juking the stats? Authoritarian information problems in China." *British Journal of Political Science* 46 (1): 11–29.

Willis, Eliza, Christopher da C. B. Garman, and Stephan Haggard. 1999. "The Politics of Decentralization in Latin America." *Latin American Research Review* 34 (1): 7–56.

World Bank. 2011. Managing a Sustainable Results Based Management (RBM) System. Get note World Bank Washington, D.C.: . https://openknowledge.worldbank.org/handle/10986/10450.