The Limits of Decentralized Administrative Data Collection: Evidence from Colombia *

Natalia Garbiras-Díaz† Tara Slough‡

August 25, 2023

Abstract

States collect vast amounts of data for use in policymaking and public administration. To do so, central governments frequently solicit data from decentralized bureaucrats. Because central governments use these data in policymaking, decentralized bureaucrats may face incentives to selectively report or misreport, limiting the quality of administrative data. We study the production of a transparency index by measuring the reporting behavior of bureaucrats in the near-universe of Colombian public sector entities. Using an original audit, we show that failure to report and misreporting vary systematically in actual transparency practices, revealing limits to the use of these data. Further, in partnership with a Colombian government watchdog agency, we implement a large-scale experiment that varies the salience of central government oversight. Increased salience changed the data bureaucrats reported to the central government. Similar dynamics across policy areas and regime types are apt to limit the quality of state information in multiple contexts.

*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, 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

The word "statistics" famously derives from the word "state." Indeed, for nearly six millennia, states have collected data through censuses, surveys, and cadasters to gather information in order to govern their populations (Grajalez et al., 2013; Scott, 1998). Modern states collect vast amounts of data for use in policymaking and public administration. Recent directives from international organizations and donors to expand data-driven governance advocate further entrenchment on both fronts, urging states to collect more data and rely more heavily on these data in policymaking to improve efficiency (van Ooijen, Ubaldi, and Welby, 2019; Bracken, Greenway, and Kenny, 2019). We argue that the link between administrative data inputs and state policy outputs renders state data collection as a political process, and thus state administrative data as a political outcome.

In addition to the population-level data characteristic of the earliest state data collection efforts, modern states rely heavily on multiple means of active and passive data collection. One important form of active data collection is the solicitation of data from bureaucrats across multiple entities, or distinct bureaucratic organizations. We conceive of this form of data production as an interaction between bureaucrats in different entities across different levels of government. Most commonly, central governments rely on data produced by decentralized entities to allocate resources (e.g., transfers or subsidies) or enforcement to decentralized entities.

While the bureaucratic politics literature emphasizes principal-agent problems between bureaucrats and politicians within a single entity or agency (e.g., Gailmard and Patty, 2012), less is known about how bureaucracies interact across levels of administration. We study one such interaction: the process of data transmission from (largely) decentralized entities to the central government, where, crucially, the latter relies on this output to perform some of its core functions. We contend that policies based upon these data—transfers or enforcement, depending on the domain—impact decentralized agents' incentives to truthfully report to the central government. These dynamics measurably worsen the quality of state administrative data that central governments use to make decisions, and turns state administrative data a political outcome.

Our theoretical framework links the behavior of bureaucrats in data-sending entities to the ac-

curacy of data that they report. We conceptualize the quality of data within a classic exposition of statistical measurement error, considering three pathologies: missingness, systematic measurement error, and non-systematic measurement error. We link each of these data issues to the behavior of bureaucrats. Non-submission of data generates missingness; intentional distortions of true (latent) measures constitute systematic measurement error; and lack of effort generates noise (non-systematic measurement error). When the agents of decentralized entities perceive that their responses may draw oversight attention from the central government with the possibility of enforcement, they may change their reporting behavior in an effort to deter this unwanted attention and potential punishment (Cook and Fortunato, 2022). Optimal reporting behavior from the perspective of decentralized entities therefore can introduce measurement error, ultimately limiting the quality of the data.

We test the idea that decentralized entities' reporting decisions are sensitive to their perceptions of the oversight process in the context of Colombia's annual National Transparency Index (known by the Spanish acronym ITA). 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 from the universe of 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 to 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 compliance with ITA. We use this manipulation to measure how data submission and reported scores change as a function of increased salience of the possibility of enforcement.

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 the bureaucrats' reports

and true levels of transparency practices. Collectively, this design allows us to learn how (decentralized) bureaucrats' anticipation of oversight conditions the data that they submit and to describe how these reports relate to true levels of transparency practices.

We present several findings indicative of political distortions to data quality. In the experiment, we show that by making the potential for oversight from the PGN more salient, more entities report, though this effect is small and statistically indistinguishable from zero. Instead, we show that entities report lower levels of compliance with transparency practices when the salience of oversight is increased. We show that this difference in reported scores can be decomposed into two effects: changes in the reporting behavior of entities that would always report regardless of their perceptions of PGN oversight, and changes in the composition of entities that choose to report because of the direct communication of oversight. First, we find that, on average, treated always-reporting entities report lower scores than their control counterparts (which we estimate using Lee bounds). Second, entities that select into reporting due to direct communication treatment report lower average scores than entities that would always report.

Our audit of public sector entities' transparency practices shows that true (latent) transparency practices correlate with three pathologies of measurement in the data. We document that high-performing entities disproportionately select into reporting, systematic distortions of data quality toward higher (more desirable) scores among low and middle-performing entities, and higher variance in reported scores among low-performing entities. The measure of variance proxies for non-systematic measurement error. We further consider how these pathologies vary in the administrative capacity of bureaucracies, which suggests that oversight and capacity jointly influence reporting behavior. Collectively, these results are consistent with the idea that central government reliance on data produced by decentralized government entities can undermine the accuracy and observability of that very data.

This paper makes four principal contributions to the literature. First, we argue that administrative data constitutes an important bureaucratic output. While recent studies have emphasized

bureaucracies' role in public goods provision (Pepinsky, Pierskalla, and Sacks, 2017; Grossman and Slough, 2022), we argue that bureaucrats' role in data production has been underemphasized. Time use surveys suggest bureucrats devote substantial effort to monitoring and evaluation, of which data collection and reporting constitutes a major task (Kalaj, Rogger, and Somani, 2020). In contrast to observations of the ubiquity of data collection as a quotidian task of bureaucrats, existing discussion of state data manipulation focuses on specific cases. For example, central government bureaucrats' career concerns in autocratic regimes are thought to encourage misreporting of economic data (Guriev and Treisman, 2019; Martínez, 2022; Trinh, 2021; Lorentzen, 2014; Wallace, 2016; Edmond, 2013). Recent work suggests that the delegation of data production to decentralized agencies can hinder data quality on politically sensitive topics in democratic regimes. For example, in the United States, police departments manipulate data in response to incentives related to measurement management (Eckhouse, 2022) or legislative oversight capacity (Cook and Fortunato, 2022). Our paper substantially expands these scope conditions with respect to strategic misreporting: we show that these dynamics need not be constrained to autocratic regimes or politically sensitive measures.

Second, we provide a framework linking the behavior of reporting bureaucrats to omission (failure to report) and data manipulation. Most studies of data production have focused on just one of these pathologies in isolation. For example, Hollyer, Rosendorff, and Vreeland (2011, 2014) study government decisions to report or not, abstracting from concerns about misreporting. In contrast, works including Martínez (2022), Gibilisco and Steinberg (2022), and Kofanov et al. (2022) emphasize misreporting or data manipulation in administrative data. Empirically, our data—like that of Cook and Fortunato (2022)—suggest that both omission and misreporting may well co-occur. When they do, it is crucial to study both phenomena to understand the quality—and limitations—of administrative data. To measure both phenomena, we propose new methodological innovations for decomposing failure to report from data manipulation.

Third, the application we study—Colombia's ITA—links our work to other literature docu-

menting central government efforts to monitor the transparency of local governments. The best-documented forms of state data collection on transparency practices (or lack thereof) are top-down audits, like those in Brazil (Avis, Ferraz, and Finan, 2018) and Mexico (Larreguy, Marshall, and Snyder Jr, 2018). In these audits, central governments deploy national public servants to validate correspondence between reported and actual expenses of decentralized government entities. However, these in-person audits are very expensive to conduct. Relying on self-reports in the data collection process we study is certainly cheaper and more scalable, but may yield far less accurate data. We demonstrate that these tradeoffs are central to governments' efforts to develop and utilize information effectively.

Finally, our work builds upon literature studying the states' collection and use of information about their citizens (Scott, 1998). Tractable data are essential to extract compliance (e.g., with taxes) or allocate resources across the state (Sánchez-Talanquer, 2020). Recent studies have stressed the connection between the legibility of citizens to the state and development and distributional outcomes (Slough, 2020; Bowles, 2020; Lee and Zhang, 2017).[1] We depart from this literature by considering a distinct, and arguably more frequent, source of state information: bureaucrats' reports. By considering the production of additional sources of (central) government information, we can better theorize the conditions under which states can benefit from using this data to make policies or improve policy implementation.

1 Theoretical Framework

1.1 Data as a state output

Prior to enumerating our account of bureaucratic data production, 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,

[1] Brambor et al. (2020) provide an overview of cross-national variation in data-collection and information-processing institutions.

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 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 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. 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.

Collectively, these three terms enter the bureaucrat's utility function in (2). In formulating the bureaucrat's utility in this way, we suppose 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 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 objective is to understand how bureaucrats' beliefs about specific policies influence the reporting behavior of bureaucrats. As such, we aim to study the data reporting to better characterize the incentives faced by decentralized bureaucrats. In Section 6, we return to a brief discussion of 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 to 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 specificying 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.

Ultimately, this analysis suggests that increasing oversight should impact reporting behavior by bureaucrats. While our simple 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 rela-

tionship 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. 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.[2]

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

Like many other national governments, the Colombian government relies heavily on self-reported data from territorial governments to inform policymaking 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."[3] Despite some efforts to consolidate these tasks, data collection, collation, and submission remains a central task of decentralized bureaucrats in Colombia. In an original survey of bureaucrats in Colombian municipal governments (*alcaldías*), for example, Slough (2023) 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 state 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).[4] By emphasizing data production, we complement existing treatments of bureaucrats as service providers by shedding light on an under-studied bureaucratic task consequential for governance and the distribution of resources.[5]

[3] All translations by authors.

[4] Bureaucrats in local governments self-report working an average of 52.3 hours per week in the time-use surveys by Slough (2023), in contrast to Colombia's workweek of 48 hours at the time of the survey. This suggests that the average bureaucrat may be overburdened.

[5] In an interview, a former National Planning Department official responsible for developing

2.1 Corruption and Transparency in Colombia

We study the collection of the 2020 Transparency and Access to Information Index (ITA), an annual measure of institutional compliance with transparency practices that was inaugurated in 2018. ITA was first mandated by Colombia's *Ley 1712 de 2014*.[6] The PGN is tasked with implementation of ITA. The PGN is the principal watchdog agency under the Public Ministry in Colombia. This central government entity investigates and sanctions any irregularity or misbehavior by publicly-elected officials, public servants, or any public sector agencies. The PGN is widely known by Colombian bureaucrats, even at the local level. Multiple survey respondents recall that all public servants are mandated to take training by the Administrative Department of Public Service (DAFP per its Spanish acronym) that explains the structure of the state and, importantly, what the PGN is and its oversight functions.

The PGN collects ITA as part of its preventative mandate to monitor public officials and entities. These efforts are intended to prevent corruption or other public misconduct. By collecting data, the PGN seeks to identify entities which may be more likely to engage in wrongdoing. Per our conversations with partners at the PGN, ITA data are used to direct preventative efforts by the PGN. Importantly, the PGN also initiates disciplinary proceedings against entities, which, upon investigation, fail to abide by laws regulating transparency practices and anti-corruption laws.

2.1.1 ITA: The Transparency Index

ITA mandates that more than 50,000 entities (organizations) report data on transparency practices annually. These subjects are classified into three categories. First, traditional subjects consist

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."

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

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 are private firms or individuals who contract with the state and political parties/social movements. We relegate the discussion of the latter two types of entities 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. These item responses are then weighted according to a formula to generate ITA. The final scores range from 0 to 100, where 100 means full compliance with the transparency practices specified on the questionnaire and 0 indicates no compliance with these regulated practices. These measures are published by the PGN in a document that consolidates ITA data.

Each year, the PGN has delegated the request for ITA data to a number of other national government agencies, that they refer to as "heads of sector." In practice, this means that entities receive the request to submit ITA data from a different entity. For example, almost all 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 are used to guide preventative anti-corruption efforts, low response rates and unknown accuracy render reliance on ITA potentially problematic. These pathologies in reporting mean that entities that honestly reveal imperfect transparency practices may be penalized while entities that do not report data or falsely report compliance with transparency best practices skirt oversight. We do not know precisely the use of the data by the PGN beyond these broad contours. Our interviews with bureaucrats who submitted ITA data revealed similar perceptions among the actors we study. Nevertheless, to the extent that ITA is used to guide

	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.

enforcement, it may yield perverse outcomes. As such, understanding how these data are produced and their accuracy is important.

3 Research Design

We conduct a field experiment in collaboration with the PGN to examine the data produced in the 2020 iteration of ITA. PGN sought to experiment to see if low-cost strategies could increase rates of complete data submission. In our effort to understand the behavior of the bureaucrats tasked with compiling ITA data, we emphasize the importance of descriptive quantities in addition to the causal estimands targeted with the experimental design. Much can be learned about the production of ITA from the relationship between actual transparency practices—as measured by an independent audit—and reported measures of compliance with these practices. These descriptive patterns are of consequence for how data can be interpreted and used. The causal effects show us the degree to which changes in the incentives of bureaucrats can change patterns of reporting to the PGN.

3.1 Sampling

Our unit of assignment is the entity, or organization. Our experimental sample consists of the near-universe (99%) of public-sector entities in Colombia. When sampling entities for the audit, we stratify on the national versus territorial (decentralized) designation of entities and oversample national entities because there are relatively few national entities. We describe the population of entities, our experimental sample, and the audited sample in Table 1.

3.2 Intervention and Assignment

We conduct an experiment to examine the effects of increased salience of oversight by the PGN on the data reporting behavior of officials within entities. We focus on two levels of treatment. Our primary manipulation emphasizes direct communication from the PGN to entities. Recall that the PGN delegates the request for data to other national entities known as "sector heads." In the status quo condition, indirect communication about ITA submission from these sector heads consisted of social media posts and other online communication. We increase the observability of the PGN's role in data collection by randomly assigning some entities to receive a direct email from PGN requesting the data. As such, the contrast we examine at the first level of treatment assignment consists of a comparison between the status quo—delegation to sector heads—versus the combination of delegated *and* direct communication from the PGN.

Direct communication from the PGN increases the perception that responses may be subject to scrutiny and, in the case of non-compliance, 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 those entities randomly assigned to communication from the PGN, we vary the content or frequency of the messages subtly using a $2 \times 2 \times 2 \times 2$ factorial design. We summarize variation in the content of these messages in Table 2 and report the full content of the messages in Table A2. We refer to the treatments in this second level of randomization 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). In contrast to the direct communication treatment which informs bureaucrats in each entity about the PGN's role and use of the data, thereby notably shifting incentives related to data submission, the nudges are simply additional sentences within these communications.[7] While our nudges are motivated by our theoretical framework, they are substantially weaker treatments than the top level randomization of direct messages.

The nudges serve practical purposes for the PGN and also offer analytic benefits. The use of direct messages is costly for the PGN because it requires staff time and expertise to tailor communications and respond to the additional volume of inquiries. In contrast, it is costless to change the content of these emails, conditional on sending them. Understanding how these communications can be optimized (at no additional cost) was important to our partners. We note that the specific nudges were informed by PGN officials' hypotheses about the sources of non-compliance, as described in Table 2. From the perspective of our design, one might worry that "direct communication" is too compound of a treatment. By varying the content, we can isolate (to some degree) the communication of oversight from some possible artifacts of the actual text that delivered that information

We use blocked random assignment to assign entities to each of the treatments. We first stratify entities on the basis of ITA completion in 2019. This creates two subgroups, thereby ensuring exact blocking on past completion of ITA. Within each subgroup, we formulated blocks of 18 entities that minimize Malahnobis 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 Malahnobis distance minimization ensures that local governments in the department of Antioquia are most

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

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.

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.

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 the items on ITA. The audit contains approximately 200 transparency practices, largely related to the online publication of information. These 200 binary items are reweighted and summed to form a 100-point scale. We audit 27.75 points of this scale, including some of the most promi-

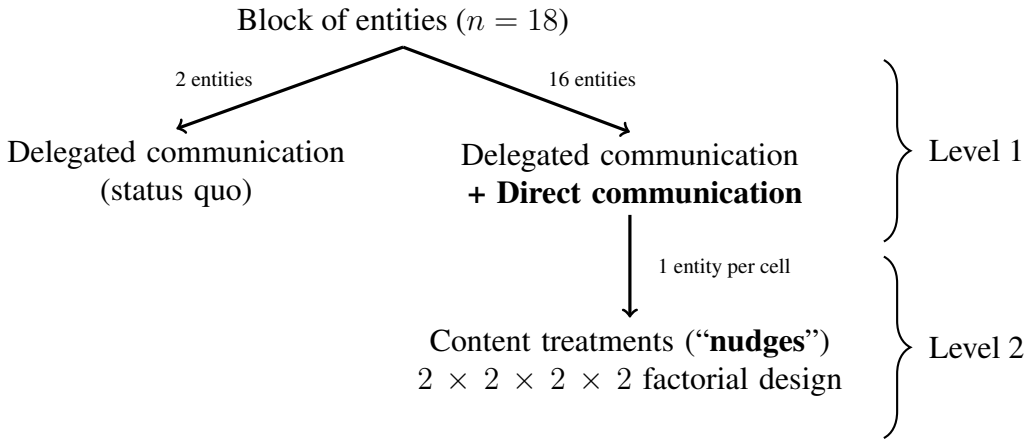

Figure 1: Treatment assignment scheme within each block of 18 entities.

nent transparency concerns. The audit was conducted by an independent firm hired by the researchers in June-July 2021.[8] Auditors were trained to search for select ITA index components through a standardized process. They recorded compliance with each item, in addition to more 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 for 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 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

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

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 true 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 outcomes. The first outcome is a binary indicator measuring completion or submission of data to the PGN. This indicator takes the value "1" whenever an entity submitted data to ITA. Our second outcome is the index score on ITA, which ranges from 0 to 100. Obviously, we only observe scores when data was 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 index like ITA for the audited items. This yields a score between 0 and 27.75. We construct indicators of whether the entity complies

with the item or not, based on the results of our audit as well as for the data reported by the entity, which we then reweight by the weights in the index. Finally, we contrast the outcome of those two calculations to measure divergence between reported and actual transparency practices. To facilitate this comparison, we construct the analogous index for audited items from the microdata, also ranging from 0 to 27.75. We discuss the quality of the microdata at more length in Appendix A4.

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 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 iden-

tical 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 changes reporting behavior

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 increased 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 this type of communication from an oversight body increased reporting, these effects are small in magnitude. There are several explanations for these small effects. First, rates of reporting are fairly high—

	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.

65%—absent direct communication (in control). It could be that it would be easier to induce entities to report when relatively few entities would otherwise report. We rely on the substantial autocorrelation of responses between 2019 and 2020 ($\rho = .42$) to test this hypothesis in our experimental sample. 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 A7). 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.

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 assumption 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 mono-

tonicity 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 A9, 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 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 lim-

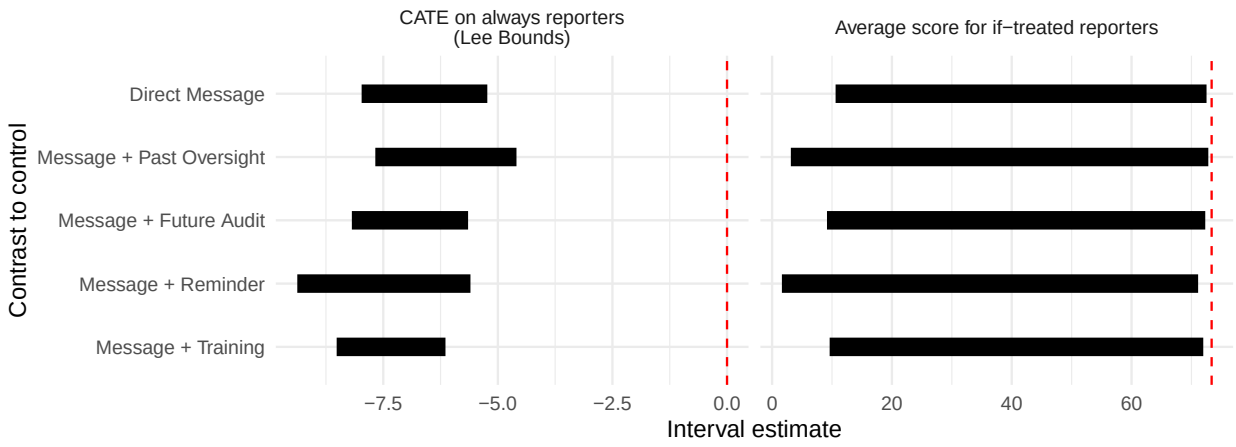

Figure 2: Decomposition of post-treatment estimands analogous to Column (3) of Table 3 (pooling over unstated message content) into a CATE on always reporters (left) and the average score among if-treated reporters (right). The CATEs are estimated using Lee trimming bounds and the interval estimate of 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.

itation 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 Entities positively select into reporting

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

We first examine 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 show that rates of reporting increase substantially in our measure of θ. Specifically, we estimate that, on average, an entity with a score of zero on the

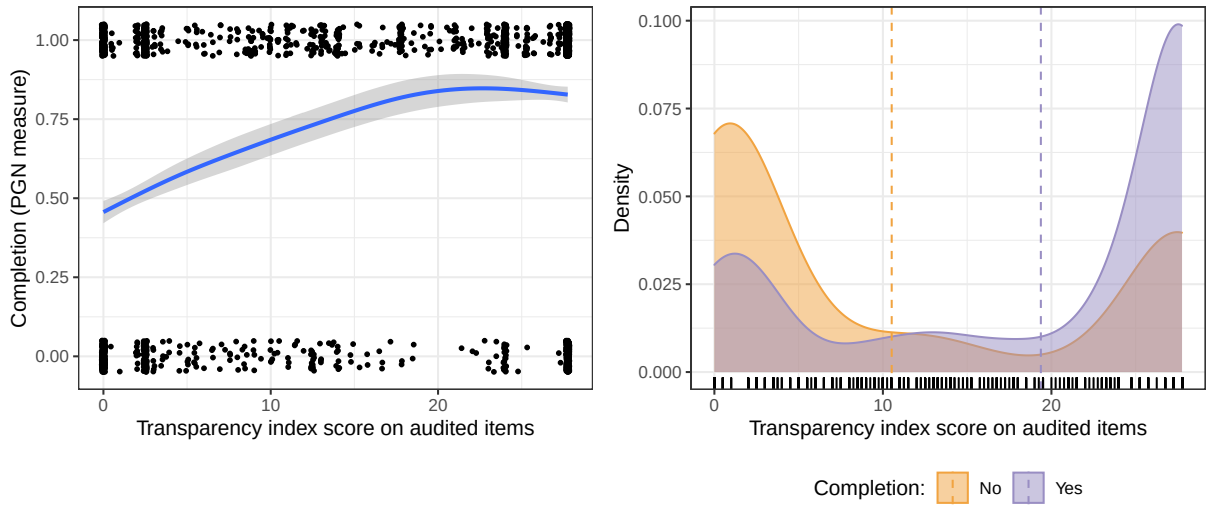

Figure 3: The association between the audit-measured transparency index and the probability of ITA data submission (left). The distribution of the audited-measured transparency index among entities that completed and failed to submit ITA data (right).

audit metric reports with probability 0.49 (95% CI: [0.45, 0.52]). An entity with a perfect score on the audit metric reports with probability 0.84 (95% CI: [0.82, 0.86]).

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.

4.3 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

	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.

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.

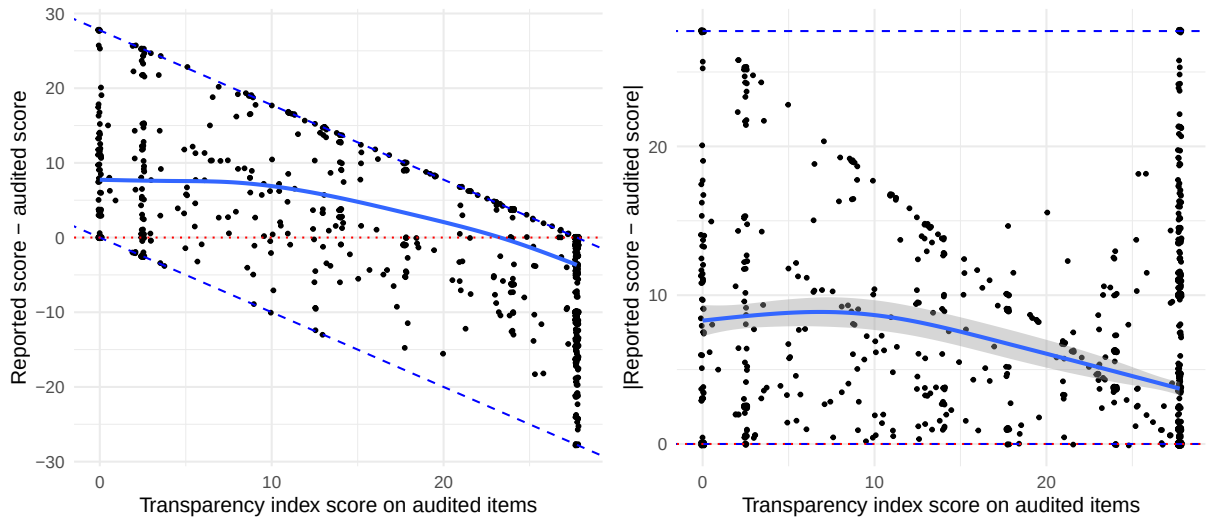

Figure 4: Discrepancies between the reported transparency practices and those detected in the audit.

We now consider the possibility of intentional misreporting (the parameter d in our model). Figure 4 examines 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 differences from reported 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 these curves are greater than—and statistically distinguishable from—zero. Because it is impossible to over-report a perfect score or under-report a score of zero, it is important to assess whether these deviations are simply mechanical. To that end, the right panel examines the association between audit-measured transparency practices (θ) and the magnitude of any distortion $|r - \theta|$. 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 practices.

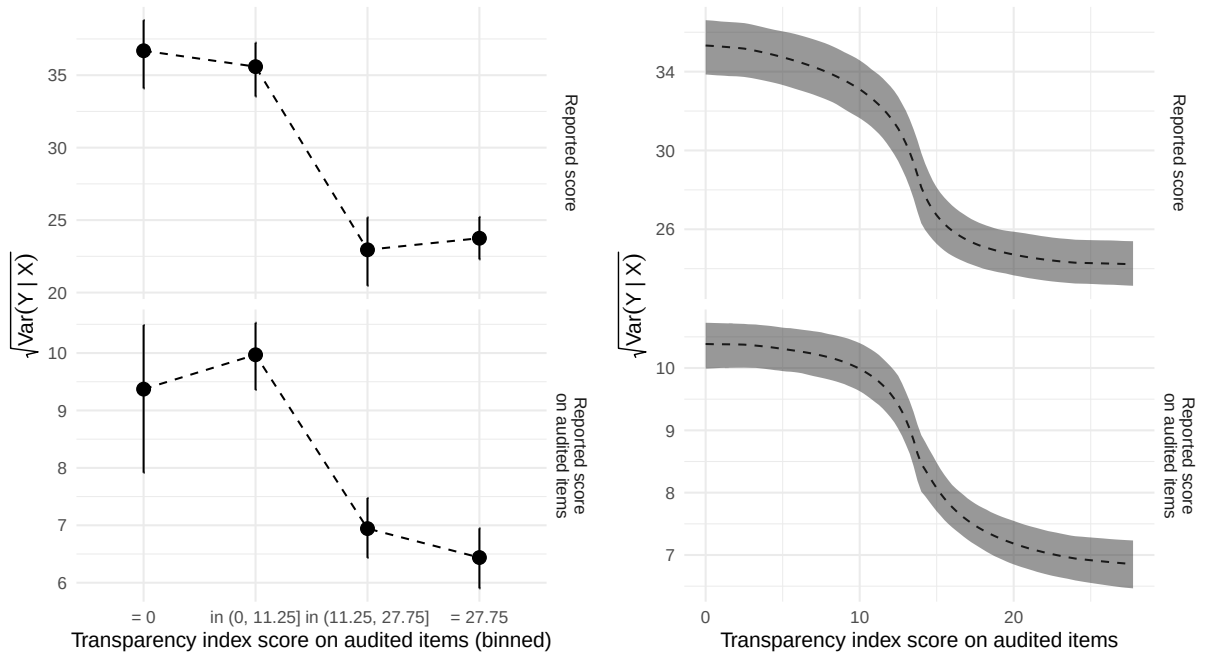

Figure 5: This plot shows how noise in reported ITA scores—measured by the standard deviation—relates to the audit-measured transparency index. The left panels bins entities by level of audit-measured transparency. The right panels employs a triangular kernel to estimate the conditional-standard deviation.

4.4 Noise in reporting

Our final 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 deviation across the support of the audit measure.

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. We now turn to a broader possible alternative explanation for our findings in Figure 5, with implications for the other findings from the audit: the administrative capacity of audited entities.

5 The Role of Administrative Capacity

Our account, to this point, focuses on entities' strategic response to oversight by the central government when they report data. Yet, per our theoretical framework, reporting to the national government should depend, to some degree, on an entity's administrative capacity. We interpret administrative capacity reduces the bureaucrat's cost of effort, $c(e)$. Higher effort reduces the probability that unintentional distortions in reports, ε, are large in magnitude. Changes in effort may, in turn, affect the choice to report and the choice to systematically distort scores.

There is substantial variation the level of administrative capacity within the public sector in Colombia. For example, the national government has greater administrative capacity than almost all territorial entities. Public sector organizations draw from differently-skilled pools of workers, as a function of geographic location, wages, and tasks. Further, some bureaucracies have greater access to technology than others. Finally, variation in the organization of these entities, or the leadership thereof, may generate further variation in administrative capacity. We now characterize the relationship between two measures of administrative capacity, transparency practices, and reporting behavior, to understand the degree to which administrative capacity may shape features of reporting behavior.

There exists no perfect measure of administrative capacity. Yet, the Colombian national government has made efforts to characterize both municipalities and public sector entities in these

terms. At the municipal level, we use the National Planning Department's (DNP's) index of municipal performance to measure geographical variation in administrative capacity.[9] This measure seeks to measure outputs of local governments, which represent a subset of the entities that we study. Second, at the institutional level, we use data from the Administrative Department of Public Administration's (DAFP's) index of institutional performance to measure administrative capacity at the entity (institution) level.

These measures have different strengths and weaknesses for the purposes of our analysis of the role of administrative capacity. Use of DNP's index of municipal performance for all entities in a municipality assumes that capacity positively covaries across institutions in the same municipality. If this covariance is weak, our estimates are likely to be attenuated by measurement error. However, we have this measure for 952 of 953 municipalities that are represented in the audit, minimizing the degree of missingness. In contrast, the institutional performance index measures institutional performance at the level of observation: the entity. Unfortunately, this index is constructed by self-reports (like the ITA matrix) and is not required for the full set of entities obligated to report ITA scores. This leads to higher levels of missingness: we have scores for only 1,210 of 2,400 audited entities. Further, pathologies of reporting from entities to the national government may be similar to those that we document in our audit of the ITA scores, which could induce correlation in the errors between the capacity measure and our reporting outcomes of interest. Given the strengths and limitations of each measure, our discussion is informed by the comparison of the signs of estimates from each measure.

Table 5 reports the associations between these measures of municipal and institutional capacity and our measures of reporting behavior from the audit. In Columns 1-2, we show that the probability of completion of the ITA matrix increases in both measures of administrative capacity. Comparing the two estimates, a one standard deviation increase in municipal capacity increases the probability of reporting by 1.8 percentage points whereas a one standard deviation increase in

[9]For more information on index construction, see Angulo et al. (2018).

	Completed ITA $\mathbb{I}(r \neq \emptyset)$		Audit score (θ)		Distortion $(r-\theta)$		| Distortion | $(\mid r-\theta \mid)$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Muncipal capacity (standardized)	0.018^{+}		0.851***		0.673***		-0.655***	
	(0.009)		(0.232)		(0.146)		(0.128)	
Institutional capacity (standardized)		0.116***		2.572***		0.568**		-1.259***
		(0.011)		(0.279)		(0.217)		(0.186)
Num. Obs.	2400	2400	2400	2400	1696	1696	1696	1696
Sample	All	All	All	All	Completed	Completed	Completed	Completed
Indicator for missing capacity measure	yes	yes	yes	yes	yes	yes	yes	yes
DV mean | Capacity measure present	0.707	0.807	16.773	21.137	-0.119	0.341	3.678	3.298
DV std. dev | Capacity measure present	0.455	0.395	11.929	9.683	7.141	6.301	6.122	5.379
DV range	$\{0,1\}$	$\{0,1\}$	$[0, 27.75]$	$[0, 27.75]$	$[-27.75, 27.75]$	$[-27.75, 27.75]$	$[0, 27.75]$	$[0, 27.75]$

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

Table 5: Associations between each of the capacity measures and measured parameters capturing reporting behavior. All regressions are estimated by OLS with heteroskedasticity-robust standard errors.

institutional capacity increases the probability of reporting by 11.6 percentage points. Note further that entities for which DAFP has constructed the institutional capacity measure are 10 percentage points more likely to report in the first place, providing suggestive evidence in favor of our intuitions that entity-level reporting behavior may be correlated across measures.

Columns 3-4 show that audit-measured transparency scores are increasing in municipal capacity. Again, a one standard deviation increase in institutional capacity is associated with a audit score that is 2.57 points higher (on the 27.75 point scale), whereas a one standard deviation increase in municipal capacity is associated with only a 0.85 point increase in audit scores. The differences in the magnitude of these estimates is consistent with the insight the municipal-level measure introduces measurement error that attenuates these associations toward zero. Here again, the entities for which we have DAFP capacity measures are positively selected, meaning that they have higher average transparency scores than the pool of all entities.

Columns 5-8 report distortions in reporting among the entities that reported ITA transparency scores. First, note that Columns 6-7 document an *increase* in the degree of over-reporting as a function of administrative capacity. Thus, even though these entities already have higher compliance with the audited transparency practices (as shown in Columns 3-4), they over-report by a greater amount than low-capacity institutions. Note that we observe this even though these entities

have less room to over-report, since they are closer to the ceiling of 27.75 points (the maximum possible score on the audit). One plausible interpretation of this finding is that administrative capacity facilitates efforts to strategically misrepresent performance to the purveyors of oversight in the national government (the PGN). Second, columns 7-8 show that despite this systematic over-reporting by higher-capacity entities, higher-capacity entities also report with *less noise* (greater accuracy) than low-capacity entities. These distortions include both intentional (d) and unintentional (ε) distortions. Combining the estimates in columns 5-6 and 7-8 suggests that lower capacity introduces large unintentional distortions that outweigh the score manipulations that we document in columns 5-6. This is consistent with our interpretation of administrative capacity as reducing costs of effort. Lower costs of effort increase effort, thereby attenuating the magnitude of unintentional distortions.[10]

These analyses report only associations between administrative capacity and reporting behavior. Myriad features that shape administrative capacity could may also shape transparency practices and/or reporting incentives. Nevertheless, these findings do provide two important takeaways. First, consistent with our theoretical framework, administrative capacity predicts reporting behavior. In the context of the Colombian ITA, higher capacity entities (or municipalities) report at higher rates, distort their scores toward desired outcomes, but ultimately report with less average noise. The latter finding—on noise—is directly consistent with our assumption about how effort shapes reported scores.[11] Second, the former outcomes are not obviously driven by variation in administrative capacity in isolation. Our finding that intentional distortions correlate positively with

[10]Table A9 reports specifications that interact audit-measured transparency scores (θ) with our capacity proxies.

[11]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 5, complementarities between quality and the cost of effort can affect the probability of completion.

administrative capacity suggests that administrative 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 a role for both administrative capacity and oversight in the production of administrative data. The importance of administrative capacity within individual entities suggests limits on what the national government could achieve through optimal design of oversight instruments in isolation.

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 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 6, 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 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.

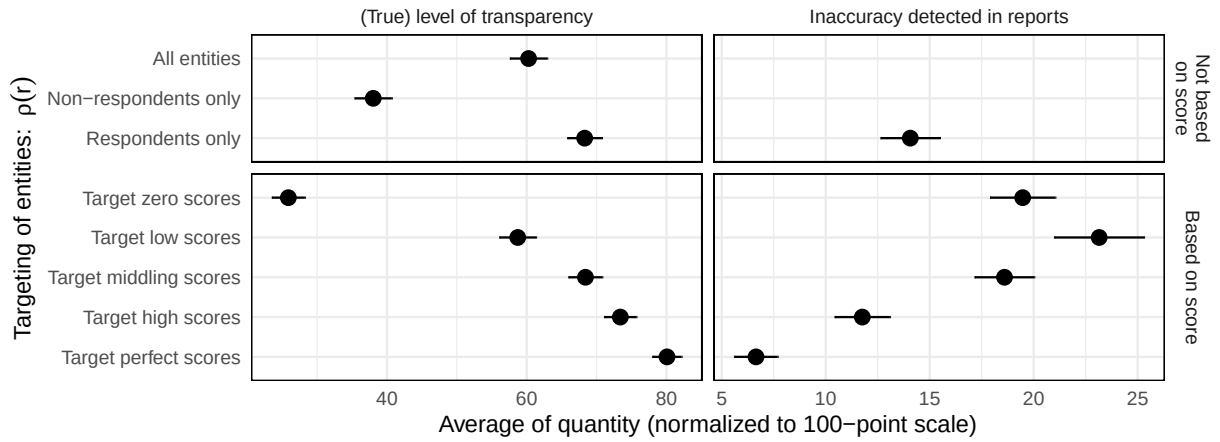

Figure 6: Average levels of θ (left) and $|r - \theta|$ right under various simulated oversight strategies, $\rho(r)$. The 95% confidence intervals correspond to a sample of 1,000 entities. Functional forms used for $\rho(r)$ appear in Appendix A10.

7 Conclusion

Proponents of data-driven governance seek to expand government use of data in policymaking. However, these data are routinely produced in the course of interactions between bureaucracies. When bureaucrats or their entities stand to win or lose from the use of the data they supply, they may have incentives to misreport. We document distortions in the quality of the data that the Colombian PGN uses to detect instances of corruption or malfeasance. These findings echo Strathern's (1997: p. 308) conclusion that "when a measure becomes a target, it ceases to be a good measure" and Goodhart's (1983: p. 96) law which states that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

Our results extend these maxims in three ways. First, we concur with Cook and Fortunato (2022) that state data should be considered a *political* output. Given the large number of data production tasks within decentralized bureaucrats' portfolios, data constitutes an important bureaucratic output. Second, we show how relationships between central and decentralized bureaucracies shape the incentives of data-reporting bureaucrats. These incentives are typically far more subtle than the targets described by Stathern (1997) and Goodhart (1983), but can produce important

distortions in data quality. Understanding these incentives is critical to learning about the accuracy/quality of state data. Finally, we argue that using data to inform government decision-making is a harder problem than acknowledged by many advocates of data-driven policymaking. Further work is needed theoretically and empirically to design mechanisms through which governments can produce and utilize data effectively.

Existing scholarship on administrative data distortions has focused primarily on autocratic regimes and generally assumes that democracies have built-in checks and balances that counteract the incentives of governments to distort data. We challenge this conventional wisdom by providing a new framework for understanding distortions in administrative data that are rooted in interactions between bureaucracies. These efforts generalize and extend insights from work on police data in the US (Eckhouse, 2022; Cook and Fortunato, 2022). Our logic shows how distortions in administrative data can present across policy domains in democracies and autocracies alike. Our framework further bridges two concerns about administrative data—missingness and manipulation—that are often treated as distinct by existing work. We show how the bureaucratic behaviors underpinning these phenomena are related.

Our study posits a number of questions for further research. First, we focus on an understudied agency problem between the national government and decentralized governments and show how national government oversight shapes administrative data outputs. Yet, there presumably exist agency problems *within* both the national and local governments that may also shape how data collectors and reporters internalize the incentives we describe. Theoretical advances will facilitate understanding of the relationship between these overlapping agency problems. Second, we provide suggestive evidence on how administrative capacity and oversight jointly shape the quality of administrative data. This interaction between capacity and incentives echoes recent work by Raffler (2022) in the domain of service provision. Understanding how the national government optimally devises data collection processes in light of variation in local capacity represents an important topic for future research in Colombia and beyond.

Studying the accuracy and quality of state administrative administrative has important implications for empirical social science. While measurement error is widely discussed in the case of survey data (i.e., Bound, Brown, and Mathiowetz, 2001) and for expert-coded data (i.e., Rozenas, 2013), existing discussions of measurement error in administrative data are more limited in scope, emphasizing certain types of data and certain political contexts. Our framework suggests that these problems are likely far more systematic. Broader acknowledgement of these limitations of administrative data produced by states—even in those known for relatively high-quality data—are important for understanding the limitations of our data and therefore our inferences.

References

Angulo, Roberto, Dallma Aniza, Alfredo Bateman, Natalie Gómez, Jorge Iván González, Javier Pérez, Juan Mauricio Ramírez, Fernando Rojas, Juliana Ruíz, Fabio Sánchez, and Carlos Sepúlveda. 2018. "Medición del Desempeño Municipal: Hacia una gestión orientada a resultados." Documentos CEDE 1148, available at https://repositorio.uniandes.edu.co/handle/1992/41037.

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

Avis, Eric, Claudio Ferraz, and Frederico Finan. 2018. "Do government audits reduce corruption? Estimating the impacts of exposing corrupt politicians." *Journal of Political Economy* 126 (5): 1912–1964.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of econometrics*. Vol. 5 Elsevier pp. 3705–3843.

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.

Bracken, Mike, Andrew Greenway, and Angeles Kenny. 2019. From Information to Actionable Intelligence: Adapting Governments to Data Analytics. Technical report Inter-American Development Bank.

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

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.

Gibilisco, Michael, and Jessica Steinberg. 2022. "Strategic Reporting: A Formal Model of Biases in Conflict Data." *American Political Science Review* forthcoming.

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* Forthcoming.

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

Hollyer, James R, B Peter Rosendorff, and James Raymond Vreeland. 2011. "Democracy and transparency." *The Journal of Politics* 73 (4): 1191–1205.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2014. "Measuring Transparency." *Political Analysis* 22: 413–434.

Kalaj, Jozefina, Daniel Rogger, and Ravi Somani. 2020. "Bureaucrat Time-Use and Productivity: Evidence from a Survey Experiment." *Working paper* .

Kofanov, Dmitrii, Vladimir Kozlov, Alexander Libman, and Nikita Zakharov. 2022. "Encouraged to Cheat? Federal Incentives and Career Concerns at the Sub-national Level as Determinants of Under-Reporting of COVID-19 Mortality in Russia." *British Journal of Political Science* pp. 1–26.

Larreguy, Horacio A, John Marshall, and James M Snyder Jr. 2018. "Leveling the playing field: How campaign advertising can help non-dominant parties." *Journal of the European Economic Association* 16 (6): 1812–1849.

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.

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

Raffler, Pia J. 2022. "Does political oversight of the bureaucracy increase accountability? Field experimental evidence from a dominant party regime." *American Political Science Review* 116 (4): 1443–1459.

Rozenas, Arturas. 2013. "Inferring Ideological Ambiguity from Survey Data." In *Advances in Political Economy, Institutions, Modeling and Empirical Analysis*, ed. Norman Schoefield, Gonzalo Caballero, and Daniel Kselman. Springer pp. 369–383.

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. 2020. "Oversight, Inequality, and Capacity." Working paper, available at `https://taraslough.com/assets/pdf/oci.pdf`.

Slough, Tara. 2023. "The Incentives of Data Producers." 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* .

van Ooijen, Charlotte, Barbara Ubaldi, and Benjamin Welby. 2019. A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance. OECD Working Papers on Public Governance 33 OECD.

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.