The Limits of Decentralized Data Collection: Experimental Evidence from Colombia

Tara Slough Natalia Garbiras-Díaz

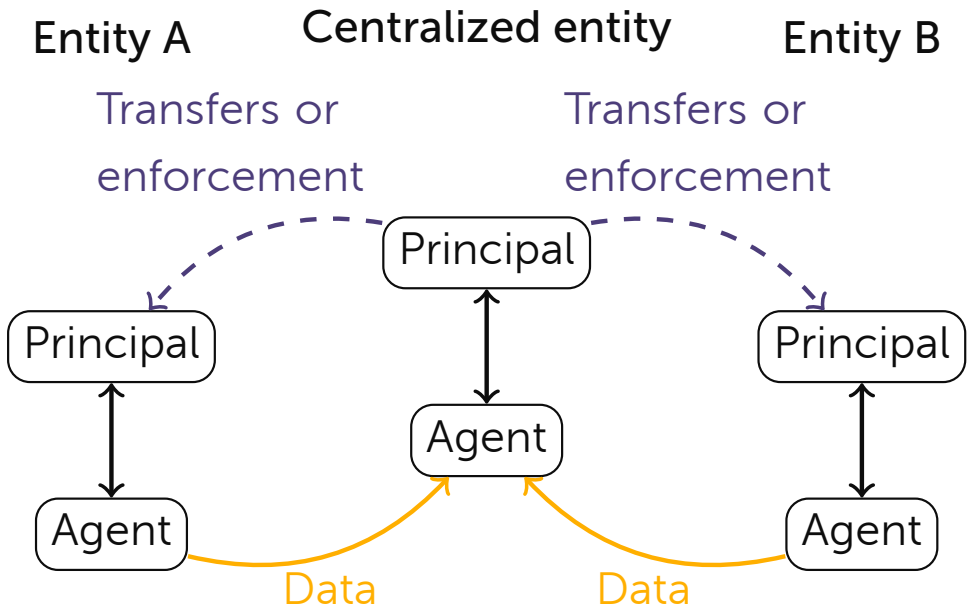
June 10, 2022

State data

- "Statistics" famously derives from the word "state."
- Modern states collect vast amounts of data for use in policymaking and public administration.
  - We see some of these outputs as **administrative data**.
  - This talk: **bureaucratic** data production.
- Policymakers and donors advance "data-driven governance" as a means to reduce corruption, waste, and inefficiency in resource allocation.
  - Posits that data inputs should be used to affect **policy**.

Decentralized data production

- Our focus: data produced by central government requests to sub-national/local government entities.
  - Data used to target spending or enforcement to decentralized entities.



- Is this really a thing?
  - Local bureaucrats complain about these requests and time-use surveys suggest that these requests are time consuming Kalaj, Rogger, and Somani (2022)
  - National bureaucrats seek advice on eliciting "honest" reports.

Overview

Framework: Simple decision theoretic model maps the behavior of decentralized bureaucrats onto statistical measurement framework.

- Bureaucrats choose how much **effort** to exert and whether to purposefully **distort** their reports to the central government.

Overview

Framework: Decision theoretic model that maps the behavior of decentralized bureaucrats onto statistical measurement framework.

Design: A field experiment in collaboration with the Colombian Attorney Inspector General's office on the annual national transparency index (ITA).

- Manipulation of the visibility of this watchdog institution allows us to measure the responses of bureaucrats in (mostly) decentralized entities.
- An independent audit of a subset of items allows us to describe data quality.

Overview

Framework: Decision theoretic model that maps the behavior of decentralized bureaucrats onto statistical measurement framework.

Design: A field experiment in collaboration with the Colombian Attorney Inspector General's office on the annual national transparency index (ITA).

Findings:

- Reporting behavior responds to the visibility of oversight in both selection into reporting and the information reported.
- The audit suggests that three pathologies of measurement: selection, intentional distortions, and the random error in reports covary with true quality.

Related Literature

- Complements literature on state data collection on individual citizens like censuses and vital statistics (Scott, 1998; Lee and Zhang, 2017; Bowles, 2020; Sánchez-Talanquer, 2020)
- Generalizes discussion of administrative data quality from two common settings:
  - Autocratic regimes, esp. with respect to economic data (Guriev and Treisman, 2019; Martínez, 2021; Trinh, 2021; Lorentzen, 2014; Wallace, 2016; Edmond, 2013)
  - Police data quality in the US (Eckhouse, 2022; Cook and Fortunato, 2022)
- Application: collection of data on corruption and transparency of state institutions (Ferraz and Finan, 2008; Larreguy, Marshall, and Snyder, 2018)

Framework

Administrative data

- Decentralized entities are asked to report some quantity, θ, to the central government.
  - E.g., public service outputs, budget execution, or transparency practices.
- A bureaucrat (or office) in the entity chooses whether to report, r.
  - If they do not make a report, $r = \varnothing$.
  - If they do report, $r \in \mathbb{R}$.
- Reported data, r can be different from θ due to:
  - Intentional distortion, d.
  - Unintentional errors/random error, $\epsilon \sim f(\cdot)$, where $f(\cdot)$ is a mean-zero density.

Data outputs

- Maps onto standard formulations of measurement error in statistics (e.g., Cochran, 1968; Rubin, 1976)
- The central government wants to know θ but observes r, which can suffer from:
  - Missingness when the bureaucrat does not report $r = \varnothing$.
  - Measurement error due to:
    - Intentional distortions (d)
    - Unintentional distortions (ϵ)

$$\underbrace{r}_{\text{Report}} = \begin{cases} \underbrace{\theta}_{\text{Truth}} + \underbrace{d}_{\substack{\text{Intentional} \\ \text{misreporting}}} + \underbrace{\epsilon}_{\substack{\text{Random} \\ \text{error}}} & \text{if bureaucrat reports} \\ \varnothing & \text{otherwise} \end{cases}$$

Central government data use

- Central government may use data to:
  - Target resources:carrots.
  - Target oversight or enforcement: sticks → our empirical setting.
- Key assumption: bureaucrats internalize (to some degree) entity outcomes.
- Two policy instruments by the central government in enforcement setting → exogenous in the experiment.
  - Reliance on observed data to target oversight with probability $\rho(r) \in (0, 1)$.
  - Penalties imposed when oversight reveals poor outcomes (θ) or reporting behavior (r): $P(\theta, r) > 0$.
- Governments struggle to set these policies, so much so that they are (sometimes) willing to randomize!

What are bureaucrats doing?

- When reporting, bureaucrats exert effort and choose intentional distortions
  - Effort, $e \geqslant 0 \rightarrow$
    - Missingness: when $e = 0$.
    - Unintentional distortions: $\uparrow e \rightarrow \downarrow \text{Var}(\epsilon)$.
  - Intentional distortions, d
- Bureaucrat's utility:

$$U_B = \underbrace{\underbrace{-\rho(r)}_{\text{Pr(Audit)}} \underbrace{P(\theta, r)}_{\text{Penalty}}}_{\text{Oversight}} - \underbrace{c(e)}_{\text{Cost of effort}}$$

- Note: bureaucrats may not know precisely how oversight is exercised.

Learning about the observed administrative data

- An exogenous increase in (perceived) oversight:
  - Should ↑ **rates of reporting**.
  - Changes the aggregate distribution of reports, conditional on reporting, by:
    1. ↑ entities reporting;
    2. Changing incentives for misreporting;
    3. ↓ variance of reports by increasing bureaucratic effort.

- Observation of the joint distribution of quality, θ, and reported quality, r, allow for learning about:
  - Bureaucrats' reporting behavior, including both effort and intentional distortions.
  - Bureaucrats' expectations about oversight by the central government.

Case Context

Our partner: The PGN

- The Office of the Attorney-Inspector General (Procuraduría General de la Nación), PGN, is the principal watchdog agency in Colombia.
- PGN is implementing of the National Transparency Law of 2014.
  - This law mandated the creation of the National Transparency Index (ITA), the measure that we study.
- PGN is also the principal user of the ITA data, as part of its preventative mandate.
  - This mandate seeks to prevent corruption or other public misconduct by monitoring of public officials and entities.

The National Transparency Index (ITA)

- The National Transparency Index, (Indice de Transparencia y Acceso a la Información), ITA, was first implemented in 2018.
  - We study the production of the 2020 index.
- $\approx 50k$ entities to report their compliance with ≈ 200 transparency practices.
  - All items are binary (yes/no).
  - Weighted to generate the 100-point ITA.
  - Used in PGN's preventative actions to guide monitoring/investigation.
- Unit of measurement, the entity, classified as:
  1. Traditional public sector entities including public entities, oversight bodies, and state-owned companies.
  2. Non-traditional entities are persons or legal entities that contract with the state to provide public services or manage public funds.
  3. Political parties or social movements.

Research Design

Overview

1. A field experiment conducted in collaboration with the PGN in the collection of the 2020 ITA index. Allows us to answer:
    - How does the reporting behavior of bureaucrats respond to changes in the salience of oversight?

2. An independent audit of a subset of responses provided by a random sample of public sector (traditional) entities. Allows us to answer:
    - How does the reported data we observe relate to the true measures of interest?

Experimental design, part 1

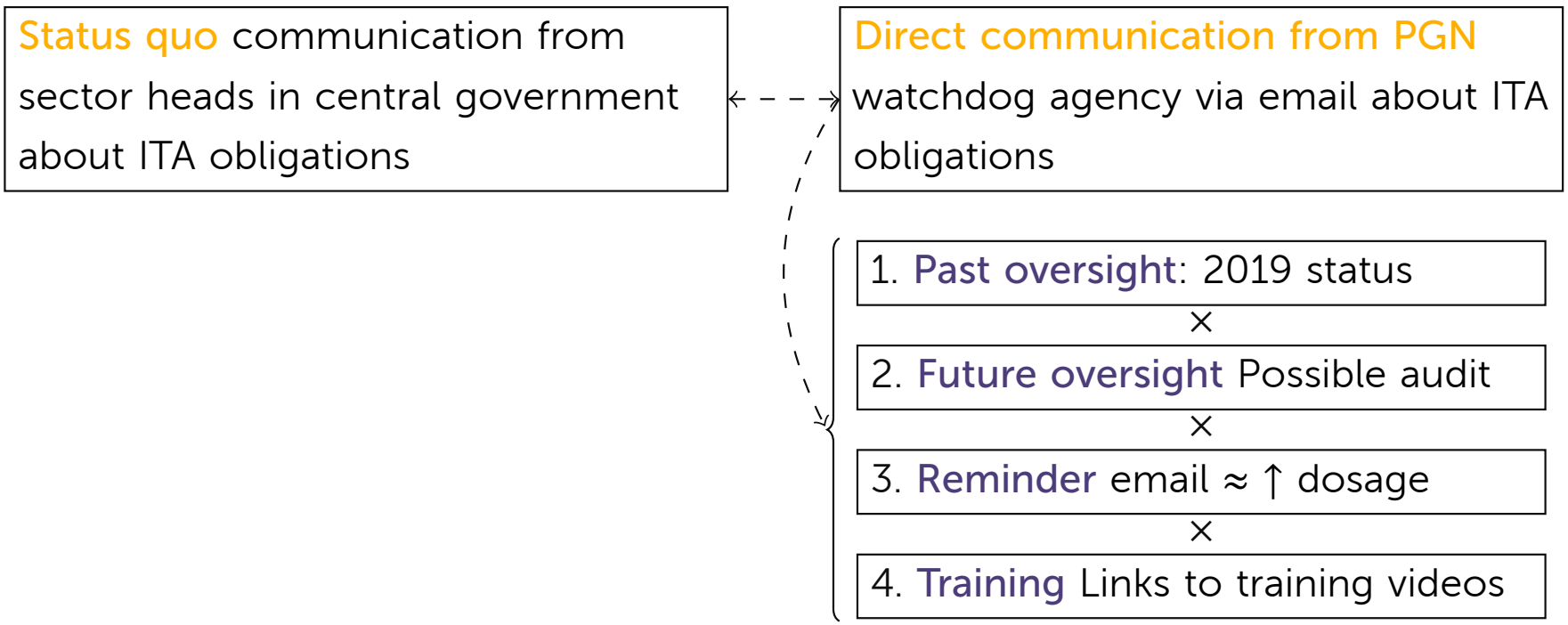
- Sample consists of 12,053 entities:
  - **Public sector** entities: 6,556 (near universe) of "traditional" entities
  - **Other** entities: 5,329 private sector + 168 political parties/social movements

Status quo communication from sector heads in central government about ITA obligations	← – – → Direct communication from PGN watchdog agency via email about ITA obligations

- Interpretation of contrast and ATE:
  - Make PGN's role and use of the data more **visible**.
  - From ex-post semi-structured interviews with bureaucrats who filled out ITA:
    - PGN is known as a watchdog agency, seen as having some teeth.
    - Entities regularly asked to send data to different central government entities.
    - ... but PGN has distinct oversight powers.

Experimental design, part 2

- Additional light-touch manipulations to content of direct communications randomly assigned in $2 \times 2 \times 2 \times 2$ factorial design



- Interpretation of message content/AMCEs:
  - Change subjects' beliefs about likelihood of oversight (#1-#3)
  - Training videos may reduce costs of effort.

Independent audit design

- Stratified random sample of 2,400 of 6,556 public sector entities
  - Crucially, we sampled entities regardless of whether they **completed the 2020 ITA**.
  - So we observe reporters and non-reporters symmetrically in the audit.
- Provides **validation** of self-reported transparency practices through an independent audit of a subset of index items:
  - Conducted outside partnership with PGN.
  - Audited items worth 27.75/100 points on the index.
  - We **measure** θ (quality) for this subset of the index.

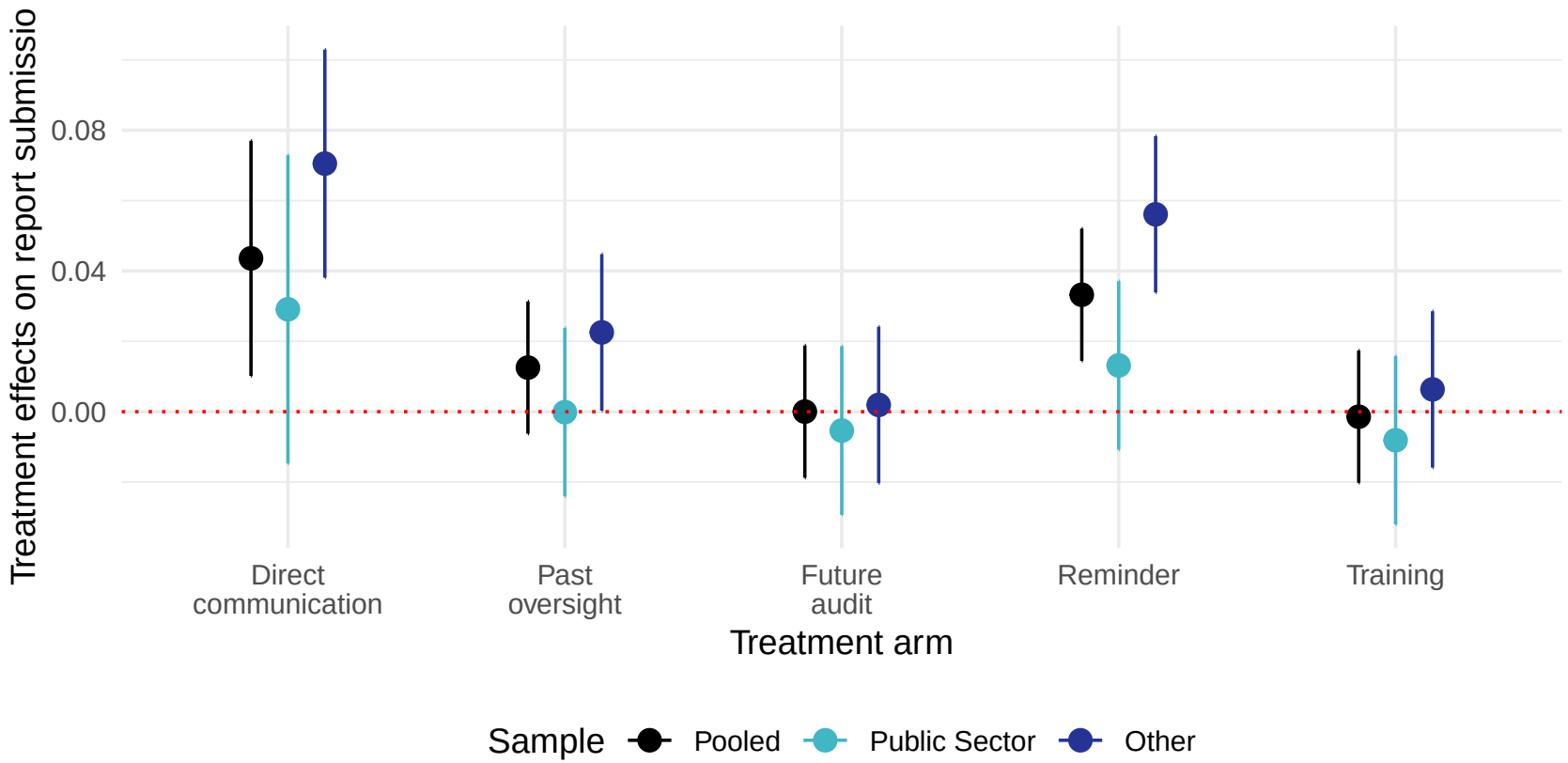
Measurement of parameters of interest

- Reported data (r):
  - Submission of the ITA matrix (yes/no)
  - ITA score (0-100) conditional on submission.
- Quality (θ for subset of index):
  - Use index weighting scheme from full index to construct a "true" score between 0 and 27.75 points.
  - We can also reconstruct the reported score on this subset from micro-data.
- Mapping to (unobserved) bureaucratic behavior:
  - Recall that $r - \theta = d + \epsilon$ (intentional + unintentional error).
  - By assumption $E[\epsilon] = 0$, so we will examine $E[r - \theta]$.
  - Link between variance of r and effort (e).

Results

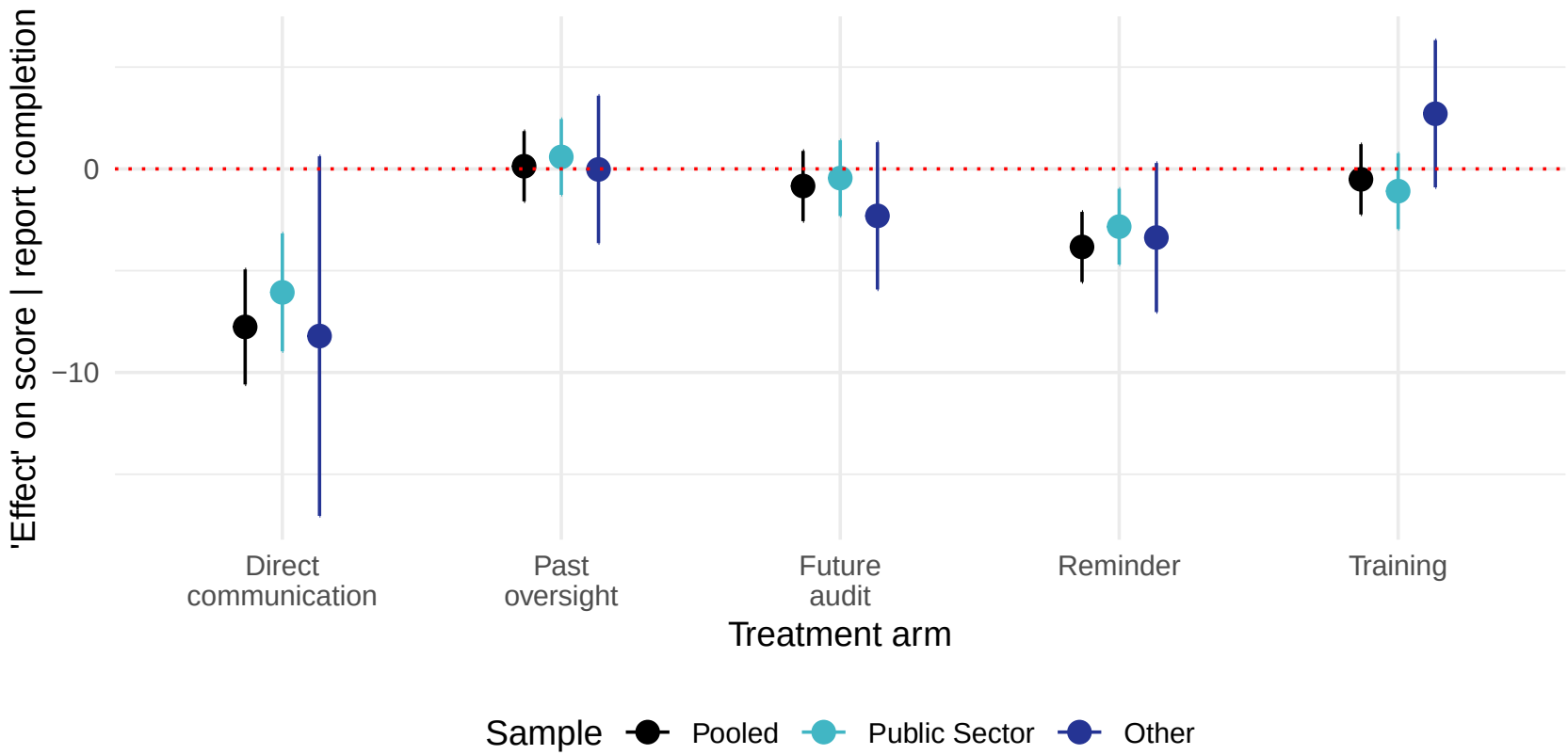
Effects on reporting

- We estimate the ATE of direct communication and the AMCEs of the messages in the factorial design. Estimator
- Outcome: Indicator for ITA data submission.



"Effects" on reported scores

- We estimate the same specification with scores as the outcome, but condition on reporting.
- Estimates of a post-treatment estimand that contains both a causal effect and a "selection" effect.



Decomposing the "effect" on scores

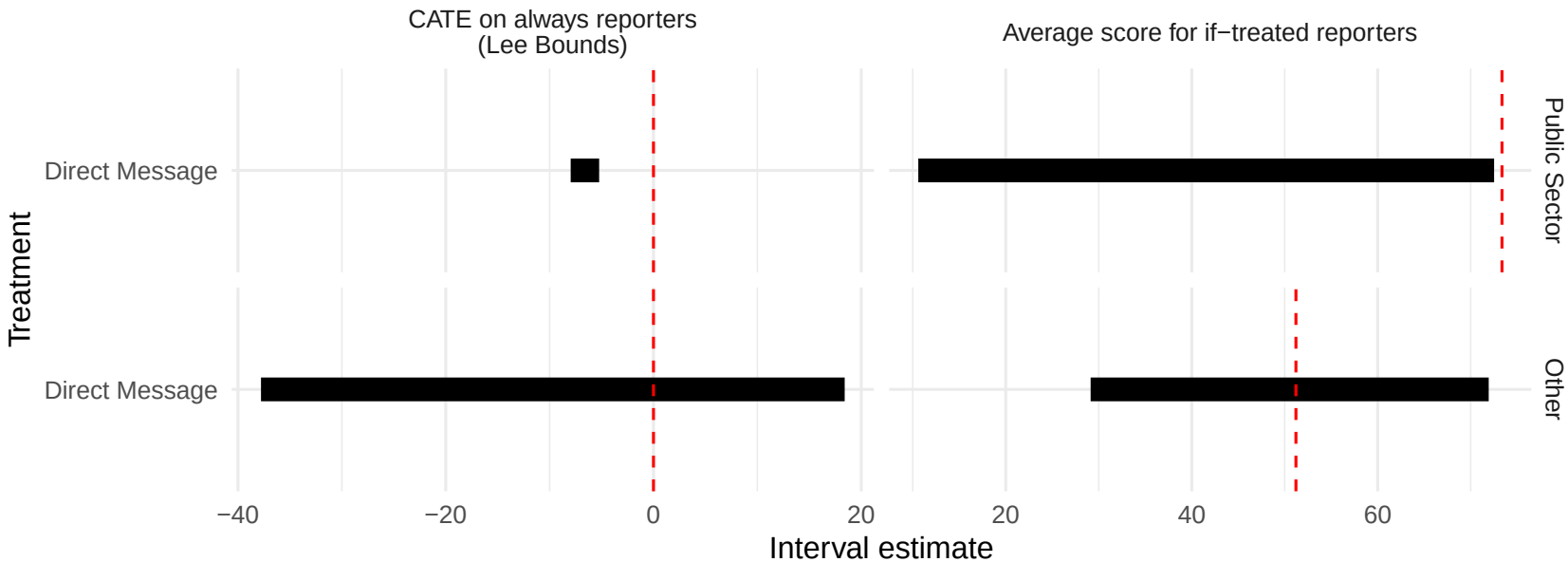
- Post-treatment estimates suggests that direct communication $\rightarrow\downarrow$ scores, but there are two possible explanations.
  1. **Treatment effect**: Those that would always report submit lower scores when subjected to oversight.
  2. **Selection/compositional change**: Those that report because of treatment report lower average reported scores than those that always report.

- Assuming monotonicity, we can **decompose** post-treatment estimand into:
  - **CATE** on always-reporters $\rightarrow$ Lee (2009) trimming bounds!
  - **Average outcomes** (scores) of if-treated reporters $\rightarrow$ boundable if we have post-treatment estimate, treatment effect on reporting, and CATEs.

Decomposition

Treatment effects or selection?

- For public sector entities, **both**.
- CATEs < 0 imply $\uparrow$ oversight $\rightarrow \downarrow$ scores among always reporters.

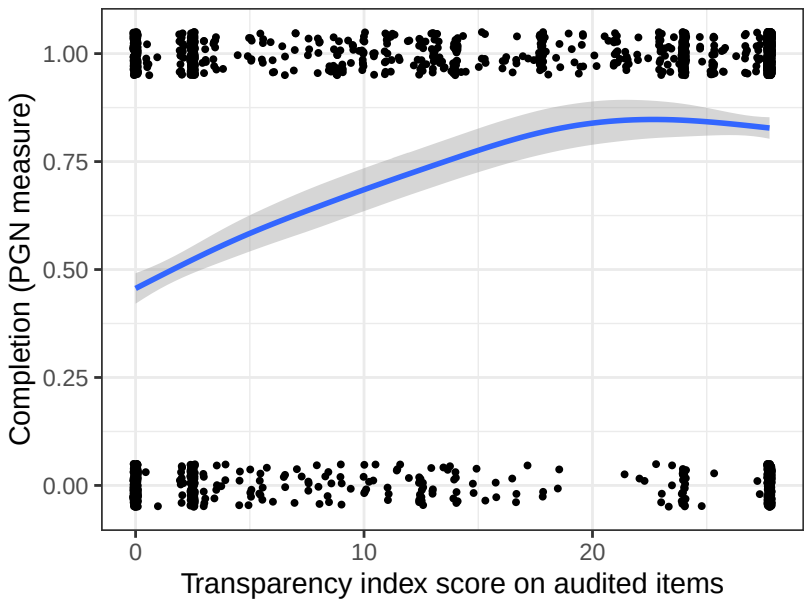


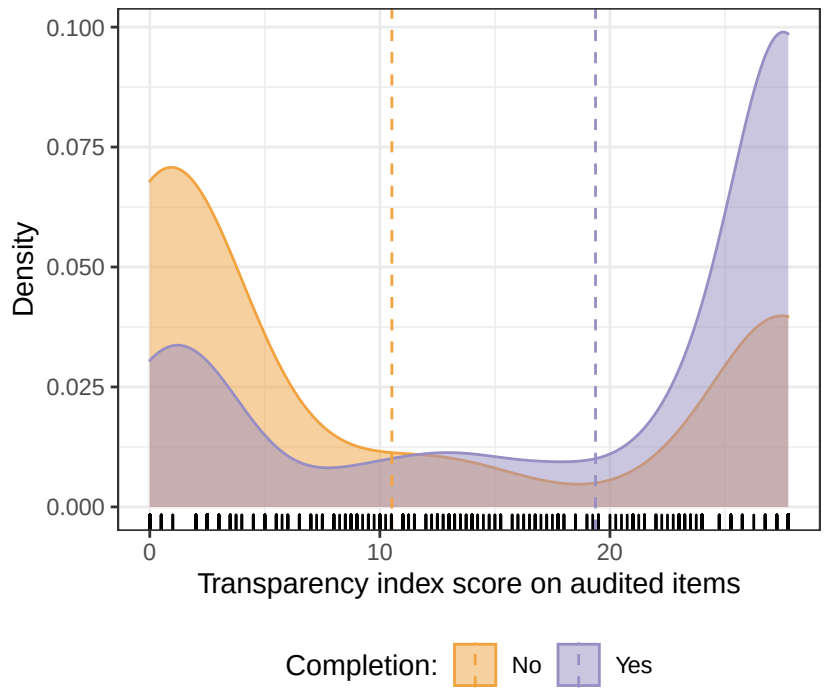
What do/don't we learn from the experiment?

- Fairly subtle manipulations of the role of the PGN/visibility of oversight:
  - Induce public sector entities to report **less desirable** scores.
  - Induce "other" entities to report at **higher rates**. Effect is stronger than for public sector entities.
- What we cannot observe from the experiment:
  - Who **selects** into reporting?
  - **Accuracy**: how do reported scores relate to actual quality?

Which entities select into reporting?

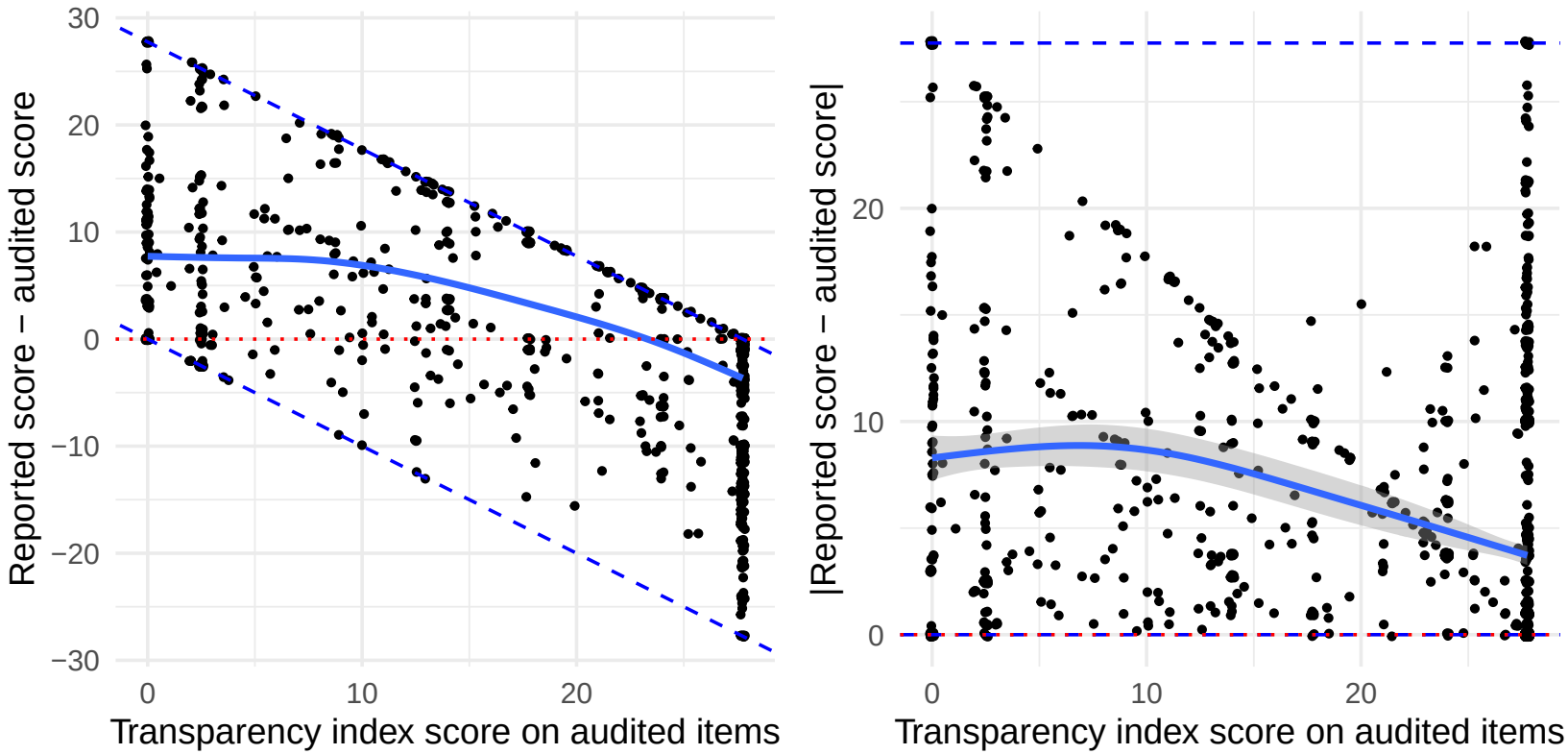
- **Positive selection** into reporting as a function of "true" transparency practices, θ.





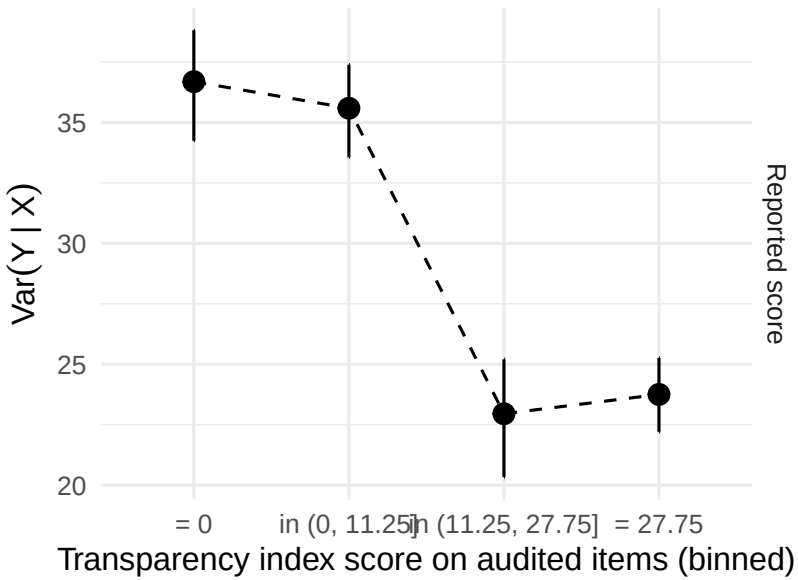
Are reported scores accurate?

- At lower levels of θ, reported scores are less accurate.
  - In general, scores are over-reported.



When are reported scores less noisy?

- Variance of reported scores is decreasing in θ.
  - Observed pattern not driven (exclusively) by state capacity or intentional distortion.
  - Within framework, suggests that lower θ entities exert less effort.



Discussion

Implications

- By experimenting, we have abstracted away from the central government's problem: When can data collected from decentralized entities be used to inform policies concerning those entities?
  - A hard problem and, anecdotally, a preoccupation of central government entities.
  - An understudied feature of intergovernmental relations.
  - Limits to "data-driven governance" as a efficiency-enhancing reform.

- Administrative data as a bureaucratic output and political outcome matters:
  - to you as a producer/consumer of empirical social science.
  - to governments that collect data in order to use it.

Thank you!

Estimator for experimental analyses

- We estimate the following treatment effects by OLS

$$\begin{aligned} Y_{ib} = &\beta_1 \text{Direct Communication}_i + \beta_2 \text{Reminder}_i + \beta_3 \text{Training}_i + \\ &\beta_4 \text{Retrospective Oversight}_i + \beta_5 \text{Prospective Oversight}_i + \psi_b + \epsilon_{ib} \end{aligned} \quad (1)$$

- The estimands of interest are:
  - β_1: The ATE of direct communication
  - β_2, β_3, β_4, β_5: The AMCEs of the factorial message treatments.

Decomposition of post-treatment estimand

- The "post-treatment" estimand is $\mathscr{P}$. We can express this quantity as:

$$\mathscr{P} = \underbrace{\frac{\pi_A}{\pi_A + \pi_T}\Big(E[S(Z=1)|j=A] - E[S(Z=0)|j=A]\Big) +}_{\text{Change in scores reported}}$$

$$\underbrace{\frac{\pi_T}{\pi_A + \pi_T}\Big(E[S(Z=1)|j=T] - E[S(Z=0)|j=A]\Big)}_{\text{Change in composition of reporters}} \tag{2}$$

$$= \frac{\pi_A}{\pi_A + \pi_T} CATE + \frac{\pi_T}{\pi_A + \pi_T}\Big(E[S(Z=1)|j=T] - E[S(Z=0)|j=A]\Big)$$

- We have point estimates for: $\mathscr{P}$, $\frac{\pi_A}{\pi_A+\pi_T}$, and $\frac{\pi_T}{\pi_A+\pi_T}$.
- We can use Lee (2019) bounds to generate an interval estimate of $CATE$.
- We can then use algebra to generate an interval estimate of $\Big(E[S(Z=1)|j=T] - E[S(Z=0)|j=A]\Big)$.