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

Natalia Garbiras-Díaz† Tara Slough‡

November 3, 2021

Abstract

States collect vast amounts of data for use in policymaking and public administration. Central governments frequently solicit data from the bureaucrats of decentralized government entities. When central governments use these data in policymaking, bureaucrats in decentralized entities may face incentives to report inaccurately, limiting the quality of state data. We study these dynamics in the production of Colombia's 2020 National Transparency Index in collaboration with the Colombian Attorney Inspector General's office. A field experiment varied the salience of direct oversight of the reported data by this watchdog agency in direct communication to individual bureaucratic entities. Bureaucrats respond to more salient oversight by changing reporting behavior. Through an independent audit of subjects' transparency practices, we describe how bureaucrats' reporting behaviors vary in the actual transparency practices of their organization. We argue that the strategic dynamics inherent to decentralized data production by bureaucrats render state data an important but understudied political outcome.

*Thanks to Carolina Torreblanca for excellent research assistance. We thank Ian Turner and APSA panel participants for very helpful comments. We are very grateful to Innovations for Poverty Action for their incredible work managing this project.

†Post-Doctoral Fellow, European University Institute, natalia.garbirasdiaz@eui.eu

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

1 Introduction

The word "statistics" famously derives from the word "state." Indeed, over nearly six millenia, states have collected data through censuses, surveys, and cadasters to gather information in order to govern their populations (Gómez Grajalez, 2013; Scott, 1998). Modern states collect vast amounts of data for use in policymaking and public administration. Recent calls by international organizations and donors to expand data-driven governance advocate further entrenchment on both fronts, calling upon states to collect more data and rely more heavily on data analyses in policymaking (van Ooijen, Ubaldi, and Welby, 2019; Bracken, Greenway, and Kenny, 2019). We argue that this link between state data inputs and state policy outputs necessitates the study of data collection as a political process, and therefore renders state administrative data as an important political output.

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. Most commonly, central governments rely on data produced by decentralized entities to allocate resources (transfers) or enforcement to decentralized entities. As such, in addition to the principal-agent relationships between principals and agents within individual entities that animate much of the bureaucratic politics literature, we are interested in data transmission from (largely) decentralized entities to the central government that are denoted with the blue arrows in Figure 1. We contend that policies based upon these data – transfers or enforcement, depending on the domain – impacts decentralized agents' incentives to truthfully report the solicited information. These dynamics can measurably worsen the quality of state administrative data that central governments use to make decisions.

Data production represents a common, albeit understudied, task of bureaucrats globally. The

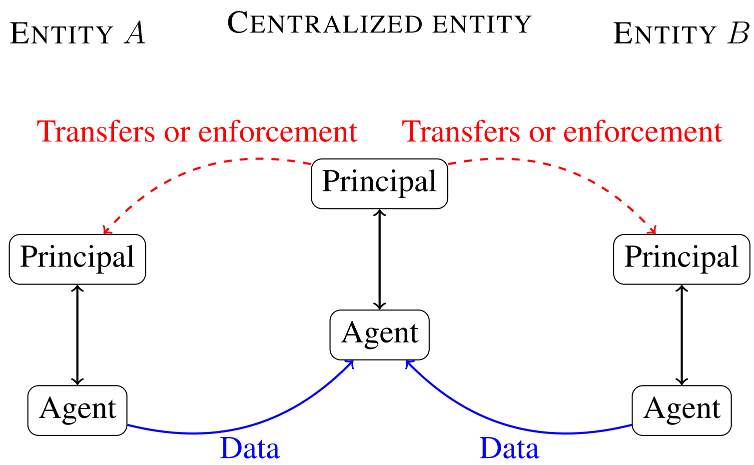

Figure 1: Schematic of data production interactions between central government agency and decentralized agencies A and B. The black arrows denote principal agent relationships within each organization.

frequency of requests for data features prominently in bureaucrats' complaints. Further, the time demands of these activities are occasionally cited as a drag on bureaucratic processes. Recent survey evidence from Ethiopia suggests that bureaucrats spend more time reporting data to monitoring and evaluation systems than delivering services (Kalaj, Rogger, and Somani, 2020). Moreover, during the COVID-19 pandemic, the UK Department of Health and Social Care (2020) decried "duplicative data requests" as a form of "excess bureaucracy" and an impediment to providing the necessary medical care during the pandemic.[1] Given the frequency of data production as a bureaucratic task and the widespread use of the resultant administrative data – both by states and scholars – understanding how bureaucratic behavior generates variation in the accuracy of state data is important.

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

[1]Often these duplicative requests are non-redundant per the conceptualizations of Landau (1969) or Ting (2003). For example, a duplicate request is non-redundant if two different central agencies query the same data for use in distinct processes.

error, and non-systematic measurement error. We link each of these data issues to the behavior of bureaucrats. Non-completion of data requests 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. Optimal reporting behavior from the perspective of decentralized entities therefore can introduce measurement error, ultimately limiting the quality of the data.

We test this 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 entity that collects and compiles the ITA annually. The PGN is a watchdog agency tasked with monitoring and pursuing public corruption charges. The ITA is an entity-level index covering approximately 200 distinct transparency practices, which are collapsed an additive index. Each year, over 50,000 public entities and contractors are asked to report their compliance (or lack thereof) with these 200 practices, generating an index score. To understand entities' sensitivity to oversight, the PGN randomly varied whether these entities received direct communication from the PGN 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 exposure to 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 the reported data 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 characterize descriptively how these reports relate to true levels of transparency practices.

We present several findings indicative of political distortions to data quality. First, 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 for public sector entities. Instead, we show that public sector entities report lower levels of compliance with transparency practices when assigned to oversight. We show that this finding can be decomposed into two effects: (i) on average, "always reporter" entities report lower scores, and (ii) entities that select into reporting because of the direct communication report lower scores than "always reporter" entities. Second, 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 positive selection 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. Collectively, these results are consistent with the idea that central government reliance on data produced by decentralized government entities can undermine the accuracy of that very data.

This paper makes four principal contributions to the literature. First, we contend 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) and especially the "last mile " delivery of services (Muralidharan et al., 2021; Callen et al., 2020), we argue that bureaucrats' role in data production has been underemphasized and understudied. Time use surveys suggest that a focus on the delivery of public goods and services captures a small fraction of what bureaucrats spend their time doing, while monitoring and evaluation demand substantial bureaucratic time and effort (Kalaj, Rogger, and Somani, 2020). Existing discussions of state data manipulation emphasize strategic distortions of data by autocratic regimes Guriev and Treisman (2019); Martínez (2021); Trinh (2021); Lorentzen (2014); Wallace

(2016); Edmond (2013). We show instead that small changes in bureaucrats' incentives for reporting yield non-trivial shifts in data quality outside of autocratic regimes. As such, studying the bureaucratic production of data is important because state reliance on data of variable quality may lead to substantial distortions in policymaking or citizen welfare.

Second, to the bureaucratic politics literature, our work advances a call for further study of strategic interactions *between* different bureaucratic organizations within the state. Much existing literature has emphasized interactions *within* bureaucratic organizations, often invoking principal-agent frameworks (see Gailmard and Patty 2012 for a review). We show that interactions between bureaucratic organizations introduce incentives to strategically report data, likely to the detriment of states as they try to use the resultant data. Indeed, we suggest that these interactions in the domain of data production and reporting adhere to observations by Goodhart (1983) and Stathern (1997: p.4) that "when a measure becomes a target, it ceases to be a good measure."

Third, the application we study – Colombia's ITA – links our work to other literature documenting central government efforts to measure transparency and limit corruption at the local level. The best-documented forms of state data collection on transparency practices (or lack thereof) are top-down audits. In these audits, central governments deploy national public servants to validate correspondence between reported and actual expenses of decentralized government entities. These audits are well documented in studies of corruption and accountability in Brazil (Avis, Ferraz, and Finan, 2018) and Mexico (Larreguy, Marshall, and Snyder Jr, 2018). However, these in-person audits are very expensive to conduct (Seabra, 2019). Relying on self-reports in the data collection process we study is certainly cheaper and more scalable, but may yield far less accurate data.[2] 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 informa-

[2]ICT systems may provide an additional low-cost alternative for states to consider to improve or validate measurement (Callen et al., 2020; Singh, 2020).

tion 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 the state" and development and distributional outcomes (Slough, 2020; Bowles, 2020), whereas others have used the former to operationalize state capacity (Lee and Zhang, 2017).[3] We depart from this literature by considering a distinct, and arguably more frequent, source of state information: bureaucratic data collection and reporting. By considering the production of additional sources of (central) state information, we can better theorize the conditions under which states can benefit from using this data to make policies or improve policy implementation.

2 Theoretical Framework

2.1 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 data are produced by state entities. Decentralized entities are tasked with reporting some measure of the quality of their performance – whether public service outputs, budgetary management, or transparency practices – to the central government. A bureaucrat (or office) within the decentralized entity determines whether to comply with the request for information, making a report or failing 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 performance quality reporting performance of $\theta + d$, where $d \in \mathbb{R}$. There may also be unintentional errors in reporting. These errors could be misunderstanding of questions,

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

typos, or failure to correctly follow directions. We represent these errors as $\epsilon \sim f(\cdot)$, where $f(\cdot)$ is a mean-zero density.

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

Equation 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. From the analyst's perspective, non-reports (denoted $r = \emptyset$) manifest as missing data.

It is worthwhile to note that some states aggregate reports when publicizing data. When this occurs, instead of observing r, an analyst or curious citizen would observe some summary statistic aggregating over multiple entities, often territorial or administrative divisions. While such aggregation may average out the effects of unintentional misreporting (given that $f(\cdot)$ is mean zero by assumption), the relationship between the true summary statistic $g(\boldsymbol{\theta})$ and the reported $g(\mathbf{r})$ quality is ambiguous in the presence of non-reporting and/or intentional misreporting. By focusing our analysis at the most disaggregate level, the entities that produce data, we provide implications for what analysts may observe from aggregate data.

2.2 Data production

We focus on the decision of decentralized government entities to report data to the central government. The actors that we focus are therefore officials within the government entities tasked with data reporting. These officials are generally bureaucrats, though in our context, it is possible that politicians (principals) usurp this responsibility in some settings.[4] Our decision-theoretic

[4]In entities administered directly by elected politicians, it is possible that the actors that report data are politicians. We discuss this possibility in the empirical context that we study in Sections 3 and 5.

framework is premised on several assumptions about these bureaucrats' incentives to report. We maintain the notation used in Equation 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 $\rho(r)$. Indeed, if $\rho(r)$ were constant, 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 $P(\theta, r) > 0$, potentially varies in true quality (θ), the reported data (r), and/or the distance between these measures. While we do not specify the precise functional form of P, it is conceivable that the penalty is set to punish poor performance (i.e., low θ) or distortions in the reported data (i.e., some function of the distance between r and θ).

Finally, we assume that collecting, collating, entering, and reporting data demands that bureaucrats exert costly effort. Readers familiar with completing paperwork to apply for state services, file income taxes (where relevant), or seeking reimbursement may be familiar with these costs. 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 $\frac{\partial Var(\epsilon|e)}{\partial e} < 0$. This implies that r is a function of e in the case when the bureaucrat chooses to report.

Collectively, these three terms enter the bureaucrat's utility function in Equation (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 work upon bureaucrats. Lastly, it could be the case that data is used principally to target resources to an institution. While this is not relevant in the empirical case we describe, one could add a benefit term to the utility function in Equation (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 given current policies influence data quality. As such, we aim to study the data reporting behaviors of entities within current policies to better understand the incentives of bureaucrats within these entities. In Section 6, we return to a brief discussion of equilibrium considerations after discussing the empirical findings.

2.3 Testable Implications: Measurement Error as a Political Outcome

Within this simple framework, we discuss several testable implications relevant to the research design that we present. The goal of the empirical exercise that we present is to learn about the properties of oversight from the central government in the production of data. First, we consider the implications of an exogenous shock to oversight. We refer to the product $\rho(r)P(\theta, r)$ as anticipated oversight. By making the bureaucrats tasked with reporting data more sensitive to the possibility of oversight, bureaucrats may respond by changing reporting behavior. Within the framework, anticipation of oversight from the central government drives the decision to devote any effort to reporting data.

Remark 1. *Testable implication with respect to a shock to perceived oversight:*

i) Increased anticipation of oversight yields higher rates of reporting among audited entities.

ii) Increased oversight yields changes in the aggregate distribution of reports, r*, conditional on* $r \neq \emptyset$ *by: (i) inducing additional entities to report; (ii) changing incentives for misreporting,*

d; or (iii) reducing the variance of reports by increasing bureaucratic effort.

In sum, Remark 1 suggests that if bureaucrats are responsive to a shock to the degree of perceived oversight, we may expect changes in the composition of entities who choose to report $r \neq \emptyset$. While our general framework does not provide concrete predictions on how the distribution of scores should change in response to oversight, it admits three mechanisms through which to understand changes in the distribution of reports.

Our framework offers further allows us to consider how bureaucratic incentives underlie three pathologies of measurement that are familiar in statistical analysis: non-response, distortions in reporting, and non-systematic measurement error. The framework allows us to identify sufficient conditions for the emergence of these pathologies in terms of bureaucrats' incentives. In turn, these conditions serve as testable implications in the data.

To understand selection into reporting, we first note that given our assumption that submitting reports requires costly effort, for or any bureaucrat to submit a report, $x \neq \emptyset$, it must be the case that $\rho(x)P(\theta, x) < \rho(\emptyset)P(\theta, \emptyset)$. Observing reports implies that some entities anticipate the possibility of oversight on the basis of the content of the data that they provide. Second, the presence of missing reports could be indicative of differences in anticipated oversight and/or the cost of effort necessary to collate and submit the necessary data.

Remark 2. *Testable implication with respect to selection into reporting:*

i) Non-zero rates of reporting imply that there exists some entities for which, given a report of x, $\rho(x)P(\theta, x) < \rho(\emptyset)P(\theta, \emptyset)$.

ii) Suppose that costs of effort are independent of quality. Systematic variation of response rates in quality, $\frac{\partial Pr(r \neq \emptyset)}{\partial \theta}$, indicate that anticipated oversight varies in quality.

Next, consider intentional distortion in reported data, the d parameter in Equation (1). Does intentional distortion of data stem from the use of data in targeting audits or perceived penalization of distortions in the data (conditional on detection)? Within our framework, for bureaucrats to choose to systematically distort the data that they send, $d \neq 0$, they must believe that the likelihood of audit depends on reported data. If auditing were independent of the data, then bureaucrats would have no incentive to systematically distort quality in their reports. In this case, if the penalty depended only on actual quality, the bureaucrat would be indifferent between truthfully reporting and systematically distorting the data. In the more likely case that penalties depend, in part, on disparities between true and reported quality, the possible punishment should attenuate intentional misreporting to zero when the possibility of oversight does not depend on the assembled data. It is important to note that for individual entities, we cannot identify d because $r - \theta = d + \epsilon$, and ϵ is the unobserved realization of a random variable. As such, our tests for intentional distortion – and thus the presence of perceptions that the central government uses data to target oversight – relies on distortions in the aggregate, i.e. $E[r - \theta]$.

The relationship between punishment and the occurrence of intentional misreporting is somewhat more subtle. Suppose that the central government used scores to direct oversight but conditioned punishment only on true quality, not reported quality. In this case, we would expect that all entities that report submit the score that minimizes the probability of oversight. In this case, uniform scores across entities are independent of, and thus uninformative of, true quality. This scenario can easily be tested. Specifically, if (1) we observe misreporting in the aggregate $E[r-\theta] \neq 0$ and reported scores covary with true quality, there is evidence that, on average, entities anticipate that misrepresentation of quality can help to evade oversight.

Remark 3. *Testable implications with respect to intentional misreporting:*

i) Observation of misreporting in the aggregate $E[r-\theta] > 0$ (resp. < 0) is a sufficient condition for $\frac{\partial \rho(r)}{\partial r} < 0$, (resp. $\frac{\partial \rho(r)}{\partial r} > 0$).

ii) A sufficient condition for a belief that penalties depend on deviations from the true data is $\frac{\theta E[r]}{\partial \theta} \neq 0$.

Finally, we consider the implications for unintentional misreporting, or non-systematic error. Recall that bureaucrats cannot control the magnitude or direction of unintentional errors only the distribution from which they are drawn. The degree to which bureaucrats work harder to minimize these errors will depend on bureaucrats' attitudes toward risk, which we have not imposed in the present framework. As such, we evaluate the relationship between quality, θ, and the conditional variance $Var(r|\theta)$ inductively.

3 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 decentralized (territorial) government entities. Following decentralization in the 1980s and particularly the 1991 Constitution, departmental and municipal government entities assumed greater responsibilities and developed more public administration capabilities (Fizbein, 1997). The deepening of Colombia's fiscal, political, and administrative decentralization 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 and to better target policies following a results-based policymaking approach (World Bank, 2011).

Like many other national governments, the Colombian government relies heavily on self-reported data from territorial governments to inform policymaking and target monitoring. To collect this data, national government entities request fine-grained data from local officials, typically with a deadline. The reliance on this type of self-reported data from local governments has imposed such a heavy administrative burden on municipalities that a series of reforms have focused on the consolidation of duplicative requests from different national governrment entites.[5] From these data, the national government generally collates or indexes the reported data for use in policymaking. To better understand the dynamics inherent to this type of data collection, 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.

[5] For example, the adoption of the Unique Terrritorial Form (*Formulario Único Territorial*) aims to centralize these requests for data for monitoring budget execution and expenses.

3.1 Corruption and Transparency in Colombia

The ITA was first mandated by the Colombia's 2014 Transparency and the Right to Public Information Law (*Ley 1712 de 2014*), henceforth the "Transparency Law." The transparency practices advanced by this legislation ostensibly respond to perceptions of rampant corruption in Colombia. Indeed, corruption has long been cited as a problem in Colombia. Per Transparency International's 2020 annual Corruption Perceptions Index, Colombia ranks near the median in the world (51st percentile), in Latin America and the Caribbean (48th percentile), and among other upper-middle income economies (50th percentile). Corruption remains a durable and salient concern on public opinion surveys. Across the 2014, 2016, and 2018 LAPOP surveys, 8-18 percent of respondents have named corruption as the most important issue in Colombia in an open-ended response. Moreover, a majority of respondents assess corruption to be "very common" among public officials (2014 and 2018 surveys) and the modal respondent surmised that "more than half" of politicians are corrupt (2018 survey). While there are many policy instruments that may reduce corruption, promotion of greater transparency is widely promoted as an antidote to corruption (Vrushi and Hodess, 2017).

3.1.1 The Office of the Attorney-Inspector General (PGN)

The Tranparency Law tasks the Office of the Attorney-Inspector General (which we refer to by PGN, its Spanish acronym) with its implementation. The PGN is the principal watchdog agency under the Public Ministry in Colombia. This national-level entity investigates and sanctions any irregularity or misbehavior by publicly-elected officials, public servants, or any public sector agency. In this role, the PGN represents civil society vis-à-vis the Colombian state.[6] The PGN has three core functions. First, it monitors public officials and entities in an effort to prevent misconduct. Second, it intervenes in some judicial proceedings on behalf of the citizens when rights are vio-

[6]The PGN coexists with two other watchdog agencies with closely-related mandates: the Attorney General's Office (*Fiscalía General de la Nación*) and the Ombudsman's Office (also part of the Public Ministry).

lated. Finally, the PGN prosecutes alleged misconduct by public servants or other subjects executing public functions (i.e., contractors).

In overseeing the collection of the ITA, the PGN works within its preventative mandate. By collecting data, the PGN seeks to detect risks in public administration. In the case of ITA, poor transparency practices may be indicative of larger issues of corruption or mismanagement. Importantly, the PGN also initiates disciplinary proceedings against the entities, that upon investigation, fail to abide by laws regulating transparency practices and anti-corruption laws.

3.1.2 ITA: The Transparency Index

Since its launch in 2018, ITA has mandated that over 50,000 entities (organizations) report data on transparency practices annually in an effort to comply with the Transparency Law. These subjects are classified into three categories. First, traditional subjects consist of public entities, oversight bodies, and public companies that belong to the State. While these public sector entities include both central and territorial (decentralized) institutions, over 90% of these institutions are territorial entities, largely departmental and municipal government institutions. Second "non-traditional" subjects comprise persons, legal entities, or legal dependents who provide any public services or manage any public funds. These subjects are mostly individuals and firms with state contracts. Finally, the remaining subjects are large political parties or social movements. Our primary focus is on the first category – public sector organizations – so we pool the "non-traditional" subjects and political parties/social movements into a residual category within our research design.

The ITA questionnaire (or matrix) 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.[7] The survey consists of approximately 200 binary responses (yes/no). These item responses are then weighted according to a formula to generate the ITA. The final scores range from 0 to 100, where 100 means full compliance with the

[7]The specific practices and response options vary slighty according to subject type (i.e., public entities, non-traditional subjects, or political parties).

transparency practices specified on the ITA questionnaire and 0 indicates no compliance with these regulated practices. These measures are published by the PGN in a document that consolidates and ranks all subjects that completed ITA in the preceding year.

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 receives the request from the Administrative Department of Public Administration. The source of communication to non-traditional entities is more varied. In sum, there are 76 national agencies tasked with procuring this data from obligated entities.

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 and 3.4% of other entities completed the 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 the matrix potentially problematic. Low response rates and a lack of understanding of the accuracy of the data 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 are not. We do not know preciesly the use of the data by the PGN beyond these broad contours. Nevertheless, to the extent that the ITA is used to guide enforcement, it may yield perverse outcomes. As such, understanding how these data are produced and their accuracy is important.

4 Research Design

We conduct a field experiment in collaboration with the PGN to examine data produced in the 2020 iteration of the ITA. PGN's goal in the experiment was to develop low-cost strategies to increase rates of complete data submission. Minimizing the cost of the intervention to the PGN, both in terms of staff time, expertise, and implementation costs was paramount in the design of the interventions.

In our effort to understand the behavior of the bureaucrats tasked with compiling the ITA, we emphasize the importance of descriptive quantities in addition to the causal estimands identified by the experimental design. Much can be learned about the production of Colombia's Transparency Index 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 experimental estimands show us the degree to which small changes in the incentives of bureaucrats can change patterns of reporting to the PGN.

4.1 Sampling

Our unit of assignment was the entity, or organization. As we discuss above, the PGN classifies entities as public sector ("traditional"), private sector ("non-traditional"), and political parties/social movements. Our public sector experimental sample consists of 6,556 of the 6,618 public sector entities in Colombia according to the PGN's classification. The remaining 62 entities were randomly sampled as a pilot group to test the feasibility of the intervention. Given constraints on staff time, PGN sought to engage in direct communication with a subset of the 36,647 private sector entities. Our sample of non-traditional entities consists of a convenience sample of 5,329 of the 5,367 prioritized firms – the remaining 38 were part of the earlier pilot intervention. All 168 political parties and social movements were included in the experimental sample.

The private audits of data quality were conducted on a subset of public sector entities. The list of entities for the audits was constructed by a stratified random sample of national and territorial (decentralized) public sector entities. We describe the population of entities, our experimental sample, and the audited sample in Table 1.

4.2 Intervention and Assignment

We conduct a factorial experiment to examine the effects of direct communication with the PGN on the data reporting behavior of officials within entities. We focus on two levels treatment. The

	All Obligated Entities*	Experimental Entities	Audited Entities
Category	Count (n)	Count (n)	Count (n)
PUBLIC SECTOR	6,556	6,556	2,400
National	237	237	200
Territorial	5,928	5,928	2,200
Undesignated	391	391	0
PRIVATE SECTOR	41,938	5,329	0
PGN Priority	5,329	5,329	0
PGN Non-priority	36,609	0	0
PARTIES/MOVEMENTS	168	168	0
Total	48,662	12,053	2,400

Table 1: Sampling of entities in experiment and audit outcome measurement. *This total omits 62 public sector and 38 private sector entities that were randomly sampled and used in a piloting pre-test of intervention implementation.

first emphasizes direct communication from the PGN to obligated entities. Recall that the PGN delegates the request for data to other national entities known as "sector heads." We increase the observability of the PGN's role in data collection by randomly assigning some entities to receive a direct email requesting data from these entities. 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. Within our theoretical framework, we argue that direct communication from the PGN increases the perception that responses may be subject to scrutiny and, in the case of non-compliance, punitive action.[8] This interpretation of the "direct communication" treatment is consistent with the *ex-post* assessment of our partners working within the PGN.[9]

[8]Note that it is possible that direct communication could have very different effects in subsequent iterations of the index. If entities observe the oversight activities of the PGN (or lack thereof), and update their perceptions of the likelihood or severity of oversight accordingly, the treatment effect of direct communication may vary over time. We view this treatment as a way to measure some aspects of data production, we do not advocate for or against this communication strategy as a policy intervention.

[9]In discussions with the PGN staff working on the ITA, we learned that the main national-level offices in charge of disseminating the requirement to fill out the ITA among the obligated entities (the Secretary of Transparency of the Presidency and the Adminstrative Department of

Within those entities randomly assigned to communication from the PGN, we vary the content or frequency of the messages quite subtly using a $2 \times 2 \times 2 \times 2$ factorial design, as summarized in Table 2. Two of the treatments aim to subtly change subjects' beliefs about the probability of oversight. Specifically, the "past oversight" treatment acknowledges whether or not subjects submitted data for the 2019 Transparency Index. The "future audit" treatment conveys that the PGN may conduct an audit of submissions, aiming to increase assessments of the probability of audit within the model. Both of these oversight-oriented treatments aim to raise subjects' assessments of $\rho(r)$ within the model. Importantly, PGN opted not to make any statement of the consequences for non-compliance. A third message provides links to training videos. These videos seek to reduce questions about how to fill out the Transparency Index, corresponding to a subtle reduction in the cost of effort, $c(e)$, by raising agents' familiarity with procedures. Finally, a reminder treatment consists of a repeated version of the email closer to the submission deadline for the ITA matrix. As in the direct communication treatment more broadly, this increases subjects' perceptions of oversight by the PGN in general. We include the text of all treatments in Appendix A3 (see Table A2).

We use blocked random assignment to assign entities to each of the treatments. We stratify the sample between traditional (public sector) versus other (private-sector, parties, etc.) entities and on the basis of ITA completion in 2019. This creates four subgroups within the sample, and ensures exact blocking on these two covariates. Within each subgroup, we formulated blocks of 18 entities that minimize Malahnobis distance between covariates using PGN's classification of organizational or entity type. This means that within each block of 18, all entities are identical in 2019 ITA completion behavior and in their traditional/non-traditional classification. The Malahnobis blocking ensures that local governments are most likely to be in the same block as other local governments, etc. Within the blocks, we randomly assign two entities to a pure control condition

Public Administration) publicized this requirement in their social media and other official online channels and by organizing an online event to provide further details about the ITA.

Arm	Levels		Motivation
Past (retrospective) oversight	0 =	No mention of past compliance with collection of ITA data.	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).
	1 =	Acknowledgement of compliance/non-compliance with 2019 ITA Matrix data collection.	
Future (prospective) oversight	0 =	No mention of possible audits to 2020 ITA submissions	Increase perceptions of the likelihood of sanction or enforcement for non-completion of the ITA matrix.
	1 =	Mention of possible audits of 2020 ITA submissions.	
Training	0 =	No information on training resources for filling out the ITA matrix.	Increase the capabilities of agents with respect to the ITA matrix task.
	1 =	Link to PGN resources (including videos) on how to fill out the ITA matrix.	
Reminder	0 =	Single direct communication from PGN to entity.	Reinforce perception of PGN oversight over matrix completion.
	1 =	Direct communication + a reminder from PGN to the entity.	

Table 2: Experimental manipulations within the direct contact communications between the PGN and entities.

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.

4.3 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 the ITA matrix. 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 prominent transparency concerns. The audit was conducted by an independent firm hired by the researchers in June-July 2021.[10] Auditors were trained to search for select ITA matrix 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 A5. Our primary measures come from the indicators for compliance with each ITA matrix item.

Crucially, we conduct the audit in parallel for entities that reported and entities that failed to report in the 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. We conducted a stratified random sample of 200 national and 2,200 decentralized entities. We opted to analyze traditional entities because of our focus on data production as an interaction between different bureaucracies. Moreover, PGN did not provide the names of the non-traditional entities that did not complete the matrix. They are identified in our data only through national ID or tax ID numbers. As such, it would not have been possible to accurately search for non-reporting non-traditional entities. Because our sampling into the audit oversamples national entities, we include indicators for national vs. decentralized entities throughout the analysis. In Tables A3-A4, we show that, conditional on the order of an entity (national vs. decentralized), assignment to the independent audit is balanced across past (2018 and 2019) ITA participation and scores as well as our treatments.

One potential concern is that, given complications identifying, contracting, and training the firm for this non-standard audit, too much time elapsed between the submission of ITA data and the audit six to seven months later. Improvements or reduction in transparency practices over this time are a source of measurement error in our measure of quality, θ. They should not bias our results, however, unless (1) entities became more transparent because of the treatment only after they submitted their data to the PGN; or (2) changes in transparency practices between the data submission and audit vary with the true level of transparency practices. In Figure A4, we show that experimental treatments do not have an effect on the underlying quality measure, allaying the

[10] This audit is completely independent of the PGN.

	Not in public microdata	In public microdata
Did not complete ITA (PGN measure)	615 (25.6%)	89 (3.7%)
Completed ITA (PGN measure)	478 (19.9%)	1,218 (50.8%)

Table 3: Confusion matrix for PGN data versus public microdata.

first concern.

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 obligated entities. We further assume that idiosyncratic errors by two different individuals – bureaucrats and independent auditors – at different times are uncorrelated.

In comparing item-level responses to their audit results to measure discrepancies between reported and true scores, we rely on ITA responses present in the public microdata that records responses to each item. When we compare entities for which the PGN has recorded a score to those in the public microdata, we observe some discrepancies. Specifically, there are fewer entities in the public microdata than entities that completed the ITA matrix according to the PGN. This is evident in the lower left cell of Table 3, where nearly 20% of audited entities completed the matrix but are not present in the microdata. A further 3.7% of the sample did not complete the ITA matrix per PGN's match but is in the public microdata. Because bureaucrats self-reported entity names – which often do not match the administrative records – PGN and the research team conducted separate hand matches between the data inputs and the scores. These 3.7% of entities is suggestive of the lack of overlap in these matches. Ultimately, this suggests that measurement error due to misattribution of scores to entities is quite limited. Our primary concern, which we discuss at greater length when interpreting results, is the absence of some entities from the public microdata.

4.4 Measures

Our primary measures seek to capture the theoretical parameters r, entities' reports of transparency practices, and θ, the true level of transparency practices. Our primary measures of r come 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 the ITA matrix. Our second outcome is the index score on the 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 the ITA index 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 the audit as well as for the data reported by the entity, which we then reweight by the weight in the true index.[11]. 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.

4.5 Estimation and Identification

In our analysis of the effects of the experimental treatments on submission of ITA matrix data, the primary estimand is the Average Marginal Component Effect (AMCE) of each of the five manipulations: the use of direct communication between the PGN and obligated entities and the four factorial manipulations of the content of those requests. We employ OLS to estimate AMCEs using Equation 3. The estimators of the AMCEs are β_1 through β_5 in:

[11] The weights used in construction of the ITA score are publicly available on the PGN's website.

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

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 completition of the matrix given our exact blocking. The sample includes all experimental subjects.

We also estimate a non-standard "post treatment" estimand in which we regress reported scores on the ITA matrix on the experimental treatments using an estimator identical to Equation 3. Because the sample is conditioned on a post-treatment outcome, the β's are no longer estimators of well-defined causal effects. However, as we show in Appendix A7, the post-treatment estimand is closely related to the sum of the conditional average treatment effects (ATEs) on entities that would report regardless which treatment they were assigned to and a term capturing differential selection into reporting as a function of the treatments. The latter term is not a causal effect. Because we are looking at differences in averages (as opposed to variance), two of the mechanisms in Remark 1 are worth measuring: selection into reporting and changes in intentional misreporting. To decompose these two effects, we use Lee (2009) trimming bounds to bound the conditional ATEs effects among always reporters. Importantly, when CATE estimates are bounded away from zero, we can show that at least some entities report different scores under different treatments. This provides experimental evidence of distortion in reported scores.

The additional testable implications in Remarks 2-3 examine relationships between "true" latent levels of transparency practices and reported transparency practices. In our non-experimental analysis, we examine the relationship between our audit measure of θ, dentoted 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 make descriptive inferences about 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 association between the these data. We also reweight some specifications by the inverse of sample inclusion probabilities to account for the fact that national entities are overrepresented among the audited sample.

Finally, we rely on one form of institutional variation in public sector entities to leverage observational variation in sensitivity to oversight by the PGN. Public sector entities vary in whether they are directed by an elected politician or a bureaucrat. As such, we examine how these patterns vary in the identity of the entity principal. Our expectation, based on discussion with the PGN and bureaucrats, is that entities led by elected by elected principals are more sensitive to oversight. To evaluate these predictions we estimate models of the form:

$$Y_i = \gamma_0 + \gamma_1 \text{Audit}_i + \gamma_2 \text{Elected principal}_i + \gamma_3 \text{Audit}_i \times \text{Elected principal}_i + \boldsymbol{\kappa}\mathbf{X}_i + \epsilon_i \tag{5}$$

5 Results

5.1 Direct Communication from the PGN changes reporting behavior

How does increasing the salience of oversight change reporting behavior? In Table 4 Panel A, we report estimates of the AMCEs on the probability of completing the transparency index data submission. We find that direct communication increases the probability of completion by 5 percentage points ($p < 0.005$). Repeated direct communication in the form of a reminder increased the probability of completion by an additional 3.4 percentage points ($p < 0.005$). Comparisons of columns 2-3 and 5-6 indicate that these effects are substantially larger among non-public sector entities (parties, social movements, and private-sector entities) than among public-sector entities. We cannot reject the null hypothesis that AMCE $= 0$ for public-sector entities for any treatment component.

In Figure A6, we estimate differences between the CAMCEs of public sector and non-public sector entities. We find that these differences are significant at the $\alpha = 0.05$ level only for the reminder messsage and at the $\alpha = .1$ level for direct communication. There are multiple possible reasons for these differences in responsiveness to the treatments. We discuss two. First, the delegated communication to public and non-public sector entities came through different agencies. Delegated communication to public entities came through the Administrative Department for Public Administration. Those to other entities came through a patchwork of different entities. Different agencies may have exerted more or less effort or maintained stronger or weaker lines of communication with the entities that they preside over.

Alternatively, given the large disparities in rates of completion in the control condition, it could be that it is 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 for public sector entities ($\rho = .42$) to test this hypothesis *within* the sample public sector entities. While public sector 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 CAMCEs are all near-zero and indistinguishable from zero (Figure A7). As such, we do not find evidence consistent with the idea that variation in "baseline" rates of completion drive variation in treatment effects. This analysis further suggests that, at least for public sector entities, 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 Panel B of Table 4, we regress scores, conditional on reporting, on the experimental treatments. This analysis conditions on reporting and is thus "post-treatment." The table suggests that

	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: COMPLETION OF TRANSPARENCY INDEX DATA SUBMISSION						
Direct communication	0.050***	0.029	0.075***	0.050***	0.030	0.074***
	(0.017)	(0.022)	(0.017)	(0.013)	(0.020)	(0.016)
Oversight of past completion	0.009	0.000	0.020*	0.008	-0.001	0.019*
	(0.010)	(0.012)	(0.012)	(0.008)	(0.011)	(0.011)
Possible future audit	-0.003	-0.005	-0.001	-0.004	-0.005	-0.002
	(0.010)	(0.012)	(0.012)	(0.008)	(0.011)	(0.011)
Direct reminder	0.034***	0.013	0.057***	0.033***	0.012	0.058***
	(0.010)	(0.012)	(0.012)	(0.008)	(0.011)	(0.011)
Training	-0.001	-0.008	0.008	-0.001	-0.007	0.006
	(0.010)	(0.012)	(0.012)	(0.008)	(0.011)	(0.011)
Num.Obs.	12053	6556	5497	12053	6556	5497
Sample	All	Public Sector	Other	All	Public Sector	Other
Block FE				yes	yes	yes
Control mean (s.d.)	0.46 (0.5)	0.68 (0.47)	0.19 (0.39)	0.46 (0.5)	0.68 (0.47)	0.19 (0.39)
DV range	{0,1}	{0,1}	{0,1}	{0,1}	{0,1}	{0,1}
PANEL B: TRANSPARENCY INDEX SCORES, CONDITIONAL ON COMPLETION						
Direct communication	-7.762***	-6.066***	-8.210*	-6.091***	-5.886***	-6.271
	(1.442)	(1.477)	(4.507)	(1.379)	(1.473)	(4.365)
Oversight of past completion	0.131	0.587	-0.026	0.584	0.540	0.672
	(0.880)	(0.950)	(1.847)	(0.826)	(0.936)	(1.737)
Possible future audit	-0.841	-0.453	-2.307	-0.878	-0.666	-1.828
	(0.880)	(0.950)	(1.844)	(0.827)	(0.936)	(1.744)
Direct reminder	-3.834***	-2.836***	-3.371*	-2.661***	-2.763***	-2.337
	(0.879)	(0.949)	(1.868)	(0.825)	(0.932)	(1.760)
Training	-0.512	-1.094	2.709	-0.455	-1.191	2.579
	(0.880)	(0.950)	(1.844)	(0.830)	(0.939)	(1.753)
Num.Obs.	5496	4446	1050	5496	4446	1050
Sample	All	Public Sector	Other	All	Public Sector	Other
Block FE				yes	yes	yes
Control mean (s.d.)	69.13 (30.63)	73.37 (29.4)	51.21 (29.23)	69.13 (30.63)	73.37 (29.4)	51.21 (29.23)
DV range	{0,...,100}	{0,...,100}	{0,...,100}	{0,...,100}	{0,...,100}	{0,...,100}

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

Table 4: AMCE estimates on completion of request for data comprising the transparency index (Panel A) and association between randomly-assigned treatments and transparency index scores (Panel B), conditional on completion. Robust clustered standard errors in parentheses. p-values correspond to two-tailed tests.

direct communication is associated with reductions in reported scores. Moreover, reminders are associated with an additional (additive) reduction in scores. We observe these associations for both public sector and other entities. While the results are weaker for "other" entities when we include block fixed effects, note that with reporting rates of 19% among other entities, there are many blocks with very limited reporting among this subgroup. As such, there is limited variation in some treatment indicators within some blocks once we condition on reporting. In sum, our results suggest that closer exposure to the PGN's role in collecting the ITA decreases the scores that entities report, conditional on reporting.

While these coefficient estimates should not be interpreted as causal effects given our sample conditioning on reporting, we show in Appendix A7 that a closely-related post-treatment estimand can be decomposed into a weighted sum of the conditional ATE (CATE) among "always reporters" and a term capturing differential selection into reporting. With respect to direct communication (for example), a non-zero conditional ATE implies that there exist entities that report different scores if they were contacted directly by the PGN than they would have if not contacted. The selection term consists of the expected potential outcome among entities that report when contacted by the PGN but not report when they are not contacted directly (or vice versa). To disentangle these two components, we use the trimming bounds by Lee (2009) to bound the effects on always-reporters. These bounds impose an assumption of monotonicity, which simplifies our decomposition. Once we have bounded the CATE, we then back out bounds on average scores among if-treated reporters algebraically following our derivation in Appendix A7.

In Figure 2, we report interval estimates of the CATE among "always reporters" and the expectation of scores among "if-treated reporters." We plot both sets of estimates for public sector and other entities in Figure 2. For public sector entities, where differences in rates of reporting are small (Table 4), we can show clearly that the CATE estimates are negative. This suggests that, on average, entities who would always report *lower* scores when exposed to oversight. 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 public sector 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 A6.

In the case of other entities, recall from Table 4 that we observed increased reports of reporting among those entities assigned to receive direct communication from the PGN. This larger treatment effect manifests in much wider interval estimates in Figure 2. Given the width of these intervals, we cannot fully determine whether the negative association between direct communication and scores is driven by treatment effects among always reporters or selection of lower-performing entities into reporting.

Collectively, our from Table 4 and Figure 2 provide evidence that reporting behavior is sensitive to oversight by the PGN. Even in the case of public sector entities in which we cannot detect non-zero average effects on data submission, we show that when exposed to oversight, some entities report lower scores than they would otherwise report. Yet, this sensitivity to oversight does not provide any evidence about the *accuracy* of reported scores as we do not have any measure of true quality, θ. To this end, we now proceed to our analysis of the audit data.

5.2 Public sector entities positively select into reporting

While the Transparency Index by the PGN allows us to assess which entities completed the data submission for the Transparency Index, it does not provide information on the transparency practices of entities that did not report in the absence of assumptions about the behaviors that generate missingness. Our audit of a subset of public-sector 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 whether entities choose to report.

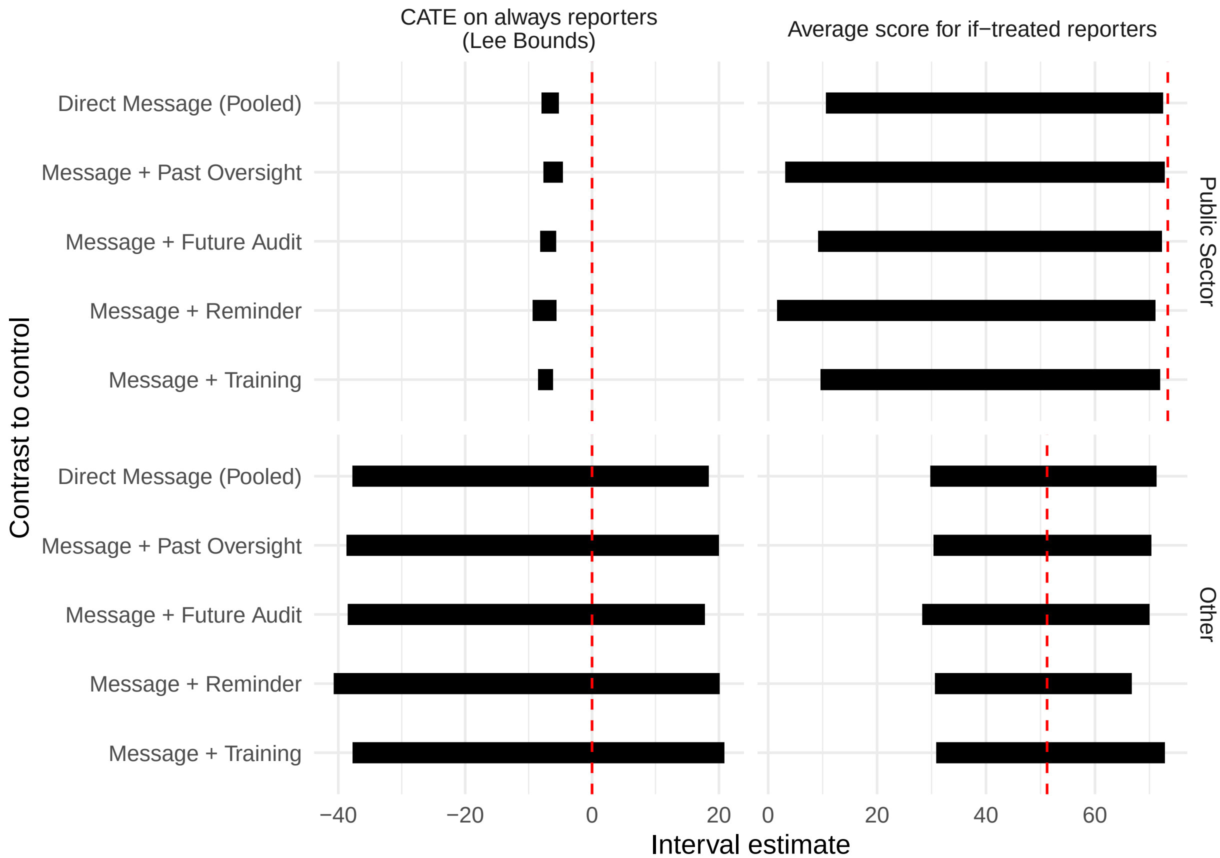

Figure 2: Decomposition of post treatment-estimands analogous to Panel B of Table 4 (pooling over unstated message content) into a conditional ATE on always reporters (left) and the expectation of the score among if-treated reporters. The CATEs are estimated using Lee trimming bounds and the expectation of the score among if-treated reporters is calculated algebraically from those bounds, using the expressions in Section A7. In the CATE plots, the vertical red line indicates an CATE of 0. In the plots depicting the expectation of the score among if-treated reporters, the red line indicates the expectation of the score among all reporters.

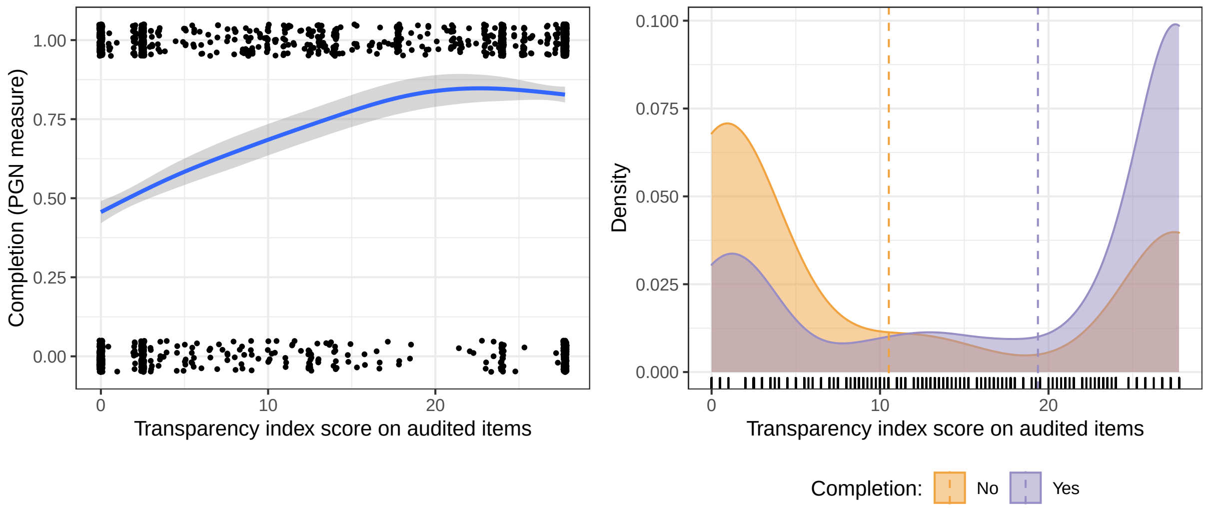

Figure 3: The association between the audit-measured Transparency Index score and the probability of submitting the Transparency Index data (left). The distribution of audited-measured Transparency Index scores among entities that completed and failed to complete the Transparency Index.

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 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 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]).

Combining the experimental results in Table 4 and the association between the audit transparency measure in Figure 3, it is reasonable to ask which entities – as characterized by their level of compliance with the audited transparency practices – are induced to report at higher levels by the direct communication treatment. Note, however, that the audit sample consisted of only public sector (traditional) entities. Recall that for public entities we cannot reject the null hypothesis of no (average) effect on completion of the transparency index. Similarly, Figure A8 compares the treatment effects of the communication treatments on completion of ITA in the full public sector sample to those in the audited sample. It shows similar AMCE estimates across both both samples, all of which are statistically indistinguishable from zero.

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 Transparency Index scores will overstate the level of compliance with constituent compliance practices among public sector institutions in Colombia.

Recall that we argue that bureaucrats in entities headed by elected politicians may perceive higher levels of oversight by the PGN than bureaucrats. In Figure A10, we disaggregate the left panel of Figure 3 between entities headed by elected and non-elected politicians may perceive different levels of oversight by the PGN. We show that because entities headed by elected politicians report at substantially higher rates than entities headed by unelected bureaucrats, positive selection into reporting is present but is greatly attenuated in magnitude (Table A7).

5.3 Misreporting of Transparency Index Data

To this point in the empirical analysis, we have not considered the accuracy of the data submitted by entities. 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 transparency index data, which is post-treatment with regard to the communication treatments. While we include the treatments as covariates in various regression specifications, the coefficients do not estimate well-defined causal effects. As such, our analysis of data quality is descriptive.

We first show that our measure of compliance with audited transparency practices correlates strongly with self-reported measures of compliance. We consider two self-reported measures.

	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

1em^{+} p $<$ 0.1, * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

Table 5: The association between audited and reported scores. Heteroskedasticity robust standard errors in parentheses.

First, we work from the available public reports to construct the reported compliance with the same subset of measures 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 5 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.[12] 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 audits, 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 audits. To this end, we seek to measure both the extent of intentional distortions and noise in reporting.

We now consider the possibility of intentional misreporting. Figure 4 examines the relationship between audit-measured transparency practices and reported transparency practices. In the

[12]If the relationship between reported and audited scores were non-monotonic, it is possible that we could estimate $\beta_{\text{Audit}} = 0$. We have no reason to believe that this relationship is non-monotonic.

top row, we plot our audit-based measure of transparency practices (θ) against their differences from reported transparency practices on the same subset of items ($r - \theta$). The generalized additive models plotted 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 bottom panel of plots 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.

The right panel of plots disaggregates distortions in the reports of bureaucrats from entities led by elected and non-elected leaders. We see substantially larger distortions in the scores of entities led by elected politicians. The difference in misreporting between institutions led by elected versus non-elected principals is significant in the middle range of scores (Figure A11). Recall that these entities report at substantially higher rates than entities led by unelected bureaucrats, especially among entities that exhibit low levels of compliance with the audited transparency practices. As such, over-reporting may serve as an alternative to selection out of reporting, particularly for poorly-performing entities.

Qualitative evidence supports the interpretation that entities (and their heads) deem the results of the ITA as high-stakes. The latter is especially true among elected officials who lead these offices. Beyond seeking to avoid oversight, some elected mayors have sought to leverage the automated ITA scores based on self-reported data to showcase transparent governance. For example upon completing the 2020 ITA matrix, the mayor of Neiva (Huila), a medium-size city, advertised the results of the ITA on several media outlets under the headline that the PGN had "given the mayor the highest rating on transparency index" (see, for example, here).[13] While this

[13]That mayor was recently found to have engaged in contracting irregularities with regard to

credit-claiming mechanism departs somewhat from our oversight-based story, it shows how the bureaucrat tasked with completing the ITA may have internalized entity-level incentives to misreport transparency practices in Huila.

5.4 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.[14] 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. Given the findings in Figure 3, if some types of poor-performing entities intentionally misreport while others do not, this variation in misreporting could increase the variance of the reported scores even if effort were held constant. If this were the case, we would expect greater reductions in the reported scores of elected entities given the higher rates of distortion. In Figure A12, we show that the standard deviation of reported outcomes decreases in audit-measured quality in both subgroups. Moreover, this decrease in the conditional standard deviation is not consistently larger for the entities led by elected politicians. As such, while variation in intentional misreporting that correlates with latent quality may help to drive these patterns, it does not account for the observed empirical patterns in isolation.

Alternatively, censoring of scores at 0 and 100 may mechanically lead to differences in variance

---

COVID-19 contracts.

[14]One can conceptualize this analysis as an effort to measure conditional heteroskedasticity of the reported scores.

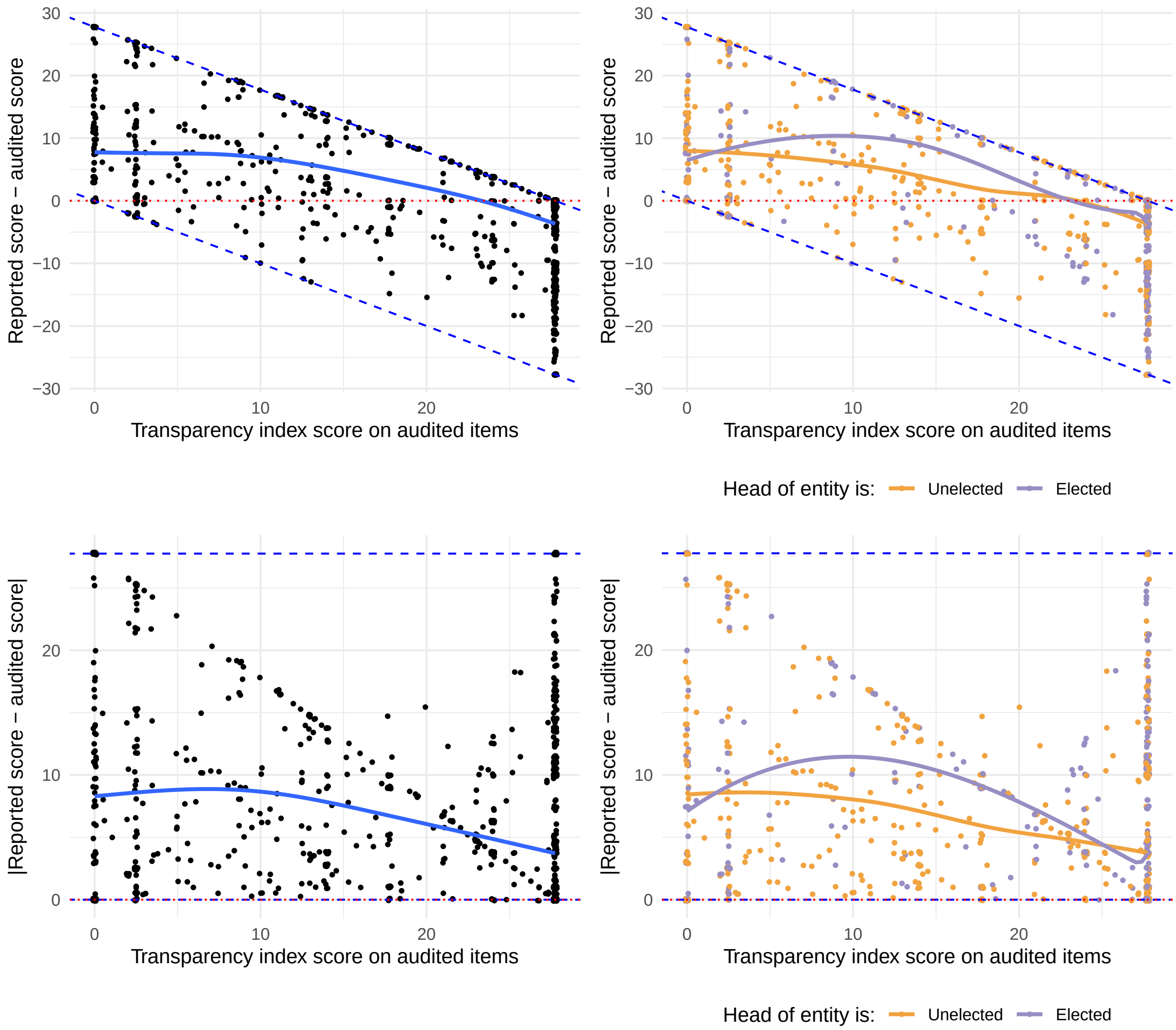

Figure 4: Discrepancies between the reported transparency practices and those detected in the audit for all audited entities (left). On the right, these discrepancies are decomposed between entities led by elected versus non-elected heads (right).

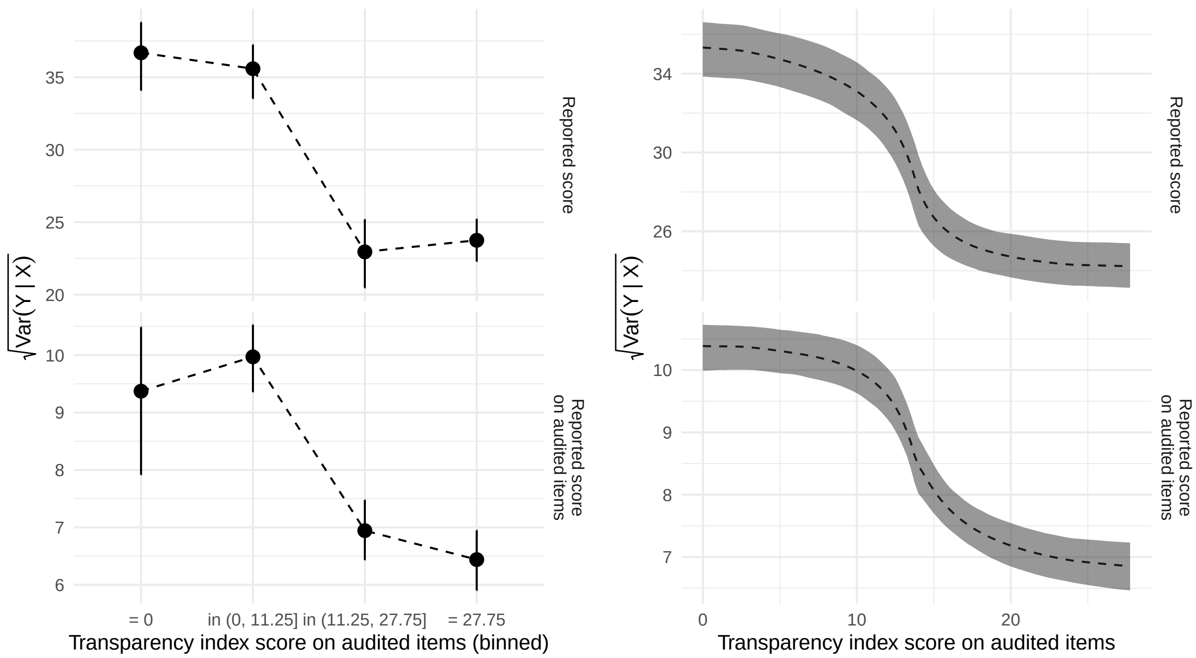

Figure 5: This plot shows how noise (standard deviation) in scores relates to the audit-measured level of transparency practices. The left panels bins entities by level of audit-measured transparency. The lowest three categories contain approximately 20% of the sample each and the top category contains 40% of the sample. The right panels employs a triangular kernal to estimate the conditional variance and reports the conditional standard deviation.

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 variance (or standard deviation). As such, censoring, in isolation, cannot explain the results in Figure 5.

Finally, transparency practices may correlate with institutional capacity. If this were the case, the cost of effort may covary with latent quality.[15] For example, bureaucrats in more transparent entities may have additional knowledge on transparency practices that facilitates accurate completion of the form. If this were the case, interpreting this outcome as a measure of aggregate effort may be misleading. Nevertheless, the observation that bureaucrats in low-performing entities issue reports with higher variance is important for understanding how the PGN can use this data to direct preventative monitoring functions.

6 Implications

We have elaborated a political process through which central states request data from decentralized entities. Our data studies the response of a large number of entities (mainly decentralized) to these requests in the context of Colombia's ITA matrix. By implementing an experiment, we have shown that small shocks to anticipated oversight of data submissions subtly changes how bureaucrats respond to these data requests from the central government. While this feature of our collaboration with the PGN facilitated the experiment, it limits our ability to understand how the design of data collection processes facilitates or mitigates the distortions to data quality that we document. Within our theoretical framework, the central government controls two policy instruments: the targeting of audits ($\rho(r)$) and the penalties imposed upon poor performance in an audit ($P(\theta, r)$). We consider three possible policy interventions that could potentially reduce the pathologies we observe in the data.

First, to incentivize completion of the ITA matrix, the PGN could target preventative action

[15]We hope to investigate this in future iterations of this paper.

against non-reporting entities. Figure 3 suggests that the transparency practices of the average non-reporting public sector entities are worse than those of reporting entities. This audit scheme uses only minimal information from the data collection activity, however, by auditing only on the basis of non-completion. By broadly communicating this use of the data, the PGN may be able to increase completion rates by imposing risks for non-completion. At the same time, this auditing strategy disregards much of the amassed data and may not deter purposeful misrepresentation among reporting entities.

Second, the PGN could conduct only random audits, auditing at a constant rate for any report, r. Importantly, by adopting this strategy, the PGN would commit to ignoring reported data entirely when monitoring the entity. Abstaining from using the reported data in preventative enforcement activities limits the utility of data to the government.[16] The goal of this policy would ostensibly reduce incentives for entities to inflate their compliance with transparency practices in an effort to avoid scrutiny. Yet, these gains may come with potential tradeoffs for completion rates in addition to the cost of less efficient targeting of audits.

Third, the PGN could impose greater penalties for distortion in the reporting of transparency practices. This could raise the expected cost of intentionally misrepresenting transparency practices or induce the bureaucrat to work harder to reduce the variance of possible unintentional errors. Yet, decentralized bureaucrats may still have incentives to misreport if the rate of monitoring depends on reported scores.

The tradeoffs associated with each of these policy interventions to improve the completion or accuracy of reported data suggest a need for additional theoretical work on how states can structure these policy instruments to motivate complete, accurate reporting. Importantly, these tradeoffs show limits to both the accurate collection and use of data, a central premise of calls for expanded use of data in governance.

[16] The ITA matrix may still be valuable because it makes transparency practices more accessible to the public.

7 Conclusion

Arguments for expansion of data-driven governance promote the collection and use of more data to inform policy decisions. Yet, large volumes of state data are produced in the course of interactions between actors who stand to benefit or lose from the use of the data they supply. Through a field experiment in collaboration with the Colombian Attorney Inspector General, we show that these incentives can distort the quality of the data that central governments ostensibly rely on to enforce laws or distribute resources. 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 contend that state data should be considered a *political* output. Given the high concentration of data production tasks within the scope of bureaucrats' tasks, measures of data production constitute important measures of bureaucrats' behavior. Second, we show that incentives in the production of state data, subtler than the targets described by Stathern (1997) and Goodhart (1983), 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.

Our discussion of the accuracy and quality of state administrative data presents important implications for empirical social scientists. While measurement error has been widely discussed in the case of survey data (i.e., Bound, Brown, and Mathiowetz, 2001) and for expert-coded data Rozenas (i.e., 2013); Fariss (i.e., 2014); Marquardt and Pemstein (i.e., 2018), expositions of the limitations of administrative data are less common. Broader acknowledgement of the limitations of administrative data produced by states – even in states known for relatively high-quality data – are important for understanding the limitations of our data and therefore our inferences.

References

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, Harvard University.

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

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

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

Department of Health & Social Care. 2020. "Busting bureaucracy: empowering frontline staff by reducing excess bureaucracy in the health and care system in England." Consultation outcome accessible at `https://tinyurl.com/42za26ku`.

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

Fariss, Christopher. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108 (2): 297–318.

Fizbein, Ariel. 1997. "The Emergence of Local Capacity: Lessons From Colombia." *World Development* 25 (7): 1029–1043.

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

Gómez Grajalez, Carlos. 2013. "Ancient censuses." *Signficance* p. 21.

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

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.

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

Landau, Martin. 1969. "Redundancy, Rationality, and the Problem of Dupliative Overlap." *Public Administration Review* 29 (4): 346–358.

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.

Marquardt, Kyle L., and Daniel Pemstein. 2018. "IRT Models for Expert-Coded Panel Data." *Political Analysis* 26 (4): 431–456.

Martínez, Luis R. 2021. "How Much Should We Trust the Dictator's GDP Growth Estimates?" *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2021-78).

Muralidharan, Karthik, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver. 2021. "Improving last-mile service delivery using phone-based monitoring." *American Economic Journal: Applied Economics* 13 (2): 52–82.

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

Rozenas, Arturas. 2013. *Advances in Political Economy, Institutions, Modeling and Empirical Analysis*. Springer chapter Inferring Ideological Ambiguity from Survey Data, 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.

Seabra, Sérgio. 2019. "How auditing can be directed against corruption: the case of Brazil's municipality-facing auditing programme." Building integrity Working Paper.

Singh, Abhijeet. 2020. Myths of Official Measurement: Auditing and Improving Administrative Data in Developing Countries. Technical report RISE Working Paper 20/042.

Slough, Tara. 2020. "Oversight, Inequality, and Capacity." Working paper, New York University.

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

Ting, Michael M. 2003. "A Strategic Theory of Bureaucratic Redundancy." *American Journal of Political Science* 47 (2): 274–292.

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 Worrking Papers on Public Goverrrnance 33 OECD.

Vrushi, Jon, and Robin Hodess. 2017. "Connecting the dots: Building the case for Open Data to fight corruption.".

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.