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

Natalia Garbiras-Díaz† Tara Slough‡

March 4, 2025

Abstract

Effective governance requires reliable information. In modern states, bureaucracies play a central role in the production of administrative data, with large volumes coming from decentralized entities' reports to the central government. However, governments' use of these data for policymaking and oversight creates incentives for bureaucrats to misreport, reducing the legibility of decentralized government actions and outputs. We examine this problem of *inter-institutional legibility*. We ask whether central government oversight can improve the reliability of bureaucratic reporting. Partnering with a national watchdog agency, we study this question in the context of Colombia's National Transparency Index and show that an exogenous shift in oversight salience changes both which entities report and their reported scores. Comparing self-reports to an independent audit, we find that true transparency practices correlate with three key measurement pathologies in the data: selection, distortion, and noise. While oversight marginally improves data reliability, limited inter-institutional legibility remains a governance challenge.

*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. Corresponding author.

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

Governance—or "seeing like a state"—requires reliable information (Scott, 1998; Callen et al., 2020).[1] Data enable states to manage populations, allocate resources, and enforce policies (Lee and Zhang, 2017; Garfias and Sellars, 2021; Callen et al., 2020), and for millennia, states have sought to collect such information through censuses, land cadasters, and administrative records (Grajalez et al., 2013; Brambor et al., 2020*a*). In modern states, bureaucrats play a crucial role in the production of administrative data, with large volumes of these data relying on reports from bureaucrats in decentralized entities—local governments or subsidiary national government agencies—to the central government.[2]

However, relying on these 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.[3] We argue that within these complex organizations characteristic of modern states, the legibility of decentralized bureaucratic agents and organizations—which we term *inter-institutional legibility*—is a central, yet understudied issue. While effective governance relies on ensuring the credibility of information from decentralized bureaucracies, how can governments incentivize accurate and complete reporting by strategic bureaucrats? In this paper, we study the role of government oversight.

We begin by developing a framework that links bureaucratic behavior 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 perceive that their responses may draw oversight attention from the central government with

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

[2]Earlier instances of bureaucratic data collection exist. For example, *hojeok* household registers in Korea's Joseon Dynasty relied on bureaucrats to verify information (Kuentae, 2007). However, the volume and types of data collected or reported by bureaucrats have expanded significantly since these early efforts.

[3]For instance, Eckhouse (2022) documents how U.S. police departments manipulate crime statistics in response to performance metrics, reclassifying or downgrading offenses to improve reported outcomes.

the possibility of enforcement, they may change their reporting behavior in an effort to deter this unwanted attention and potential punishment. Optimal reporting behavior from the perspective of decentralized entities, therefore, can introduce measurement error, ultimately limiting the quality of data and legibility of decentralized governments to the central government.

We test the idea that decentralized entities' reporting decisions are sensitive to their perceptions of the oversight process in the context of the production of Colombia's National Transparency Index (ITA), a measure of transparency practices across public entities. Specifically, we partner with the Office of the Attorney-Inspector General (PGN), a national-level watchdog entity that collects and compiles ITA annually from 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 obligation to report. 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 enforcement capabilities over ITA compliance. This experimental variation allows us to examine how increased salience of oversight and the possibility of enforcement affect data submission and reported scores.

After the intervention, and outside the scope of our partnership with the PGN, we also conduct an independent audit of a subset of items in the index to approximate a true latent measure of transparency practices at the entity level. We compare reported transparency practices to the independent audit-based measure to characterize the relationship between bureaucrats' reports and true levels of transparency practices. Collectively, 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.

We present several findings that highlight the limits of inter-institutional legibility. First, in the experiment, while increased oversight salience leads to slightly higher reporting rates, this effect is small and statistically indistinguishable from zero. Instead, entities report lower levels of com-

pliance with transparency practices when oversight is more salient.[4] Using a novel decomposition, we break down this difference in reported scores into two effects: (1) changes in the reporting behavior of entities that always report regardless of their perceptions of PGN oversight, and (2) changes in the composition of entities that choose to report because of direct communication about oversight. First, we find that, on average, treated always-reporting entities report lower scores than their control counterparts (estimated using Lee bounds). Second, entities that select into reporting due to direct communication report lower average scores than those that always report.

Next, our audit of public sector entities' transparency practices reveals the extent of variation in bureaucratic reporting behavior. Notably, we document that true (latent) 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 distort their reported scores, and (3) low-performing entities exhibit greater variance in their reported scores, which, in our framework, proxies for non-systematic measurement error. We further show that, contrary to common explanations of poor data quality, these pathologies are not merely a function of administrative state capacity. Instead, we find that they persist across all levels of capacity, indicating that strategic misreporting is a broader challenge of bureaucratic incentives. Taken together, these 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.

We make several contributions. First, we contribute to the literature on state legibility. Traditional accounts emphasize how governments collect data on individuals, households, and economic activity to enhance taxation, resource allocation, and administrative control (Scott, 1998; Lee and Zhang, 2017; Sánchez-Talanquer, 2020).[5] Recent research links citizen legibility to state development and governance outcomes, showing how variation in data-collection institutions shapes

[4]The ITA index goes from 0-100, with higher scores representing more compliance with transparency practices.

[5]Studies on the properties of censii, cadasters, and vital statistics include Mikkelsen et al. (2015), O'Hare (2019), among others.

policy implementation (Bowles, 2020; Brambor et al., 2020*b*). Our study expands this literature by introducing inter-institutional legibility, an understudied but crucial aspect of intergovernmental relations. We provide a framework demonstrating how governments' reliance on bureaucratic self-reports introduces risks of strategic misreporting and, empirically, show that data pathologies are closely tied to the phenomenon the state seeks to measure. These findings not only broaden our understanding of state informational capacity but also highlight important caveats for policymakers, scholars, and practitioners relying on administrative data.

Second, we contribute to the bureaucratic politics literature on principal-agent problems. While existing work extensively documents information asymmetries between bureaucrats and politicians within a single agency, less attention has been paid to inter-institutional principal-agent dynamics, where higher-level bureaucracies oversee decentralized entities with their own informational advantages (Gailmard and Patty, 2012; Prato and Turner, 2022). Our study shifts the focus to inter-institutional monitoring, where decentralized government units strategically report data to the center. We show that oversight shapes what bureaucracies report rather than whether they report, underscoring the broader governance challenges posed by multilevel informational asymmetries.

Third, we contribute to research on strategic misreporting of government data. Existing studies primarily focus on autocratic regimes, where bureaucrats manipulate economic data to align with political incentives, often due to career concerns or regime survival strategies (Guriev and Treisman, 2019; Martínez, 2022; Trinh, 2021; Lorentzen, 2014; Wallace, 2016; Edmond, 2013). Recent work documents data manipulation in democratic settings, particularly in politically sensitive domains such as crime reporting by U.S. police departments (Eckhouse, 2022; Cook and Fortunato, 2022). We substantially expand these scope conditions by showing that even routine government reporting—such as transparency compliance data—is subject to manipulation. Our findings challenge the assumption that high-quality data is an inherent feature of democracies, demonstrating that bureaucratic incentives to misreport extend beyond politically sensitive measures.

Finally, we make two applied methodological contributions to the study of selection. First, by

pairing an original audit with state data (with some non-reporting), we are able to gauge how of missingness in state data distorts the Colombian government's aggregate understanding of transparency practices in decentralized entities. Second, in our experimental design, we derive new bounds (building on Lee 2009) to decompose the sources of changes in reported data, distinguishing between 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 Information as a bureaucratic output

Prior to enumerating our account of bureaucratic information sharing, it is useful to consider the ultimate output that we observe: administrative data. Suppose that decentralized entities are tasked with reporting some measure of the quality of their performance—whether public service outputs, budget execution, or compliance with regulations or policy objectives in a sector—to the central government. A bureaucrat (or office) within the decentralized entity determines whether to comply with the request for information by making a report or declining to submit information. We will denote a non-report by $r = \emptyset$.

When the bureaucrat reports the quality of performance, their report, $r \in \mathbb{R}$, is a function of true quality, as well as intentional and unintentional errors or distortions. The true quality of performance is represented by the parameter $\theta \in \mathbb{R}$. A bureaucrat within an entity may choose to *intentionally* misreport quality, by reporting performance of $\theta + d$, where $d \in \mathbb{R}$ captures the intentional distortion. There may also be unintentional errors in reporting. These errors could be misunderstanding of questions, typos, or failure to correctly follow directions. We represent these

errors as $\varepsilon \sim f(\cdot)$, where $f(\cdot)$ is a mean-zero density.

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

The expression in (1) follows directly from conventional expositions of measurement error and missingness in statistics (Cochran, 1968; Rubin, 1976). In terms of measurement error, d and ε capture systematic and non-systematic measurement error. Non-reports (denoted $r = \emptyset$) manifest as missing data.

1.2 Data production

We now focus on the decision of decentralized government entities to report data to the central government. The actors that we study are therefore officials within the government entities tasked with data reporting. These officials are generally bureaucrats. Our decision-theoretic framework is premised on several assumptions about these bureaucrats' incentives to report. We maintain the notation used in (1). Without loss of generality, we will assume that the central government prefers higher values of the true quality, θ.

First, we assume that the central government can use the reported data, r, to target some type of enforcement (e.g., sanctioning noncompliance with a law) or a data validation exercise. We parameterize the probability that the central government targets an entity for further investigation or validation as $\rho(r) \in [0, 1]$. We do not make any further assumptions about the functional form of $\rho(r)$. If $\rho(r)$ were equivalent for all r, then the likelihood of being audited would be independent of the reported data.

Second, we assume that there is some penalty that can be imposed on entities in the course of targeted audits on the basis of the information that is uncovered. Audits provide some additional information about the true quality or state, θ, that the central government seeks to measure through reports. We assume that the size (magnitude) of the penalty imposed $P(\theta, r) > 0$, may vary in

true quality (θ), the reported data (r), and/or the difference between these measures. While we do not specify the precise functional form of P, it is highly plausible that the penalty is set to punish poor performance (i.e., low θ) or distortions in the reported data (i.e., an increasing function of the distance between r and θ).

Finally, we assume that collecting, collating, entering, and reporting data demands that bureaucrats exert costly effort. We parameterize effort as $e \geq 0$, and the cost of effort as $c(e)$ where $c'(e) \geq 0$. If a bureaucrat chooses not to report data, then $e = 0$. When a bureaucrat reports data, we assume that increased effort reduces the extent of idiosyncratic error, ε, formally $\frac{\partial Var(\varepsilon|e)}{\partial e} < 0$, for $e > 0$. Empirically, the cost of effort likely varies substantially across bureaucracies as a function of administrative capacity, which includes the human capital of bureaucrats, resources available for this type of data collection and reporting, and access to technology.[6] There may be other sources of variation in the cost of effort beyond capacity insofar as a given data-reporting task might be more difficult for some types of organizations or functions than others.

These three terms enter the bureaucrat's utility function in (2). In formulating the bureaucrat's utility in this way, we assume that the bureaucrat internalizes any penalty applied to their entity through the $P(\theta, r)$ term. It may be the case that a bureaucrat is punished for providing faulty data or failing to report. Further, oversight activities even at high-performing entities may impose cumbersome additional administrative work upon bureaucrats. It is important to note that bureaucrats may not know precisely $\rho(r)$ or $P(\theta, r)$; in these cases, what matters is their beliefs about these policies. Our experimental design targets these beliefs of decentralized bureaucrats. While our central focus is on intergovernmental oversight, it is important to note that variation in $c(e)$ should also shape reporting behavior, a point to which we will return in Section 5. It could be the case that data is used principally to target resources to an institution. Such resources are not relevant in

[6]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.

the empirical case we describe, one could add a benefit term to the utility function in (2).

$$U_B(r, e; \theta) = -\rho(r)P(\theta, r) - c(e) \tag{2}$$

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 given. In Section 6, we revisit equilibrium considerations after presenting our empirical findings.

1.3 Measuring the quality of administrative data

Our simple framework of data production guides our assessment of the quality of administrative data. While data is shaped by bureaucrats' decisions to exert effort (e) and to distort their reports (d), neither behavior is directly observable to the central government or the analyst. Instead, both observe reports, r, which are a function of both behaviors, as clarified in (1).

Enhancing oversight: What is the effect of a shock to anticipated oversight over data, formalized by $\rho(r)P(\theta, r)$? Without specifying functional forms—which would likely vary across data collection processes—(2) does not generate unambiguous testable predictions. However, it does allow us to identify a set of mechanisms through which reporting behavior might respond to greater (perceived) oversight.

Consider first the government's monitoring rate: $\rho(r)$. This function describes how the central government uses reports to target oversight. If non-reports (i.e., $r = \emptyset$) are subject to additional scrutiny, enhanced oversight might induce otherwise non-reporters to exert effort to complete reports. Moreover if low scores are targeted by the government, entities may respond by exaggerating reported scores to pool with higher performing entities by choosing some $d > 0$.

Now, consider the penalty that might be imposed, whether for failure to report, misreporting, or poor performance, $P(\theta, r)$. Anticipated penalties for non-reporting could induce bureaucrats

in marginal entities to exert the effort sufficient to report. Penalties for misreporting could induce bureaucrats to work harder to avoid unintentional errors (since Var(ε) is decreasing in e) or reduce the magnitude of misreporting (i.e., reduce $|d|$). Finally, penalties imposed for poor performance—a characteristic that is not manipulable in the short-run by bureaucrats or their organizations—could reinforce incentives to avoid scrutiny as discussed above.

Overall, this analysis suggests that increasing oversight should impact reporting behavior by bureaucrats. While our model clarifies a set of mechanisms that produce this effect, it suggests that these mechanisms can produce different effects under different institutional settings (i.e., different formulations of oversight).

Describing aggregate reporting behavior: What could be learned about reporting behavior if we could measure θ through means other than reports by bureaucrats? While θ is often inaccessible to national governments (or at least prohibitively costly to obtain at scale through other means), it allows for additional learning about bureaucratic behavior. At the level of the individual observation, it is, of course, not possible to observe learn d or ε, since $r - \theta = d + \varepsilon$. However, given our assumption assumption that $E[\varepsilon] = 0$ and independent of d, we can measure distortions in the aggregate by measuring $E[r - \theta]$. With measures of both r and θ, we suggest that three quantities are informative about bureaucratic reporting behavior:

- First, examining selection into reporting as a function of θ provides information on the relationship between observed reports and true quality in the aggregate. Within our simple framework, if selection into reporting varies systematically in θ, it suggests that either the cost of effort varies in θ or reporting bureaucrats anticipate varying levels of scrutiny or oversight as a function of their reports, θ.
- Second, one can measure the aggregate distribution of intentional distortions in reported data, as a function of θ by estimating $E[r - \theta|\theta]$. This provides our best summary of d across bureaucrats/entities. Identifying intentional distortions in the aggregate provides evidence that bureaucrats perceive that oversight from the national government depends on

their reporting behavior. In principle, intentional distortions are used to hide from scrutiny by pooling with other entities that are less likely to be scrutinized. If national governments use scores to target scrutiny, then we should observe variation in misreporting as a function of θ.

- Finally, one can examine how effort varies in θ by measuring the conditional variance of reports as a function of θ, e.g. $\text{Var}[r|\theta]$. Here, the idea is that lower effort corresponds to more drastic unintentional errors. These unintentional errors manifest in the data as higher variance in reports. This provides our most direct measure of bureaucratic effort, though how effort should relate to θ is ultimately an empirical question.

2 Case Context

Colombia is the most populous unitary state in the Americas. As such, our focus is on the central government's collection of data from (generally) decentralized government entities. Deepening of Colombia's fiscal, political, and administrative decentralization in the 1980s and 1990s increased efforts by the central government to collect data at the local level to monitor the delivery of national-government funded public goods and services (World Bank, 2011). Our discussion of the case context is informed by our discussions with our partner, the PGN, and semi-structured interviews with bureaucrats who submit data to the national government.[7]

Like many other national governments, the Colombian government relies heavily on self-reported data from territorial governments to inform policy making and target monitoring. One secretary of planning in a small municipality complained: "Data requests from [the national government] take so much time to complete. For instance, some entities hire people just for the purpose of filling out all such forms, but others that are smaller, are bound by law and cannot hire external contractors to do so. This means we have to do it with our own resources."[8] Despite some efforts to consolidate these tasks, data collection, collation, and submission remains a central task

[7] See Appendix A5 for discussion of our sampling strategy for these interviews.
[8] All translations by authors.

of decentralized bureaucrats in Colombia. In an original survey of bureaucrats in Colombian municipal governments (*alcaldías*), for example, Slough (2024) finds that 48% of local bureaucrats report meetings or calls with national agencies in the past week, totaling an average of 2 hours. Moreover, these bureaucrats spend a majority of their time (53%) completing administrative tasks like reporting, monitoring, and evaluation, rather than tasks more directly associated with service provision (e.g., field visits or interfacing with citizens).

Our focus on official data as a bureaucratic output is distinct from a recent focus on service provision by bureaucrats in low- and middle-income countries (Pepinsky, Pierskalla, and Sacks, 2017; Grossman and Slough, 2022). The secretary of planning in the previous paragraph suggests that some entities may allocate tasks (e.g., data collection and service provision) to different officials or hire contractors to alleviate pressures to produce data. In other entities, national government requests for data may overburden bureaucrats, leading to tradeoffs or poor implementation in one or more domains (Dasgupta and Kapur, 2020).[9] By emphasizing data production, we complement existing discussions of bureaucrats as service providers, shedding light on an understudied bureaucratic task with significant implications for governance and resource distribution.[10]

2.1 Transparency in Colombia 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. ITA was mandated by Colombia's transparency and open data law (*Ley 1712 de 2014*).[11] The *Procuraduría General de la Nación* (PGN), Colombia's principal watchdog agency under the Public Ministry, is responsible for ITA's implementation. This central government entity investigates

[9] Bureaucrats in local governments self-report working an average of 52.3 hours per week in the time-use surveys by Slough (2024), in contrast to Colombia's workweek of 48 hours at the time of the survey. This suggests that the average bureaucrat may be overburdened.

[10] 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."

[11] See Appendix A1 for further information on transparency in Colombia.

and sanctions irregularities or misconduct by publicly elected officials, public servants, and public sector agencies. The PGN is widely recognized among Colombian bureaucrats, even at the local level. Multiple interviewees recall that all public servants must complete mandatory training administered by the Administrative Department of Public Service (DAFP, per its Spanish acronym), which explains the structure of the state and, importantly, the PGN's role and oversight functions.

The PGN collects ITA data as part of its preventive mandate to monitor public officials and entities, aiming to reduce corruption and other public misconduct. By collecting data, the PGN seeks to identify entities at higher risk of wrongdoing. According to discussions with PGN officials, ITA data are used to direct preventative efforts. Importantly, the PGN also initiates disciplinary proceedings against entities that, upon investigation, fail to comply with transparency and anti-corruption laws.

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 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 transparency practices specified in the questionnaire, while 0 reflects non-compliance with these regulations. The PGN publishes these measures in a consolidated report, which compiles ITA data across entities.

Each year, the PGN fully delegates the request to complete the ITA to various national government agencies, referred to as "heads of sector." In practice, this means that entities receive the ITA submission request from a different entity. For example, 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. While the PGN states that these data inform preventative anti-corruption efforts, low response rates and unknown accuracy render reliance on ITA potentially problematic. These reporting pathologies create perverse incentives: entities that honestly disclose imperfect transparency practices may be penalized, while those that fail to report or falsely claim compliance can skirt oversight. Beyond these broad contours, the exact use of ITA data by the PGN remains unclear. Interviews with bureaucrats who submitted ITA data revealed similar perceptions among the actors we study. Nevertheless, to the extent that ITA is indeed used to guide enforcement, it may produce unintended consequences. Thus, understanding how these data are produced and their accuracy is essential.

3 Research Design

We conduct a pre-registerd field experiment in collaboration with the PGN, which sought to determine whether low-cost strategies could increase rates of complete data submission. To better understand the behavior of bureaucrats responsible for compiling ITA data, we emphasize the importance of descriptive quantities alongside the causal estimands targeted in the experimental design. Much can be learned about the production of ITA data from the relationship between actual transparency practices—measured through an independent audit—and reported compliance measures. These descriptive patterns are critical for interpreting and using the data effectively. Meanwhile, the causal effects estimate the extent to which changes in bureaucratic incentives alter reporting patterns to the PGN.

3.1 Sampling

Our unit of assignment is the entity or organization. Our experimental sample includes the near-universe of public sector entities in Colombia (99%). When sampling entities for the audit, we stratify by national versus territorial (decentralized) entities and oversample national entities due to their relative low incidence. Table 1 provides details on the population of entities, the experimental sample, and the audited sample.

	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 public sector entities that were randomly sampled and used in a piloting pre-test of intervention implementation.

3.2 Intervention and Assignment

We conduct an experiment with two levels of treatment. The first level examines the effects of increased oversight salience by the PGN on officials' data reporting behavior. Our primary manipulation emphasizes direct communication from the PGN to entities. In the status quo—and thus our control condition—the PGN delegates data requests to sector heads, national entities responsible for transmitting the request. Indirect communication from sector heads typically consists of social media posts and other online messaging. To increase the observability of the PGN's role in data collection, we randomly assign some entities to receive a direct email from the PGN requesting the data. Thus, our first-level treatment assignment compares the status quo—delegation to sector heads—against a combination of delegated communication *and* direct communication from the PGN. Direct communication from the PGN heightens the perception that responses may be scrutinized and that non-compliance could result in punitive action.

Interviews with bureaucrats who submitted 2020 ITA data on behalf of their entities suggest bureaucrats' thought process closely resembles our link between direct communication and increased salience of oversight. For example, an official at a public university stated: "There can exist sanctions, surely, 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 there can be imposed for those who either do not complete the form or fill it out inaccurately." Our communication of the PGN's role in collecting and using the data seeks to reduce this uncertainty. These observations suggest that the direct communication treatment can be interpreted as a shock to perceived oversight.

Within entities randomly assigned to receive direct communication from the PGN, we subtly vary the content or frequency of the messages using a $2 \times 2 \times 2 \times 2$ factorial design. Table 2 summarizes the variation in message content, and Table A2 provides the full text of the messages. We refer to these second-level treatments as "nudges." We follow 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 informs bureaucrats about the PGN's role and use of the data, thereby shifting incentives related to data submission, the nudges consist of additional sentences within these communications.[12] While our nudges are motivated by our theoretical framework, they are substantially weaker interventions than the top-level randomization of direct messages.

The nudges serve both practical purposes for the PGN and analytical benefits for our study. Direct messaging is costly for the PGN, as it requires staff time and expertise to tailor communications and respond to an increased volume of inquiries. In contrast, modifying the content of these emails is costless once they are being sent. Understanding 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

[12]The reminder treatment instead varies the frequency and timing at which the message was received.

2. From a design perspective, one might be concerned 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 entities in our experimental sample. First, we stratify entities based on ITA completion in 2019, creating two subgroups to ensure exact blocking on past completion status. Within each subgroup, we formulated blocks of 18 entities that minimize Mahalanobis distance between covariates using (1) PGN's classification of organizational or entity type and (2) department indicators. This means that within each block of 18, all entities are identical in 2019 ITA completion behavior. For instance, the Mahalanobis distance minimization ensures that local governments in the department of Antioquia are most likely to be in the same block as other local governments in Antioquia, etc. Within the blocks, we randomly assign two entities to a pure

control condition and the other 16 entities to each cell in the $2 \times 2 \times 2 \times 2$ factorial. This means that $\frac{8}{9}$ of subjects receive some form of direct communication. We report balance on observable covariates in Figure A2. Figure 1 summarizes the experimental design graphically.

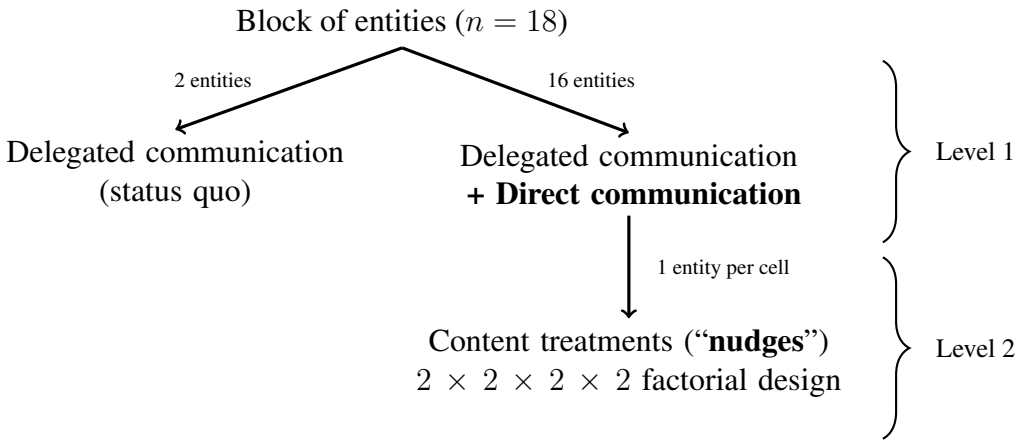

Figure 1: Treatment assignment scheme within each block of 18 entities.

3.3 Independent Audit of Data Quality

One of the central features of our research design is the independent audit of compliance with a subset of ITA items. The audit evaluates approximately 200 transparency practices, primarily related to the online publication of information. These binary items are reweighted and summed to form a 100-point scale. We audit 27.75 points of this 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.[13] Auditors were trained to search for selected ITA index components using a standardized process. They recorded compliance with each item, along with subjective assessments of quality and ease of access. We describe the audited items in Appendix A4. Our measure of latent quality is constructed from the indicators of compliance with each ITA index item.

Crucially, we conduct the audit in parallel for entities that reported and entities that failed to report in ITA data collection. Given the large number of entities in our study and the time requirements of the audit, we restricted the audit to 2,400 public sector (traditional) entities, a

[13]This audit was conducted completely independently of the PGN.

stratified random sample of 200 national and 2,200 decentralized entities. This sampling into the audit oversamples national entities, so we include indicators for national versus decentralized entities throughout the analysis of the audit data. In Tables A4-A5, we show that, conditional on this indicator, assignment to the independent audit is balanced across past (2018 and 2019) ITA submission and scores, as well as our treatments.

One potential concern is that, given complications identifying, contracting, and training the firm for this non-standard audit, too much time elapsed between the submission of ITA data and the audit six to seven months later. Improvements or reduction in transparency practices over this time are a source of measurement error in our measure of quality, θ. They should not bias our results, however, unless (1) entities became more transparent because of the treatment only after they submitted their data to the PGN; or (2) changes in transparency practices between the data submission and audit vary with the true level of transparency practices. In Figure A4, we show that experimental treatments do not have an effect on the underlying quality measure, allaying the first concern. To the second concern, our interviews suggest that, if anything, entities tailor their websites before—rather than after—submitting reports. Moreover, the PGN did not start to use the 2020 ITA data in oversight functions until the second half of 2021, after our independent 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: purposeful misrepresentation 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. We transform these scores to create two outcome measures. The first is a binary indicator that captures data submission to the PGN, taking the value "1" if an entity submitted data to ITA. The second outcome 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 27.75. We construct binary indicators of whether an entity complies with each audited item, based on both audit results and the entity's self-reported data. These indicators are then reweighted according to the index weights. Finally, 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 27.75.[14]

3.5 Identification and Estimation

The two-level randomization permits the identification of different estimands. The first level of treatment is a simple two-arm design that permits identification of the average treatment effect (ATE) of direct communication. In the second level of randomization, we estimate average marginal component effects (AMCEs) of each of the four factorial nudges through the content of those requests. We employ OLS to estimate these estimands using Equations 3. The estimator of

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

the ATE of direct communication is β_1 and AMCEs of message content are β_2, β_3, β_4, and 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. Note that in all complete blocks, there are at least two units in each treatment condition for each treatment indicator. The block indicators subsume past completion of ITA given our 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.

We further regress reported ITA scores on the experimental treatments using an estimator identical to (3). Because the sample for this outcome is conditioned on submission, the β'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 mechanisms we discuss in Section 6. 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. In some specifications we allow for higher-order polynomials and flexible specifications to characterize potential non-linearities in the associations between the these variables. We also reweight these specifications by the inverse of sample inclusion probabilities to account for the fact that national entities are overrepresented among the audited sample.

3.6 Ethical considerations

Our research design involves intervention in a government data-collection exercise. While this experiment was designed in consultation with and implemented by our partner, the PGN, two ethical concerns merit further discussion. First, the PGN did not seek informed consent from bureaucrats—all public officials—when implementing the experiment. Seeking consent would depart from their standard interactions with other government entities. Second, because ITA is used in Colombian state functions, intervening in its collection could present downstream social impacts or harms to the PGN or the subject entities. To limit this possibility, the treatments were designed in consultation with the PGN. This means that the PGN knows how the data were produced, and if we were to detect substantial changes in data quality, would be able to adjust their use of the data accordingly. At the very least, our use use of a status-quo control mitigates the possibility that creating a control group would *lower* response rates. We discuss these considerations at greater length in Appendix A3.

4 Results

4.1 Direct Communication from the PGN and Bureaucrats' Observed Reporting of Transparency Practices

How does increasing the salience of oversight change reporting behavior? In Table 3, columns (1)-(2), we report estimates of the ATE of direct communication and, in Panel B, the AMCEs of the nudge treatments on the probability of submitting ITA data. We find that direct communication increases the probability of reporting by 3 percentage points, though this increase is only marginally statistically significant ($\alpha < 0.1$) 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 is similarly not significant. Combined, however, these estimates suggest that a higher dosage of direct communication from the PGN increases rates of submission by 4.2 percentage points ($p < 0.037$ in a two-tailed test). The estimated AMCEs of the other nudge treatments are very near zero and are not significant.

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. Several factors may explain these modest effects; here, we explore one. Rates of reporting are already fairly high—65%—in the control group. It may be easier to induce reporting when initial compliance rates are lower. We test this hypothesis leveraging the substantial autocorrelation of responses between 2019 and 2020 ($\rho = .42$). While entities that did not complete the data submission in 2019 were 48% less likely to complete the 2020 version than their peers who completed the data submission in 2019, differences in the ATE and AMCEs between these subgroups are all near-zero and statistically indistinguishable from zero (Figure A6). This analysis further suggests that differences in rates of completion are not simply a function of awareness of a requirement to report. If this were the case, we might expect representatives of entities that did not report in 2019 to respond more strongly to the direct communication. We observe no evidence of this pattern.

	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.

In Columns (3)-(4) of Table 3, we regress scores, conditional on reporting, on the experimental treatments. This analysis conditions on reporting and is thus "post-treatment." The table suggests 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. While these coefficient estimates should not be interpreted as causal effects because of our sample conditioning on reporting, recall that the post-treatment estimand can be decomposed into a weighted sum of the conditional ATE (CATE) among "always reporters" and the average reported score among if-treated reporters (see Appendix A8).

With respect to direct communication (for example), a non-zero CATE implies that there exist entities that report different scores because they were contacted directly by the PGN than they would have if not contacted. The selection term consists of the expected score among entities that report when contacted by the PGN but would not report when they are not contacted directly.

Before reporting the decomposition of this post-treatment estimand, we evaluate the assump-

tion that selection into reporting is monotonic. In this context, monotonicity holds that there does not exist a subject who reported *because* they were not assigned to direct communication or who failed to report *because* they were assigned to direct communication. The assumption of monotonicity allows us to invoke Lee (2009) bounds to generate an interval estimate of CATE among always reporters. To validate this assumption, we use all pre-treatment covariates provided by the PGN to estimate heterogeneous treatment effects on selection into reporting using generalized random forests (Athey, Tibshirani, and Wager, 2019). In this analysis, we predict CATEs for all units in our sample. In Figure A8, we show that there are no units for which we can detect a negative treatment effect (at the $\alpha = 0.05$ level). In contrast, we estimate positive and significant treatment effects of direct communication on reporting for 886 of 6556 entities. This analysis supports our 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 send *lower* average scores when exposed to oversight through direct communication. Our interval estimate on the average scores of if-treated reporters is very 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 sample selection and changes in the scores reported by bureaucrats. We provide bootstrapping-based uncertainty estimates of the Lee Bounds in Table A7. 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. Even though we do not find evidence of average effects on ITA

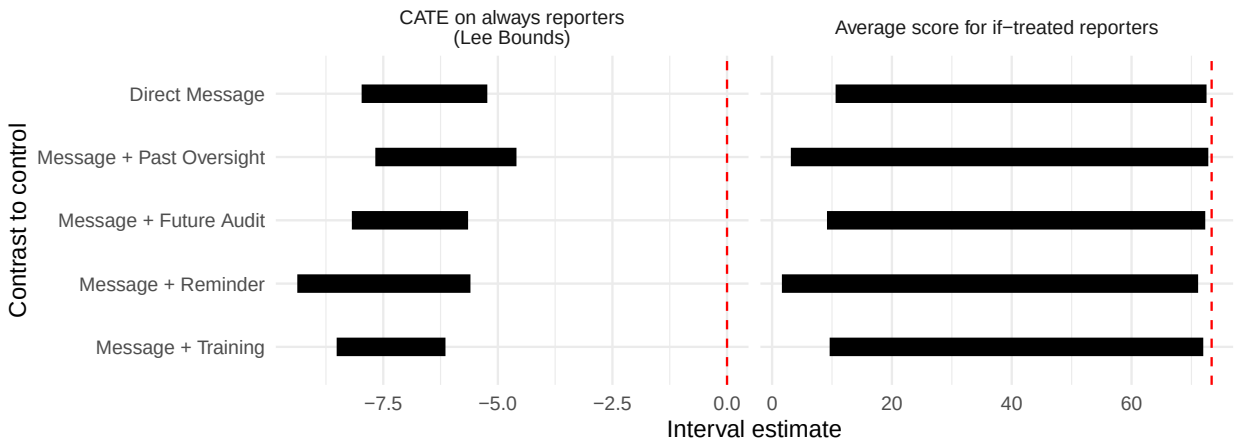

Figure 2: Decomposition of post-treatment estimands analogous to Column (3) of Table 3 (pooling over unstated message content) into a CATE on always reporters (left) and the average score among if-treated reporters (right). The CATEs are estimated using Lee trimming bounds and the interval estimate of average scores among if-treated reporters is calculated algebraically from those bounds, following Appendix A8. In the CATE plots, the vertical red line indicates a CATE of 0. In the plots depicting the average score among if-treated reporters, the red line indicates the average score among all reporters.

submission, we show that when exposed to oversight, some entities report lower scores than they would otherwise report. Where do these lower scores come from? Given our randomized design, entities' values of true quality, θ, should be independent of increased oversight. However, a limitation of the experimental data is that our measures do not, in isolation, provide evidence about the *accuracy* of reported scores because we lack a measure of θ. Thus, to explore whether data distortions explain, at least in part, the lower reported scores of treated facilities, we now turn to our analysis of the audit data.

4.2 Bureaucrats' Reporting Behavior and Latent Transparency

Our audit of a subset of entities provides an empirical measure of actual transparency practices, θ, for a subset of index components. Notably, we observe this audit-based measure regardless of entities' decision to report, as sampling into the audit was independent of reporting behavior.

1. Entities' selection into reporting: We first examine the propensity to report as a function of actual transparency practices. The left panel of Figure 3 plots the probability of completing the transparency index across the domain of our audit measure (formally, $\Pr(r \neq \emptyset|\theta)$). We find positive selection: reporting rates increase substantially with higher values of θ. Specifically, an entity with a score of zero on the audit metric reports with probability 0.49 (95% CI: [0.45, 0.52]), while an entity with a perfect score reports with probability 0.84 (95% CI: [0.82, 0.86]).

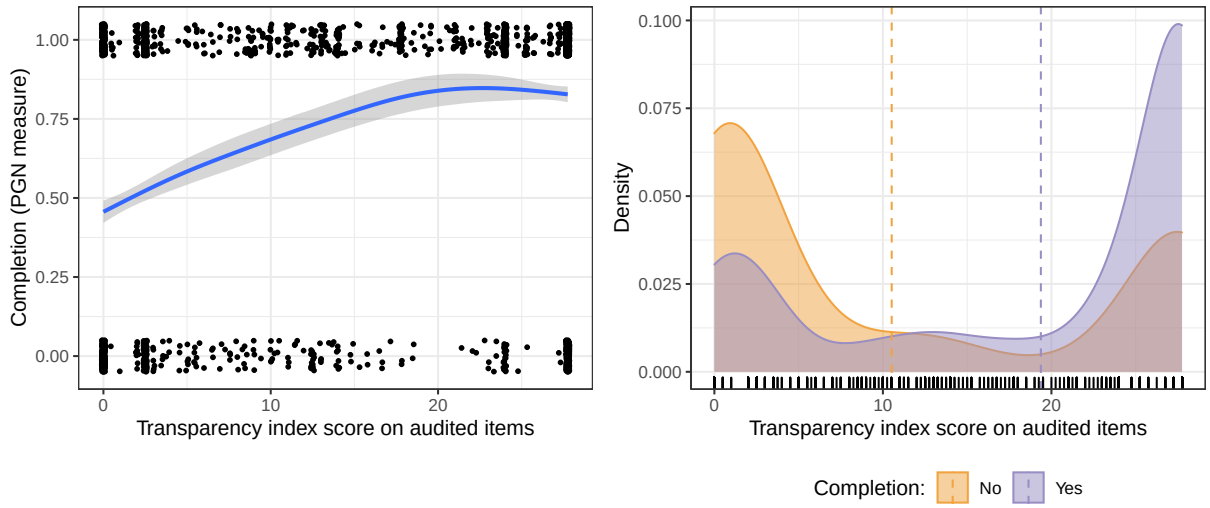

Figure 3: The association between the audit-measured transparency index and the probability of ITA data submission (left). The distribution of the audited-measured transparency index among entities that completed and failed to submit ITA data (right).

From the perspective of the PGN or another data analyst, this pattern of selective reporting yields scores on audited items that are distributed according to the purple conditional density in the right panel of Figure 3. The unreported scores are distributed according to the orange conditional density. The vertical lines denote the means of each distribution. The difference between these means (8.85 points) is equivalent to 0.74 standard deviations of the audit-based measure of transparency practices. As such, without considering selective reporting, aggregate summaries of ITA scores will substantially overstate the level of compliance with transparency practices.

2. Misreporting of Transparency Index Data: 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, which is post-treatment with regard to our experimental treatments. While we include the treatments as covariates in various regression specifications, the coefficients do not estimate well-defined causal effects. As such, our analysis of accuracy is purely descriptive.

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 27.75 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. This is not the case: we soundly reject the null hypothesis that $\beta_{\text{Audit}} = 0$ for both outcomes. 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 similarly reject a null hypothesis that $\beta_{\text{Audit}} = 1$ for both outcomes ($p < 0.001$ in all tests). This is unsurprising, but it does not allow us to decompose inaccuracy in reporting from the measurement error in the audit. To this end, we seek to measure both the extent of intentional distortions and noise in reporting.

We now examine the possibility of intentional misreporting (the parameter d in our model). Figure 4 explores the relationship between audit-measured transparency practices and self-reported transparency practices.

In the left panel, we plot our audit-based measure of transparency practices (θ) against the

	Reported score on audited items			Total reported score		
Audit score	0.509***	0.509***	0.504***	0.731***	0.733***	0.731***
	(0.025)	(0.025)	(0.025)	(0.073)	(0.073)	(0.075)
Intercept	9.886***	10.637***	10.535***	59.240***	64.274***	64.207***
	(0.607)	(0.804)	(0.809)	(1.767)	(2.433)	(2.455)
Num. Obs.	1307	1307	1307	1696	1696	1696
National Indicator	yes	yes	yes	yes	yes	yes
Experimental treatment indicators		yes	yes		yes	yes
Elected entity head indicator			yes			yes
Adjusted R^2	0.339	0.339	0.339	0.121	0.125	0.124

$^{+}$ 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.

difference between self-reported and audit-measured transparency practices on the same subset of items ($r - \theta$). The plotted 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. Since entities cannot over-report a perfect score or under-report a score of zero, it is important to assess whether these deviations are mechanical. To that end, the right panel examines the association between audit-measured transparency practices and the magnitude of any distortion. Here, we show that distortions are decreasing in true levels of transparency. Collectively, these plots suggest that bureaucrats tend to over-report compliance with transparency practices, but only at low and middling levels of transparency.

3. Noise in reporting: Our next analysis considers the magnitude of unintentional errors in reporting as a function of underlying transparency practices. Under the assumptions of our framework, greater variance in reported scores is indicative of lower effort devoted to reporting data. In Figure 5, we estimate the standard deviation in scores—both on the subset of audited items from the microdata and on overall scores—as a function of audit-detected quality. We show that the standard deviation (and thus variance) in scores is greater where transparency practices are weaker. This finding is apparent when examining the standard deviation within bins of audit-measured transparency practices (left) and when using a triangular kernel to estimate the conditional standard

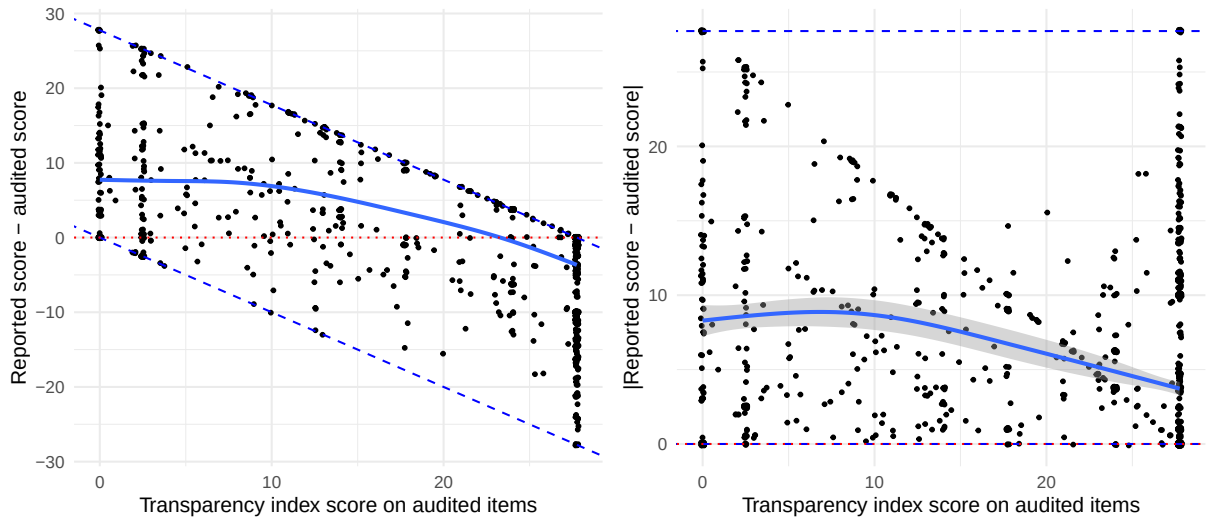

Figure 4: Discrepancies between the reported transparency practices and those detected in the audit.

deviation across the support of the audit measure.

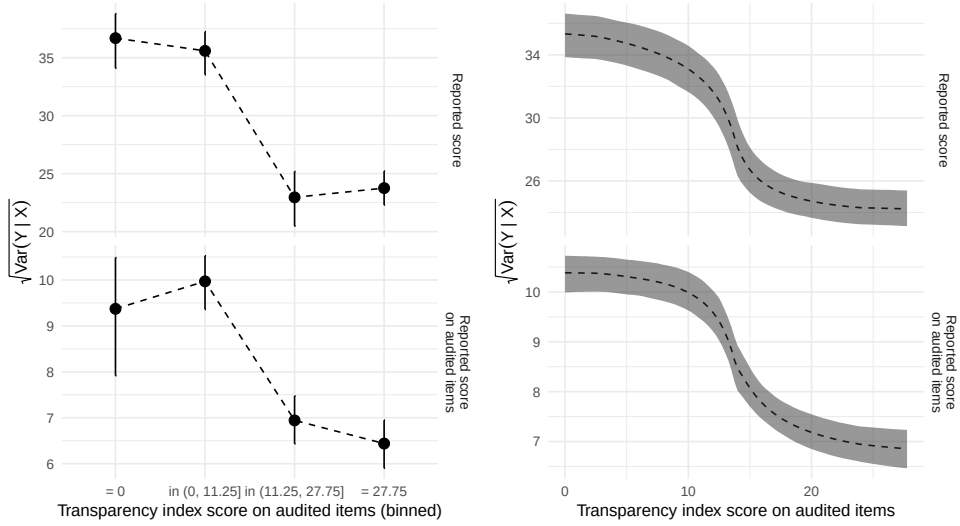

Figure 5: This plot shows how noise in reported ITA scores—measured by the standard deviation—relates to the audit-measured transparency index. The left panels bins entities by level of audit-measured transparency. The right panels employs a triangular kernel to estimate the conditional-standard deviation.

Several alternative explanations to limited effort are warranted. First, censoring of scores at 0

and 100 may mechanically lead to differences in variance as a function of scores, since institutions at these two modes in the data cannot under or over-report scores, respectively. However, if this were the case, we would expect the variance to be greatest in the middle of the distribution. We do not observe non-monotonicity in the conditional standard deviation. As such, censoring, in isolation, cannot explain the results in Figure 5.

4. Oversight: Our experiment shows that increasing the salience of oversight shapes reporting behavior, by changing *which* entities report and *what* entities report. We now bring all the pieces together, aided by audit-based measures, to examine whether entities' responses to oversight are conditioned by their true levels of transparency. 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. Faced with more information about the use of ITA scores, non-reporters with transparency practices around the mean of reporting entitites (19.37/27.75) increase their likelihood of reporting when subject to oversight (see Figure A10). This is consistent with our bounds in the right panel of Figure 2 that suggest that the average scores of if-treated reported are slightly lower than the overall sample mean.

In turn, the right-hand 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 mid-levels of true transparency practices (see Figure 4). Interestingly, we find that for entities with near-zero transparency, direct communication induces upward distortion. We interpret both results cautiously, since they condition 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.

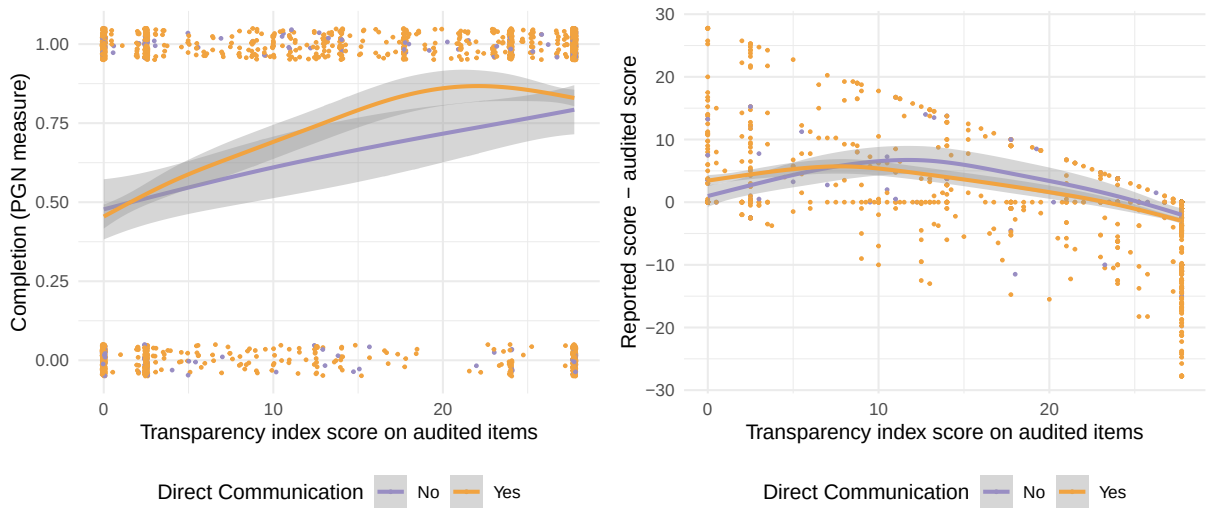

Figure 6: Conditional average treatment effects of oversight on the probability of ITA data submission at different levels of audit-measured transparency practices (left). Discrepancies between the reported transparency practices and those detected in the audit by treatment condition at different levels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

5 The Complementary Role of Administrative Capacity

Until now, our analysis has focused on entities' strategic responses to oversight by the central government when reporting data. However, according to our theoretical framework, an entity's administrative capacity should also influence its decision to report to the national government. Administrative capacity can reduce the bureaucrat's cost of effort, $c(e)$, and higher effort lowers the likelihood that unintentional distortions in reports (ε) are large in magnitude. Changes in effort may, in turn, affect both the choice to report and the decision to systematically distort scores.

There exists substantial variation in administrative capacity within the Colombian public sector. For instance, the national government generally has greater administrative capacity than most territorial entities. Public sector organizations draw from differently skilled labor pools, influenced by geographic location, wages, and tasks. Additionally, some bureaucracies have greater access to technology than others. Finally, variation in organizational structures and leadership may further shape administrative capacity across entities.

Using two measures of administrative capacity—one at the municipal level and one at the entity level (for a subset of entities)—Table A8 examines how reporting behavior varies in administrative capacity. We find that higher capacity entities (or entities in higher capacity municipalities) report at higher rates, distort their scores toward desired outcomes by a greater degree, but ultimately report with less average noise. The latter finding—on noise—is directly consistent with our assumption about how effort shapes reported scores.[15] Our finding that intentional distortions increase in administrative capacity suggests that capacity may facilitate this type of strategic misreporting. This finding, coupled with our experimental evidence that increases in perceived oversight change the reporting behavior of decentralized entities, suggests complementary roles for both administrative capacity and oversight in the production of administrative data.

6 Discussion

We have shown two central results in the context of Colombia's ITA data collection. First, the reporting behavior of decentralized entities responds to changes in communication of the PGN's role in data collection. Second, non-response and distortions in ITA data vary in the true (latent) level of transparency practices of these entities, 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 and that they had some uncertainty about how to pursue these goals.

In most equilibrium data collection processes, the central government should be viewed as a strategic actor. To extend our theoretical framework, an enforcement or control agency within the central government controls two policy instruments: the targeting of audits ($\rho(r)$) and the penalties imposed upon poor performance in an audit ($P(\theta, r)$), in addition to communication of

[15]The first finding, on completion rates, is also consistent with our interpretation about administrative capacity and the costs of effort if $Cov(\theta, c(e)) = 0$. When performance (θ) and the cost of effort ($c(e)$) covary, as suggested by Table A8, complementarities between quality and the cost of effort can also affect the probability of completion.

these policies. In this setting, governments can choose policies to influence reporting behavior, and thus shape the quality of the ultimate data they observe. As we have shown empirically, decentralized entities are likely to respond strategically to these policies, at least to the extent that they understand how the data is used.

In the present experiment, in contrast, we fix the central government's behavior by randomizing communication with decentralized entities to isolate the reporting behavior of decentralized bureaucrats. To this end, our experimental results measure partial equilibrium changes in the reporting behavior of bureaucrats. Nevertheless, our framework and audit data allows us to speculate about what the central government might uncover under different oversight strategies. In particular, we focus on $\rho(r)$, the targeting of oversight. While we do not know precisely the PGN's objective in its preventative oversight efforts based on the ITA data, two possibilities seem highly plausible. First, the PGN may seek to focus effort on entities with low levels of transparency practices (low θ in the model). Second, they may want to maximize the accuracy of the data (by minimizing $|r - \theta|$). Importantly, these seemingly-aligned goals—identifying the non-compliant entities and maximizing accuracy of data—might suggest the use of different policy instruments.

In Figure 7, we use the theoretical model to consider how the government might best use the ITA scores to target oversight, given our audit data. We consider entities that received the "direct communication" treatment that the PGN set out to study. Consistent with the results in Figure 3 and Table 4, auditing strategies that audit (i) entities reporting a zero score or (ii) non-respondent entities are best able to target low-transparency entities. In contrast, if the goal were to maximize data quality, an auditing strategy that audits low—but non-zero—scores with higher probability is apt to detect larger distortions in the data, consistent with Figure 4. Note that we may expect entities to respond differently over time as they learn about how data is being used by the central government.

While fixing the behavior of each actor may be useful analytically, it reveals how challenging these patterns may be to detect in administrative data. Indeed, if bureaucrats learned that the PGN

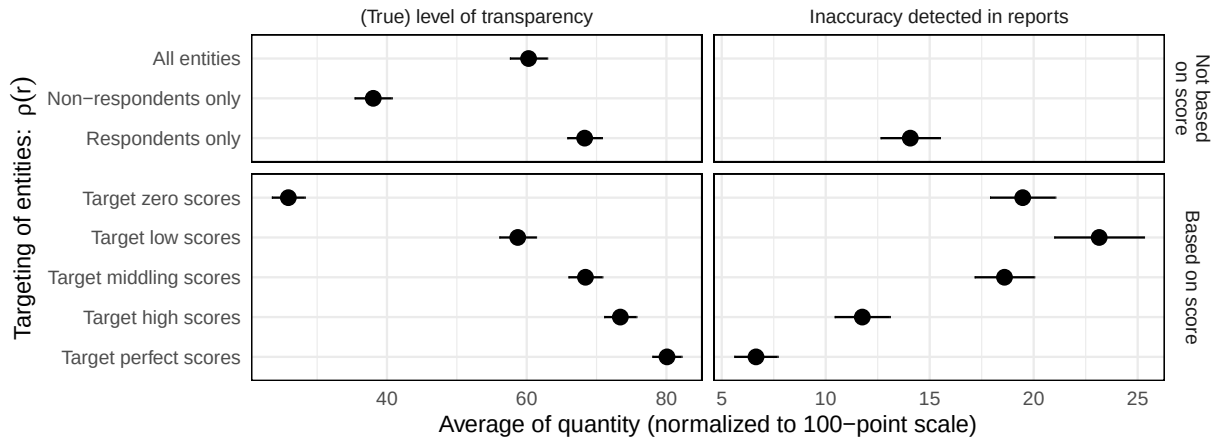

Figure 7: Average levels of θ (left) and $|r - \theta|$ right under various simulated oversight strategies, $\rho(r)$. The 95% confidence intervals correspond to a sample of 1,000 entities. Functional forms used for $\rho(r)$ appear in Appendix A10.

were, for example, to audit only entities that report an ITA score of zero, we would expect fewer entities to report zero scores, perhaps abstaining from reporting entirely. More theoretical development is necessary to better understand *equilibrium* data production—accounting for the strategic behavior of both central and decentralized governments—to better understand the properties of and optimal uses of administrative data.

7 Conclusion

Bureaucratic principals' access to information about the outputs of their subordinates is a central feature of most models of bureaucratic politics (Gailmard and Patty, 2012). Yet, we know little about the source of this information, particularly in the context of large bureaucratic organizations. We examine one such source widely used in central-decentralized government relations: self-reports from decentralized governments. When central governments rely on decentralized entities—whether regional governments or arms-length agencies—to implement policies, they require accurate, credible information about outputs. However, the central government has limited capacity to directly observe decentralized performance and must instead rely on bureaucratic self-reports. We develop a simple framework illustrating how this form of data production introduces

risks of strategic misreporting.

We study the production of ITA, a transparency index generated through self-reports from decentralized bureaucrats in Colombia. Using an original audit of true transparency practices across entities in Colombia, we document three key data pathologies: (1) positive selection into reporting, (2) modest overreporting of performance (transparency practices), and (3) general noise, which is highest at low levels of performance. Our experiment and interviews suggest that reporting is sensitive to central government oversight—here, by the PGN—consistent with strategic reporting. This finding aligns with Strathern's (1997: p. 308) conclusion that "when a measure becomes a target, it ceases to be a good measure," as well as Goodhart's Law (1983: p. 96), which states that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

Moving beyond these observations, we show through a simulation that bureaucratic reporting behaviors limit the central government's ability to effectively use the data it collects on arm's-length bureaucratic organizations. Together, our quantitative and qualitative evidence suggests that inter-institutional legibility is a persistent governance challenge. Crucially, these trade-offs are inherent to self-reports, a data collection method prevalent across diverse settings, rather than an idiosyncratic feature of weakly institutionalized contexts.[16]

Our work posits two important avenues for future study. First, we argue that the study of state legibility should be broadened to incorporate data 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

[16]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.

between these overlapping agency problems. Second, our study highlights that information itself is an important bureaucratic output, distinct from traditional measures of service provision or policy implementation. More research is needed to incorporate information as an output into theoretical and empirical studies of bureaucratic politics.

References

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

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`.

Brambor, Thomas, Agustín Goenaga, Johannes Lindvall, and Jan Teorell. 2020*a*. "The Lay of the Land: Information Capacity and the Modern State." *Comparative Political Studies* 53 (2): 175–213.

Brambor, Thomas, Agustín Goenaga, Johannes Lindvall, and Jan Teorell. 2020*b*. "The lay of the land: Information capacity and the modern state." *Comparative Political Studies* 53 (2): 175–213.

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* .

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.

Gailmard, Sean, and John W Patty. 2012. "Formal models of bureaucracy." *Annual Review of Political Science* 15: 353–377.

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.

Goodhart, C.A.E. 1983. *Monetary Theory and Practice: The U.K. Experience*. London: MacMillan Press.

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

Grossman, Guy, and Tara Slough. 2022. "Government Responsiveness in Developing Countries." *Annual Review of Political Science* 25: 131–153.

Guriev, Sergei, and Daniel Treisman. 2019. "Informational autocrats." *Journal of Economic Perspectives* 33 (4): 100–127.

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

Kuentae, Kim. 2007. "Distinctive Characteristics of the Joseon Dynasty's Fiscal Policy in the Nineteenth Century." *Korea Journal* 47 (2): 99–135.

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. 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.

Mikkelsen, Lene, David E Phillips, Carla AbouZahr, Philip W Setel, Don De Savigny, Rafael Lozano, and Alan D Lopez. 2015. "A global assessment of civil registration and vital statistics systems: monitoring data quality and progress." *The Lancet* 386 (10001): 1395–1406.

O'Hare, William P. 2019. "The importance of census accuracy: Uses of census data." *Differential Undercounts in the US Census: Who is Missed?* pp. 13–24.

Pepinsky, Thomas B, Jan H Pierskalla, and Audrey Sacks. 2017. "Bureaucracy and service delivery." *Annual Review of Political Science* 20: 249–268.

Prato, Carlo, and Ian R Turner. 2022. "The institutional foundations of the power to persuade." *American Journal of Political Science* .

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. "Bureaucratic Incentives and Administrative Data Production." Working paper, New York University.

Stathern, Marilyn. 1997. "'Improving ratings': audit in the British University system." *European Review* 5 (3): 305–321.

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

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.

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.