Government Oversight and Inter-Institutional Legibility: Evidence from Colombia

Codebook

Natalia Garbiras-Díaz* Tara Slough†

December 11, 2025

This codebook describes the processed datasets included in this replication package. All datasets described here are located in the `data` folder in the CodeOcean capsule or on Dataverse.

`exp_data.Rdata`

`exp_data.Rdata` (located in `data/`) contains one dataframe necessary to conduct all experimental analyses. See Table C1 for a list of variables. All data is original to the present experiment. This includes data that was shared by our national government partner, the PGN.

`audit_processed.Rdata`

`audit_processed.Rdata` (located in `data/`) contains one dataframe necessary to conduct most analyses of the audit of the ITA items. See table C2 for a list of variables. All data is original to the present study with the exception of:

- `record` and `distortion` are coded from microdata scraped from the PGN's website as of December 2020.
- `mdm_std` and `idi` are the 2020 *Medición de Desempeño Municipal* (standardized) and the 2020 *Indice de Desempeño Insitucional*, both collected by the Departamento Nacional de Planeación.

`siri.Rdata`

`siri.Rdata` (located in `data/`) contains four variables from the Procuraduría General de la Nación's Sistema de Información de Registro de Sanciones y Causas de Inhabilidad (SIRI). The data was obtained by an author's FOIA request. Because the data contains personally identifying information (*cédula* number and name), we (1) anonymize a unique ID number for each individual who appears; and (2) report only the variables necessary to construct Table A3 (the distribution of sanctions across types of officials). See Table C3 for a list of variables.

`processed_penalties.Rdata`

`processed_penalties.Rdata` (located in `data/`) contains two dataframes used in the motivation of the paper and in a balance test (Figure A2) on past PGN investigation of and or sanction of officials in select entities. These data come from two sources, which we match to the entities in our study via fuzzy name matching. The sources, by dataframe, are as follows:

*ngarbirasdiaz@hbs.edu
†tara.slough@nyu.edu

- `pub_sector_siri`: This data comes from the Procuraduría General de la Nación's Sistema de Información de Registro de Sanciones y Causas de Inhabilidad (SIRI). We have used fuzzy text matching to match these sanctions to the public sector entities in which implicated officials served. We report counts of sanctioned individuals at the entity level for all public sector entities. (Some have not been sanctioned during the time period covered by SIRI.) See Table C4 for a list of variables.
- `pd_merged`: We use data on investigations by the Procuraduría General de la Nación on entities headed by an elected principal to measure the frequency of investigation. We use fuzzy text matching to match these investigations to the public sector entities in our dataset. The data include counts of investigations an disciplinary investigations. See Table C5 for a list of variables.

`crossnational_indices.Rdata`

`crossnational_indices.Rdata` (located in `data/`) contains a number of country-level covariates used in Table A1. See Table C6 for a list of variables. The sources, by variable, are as follows:

- Income and regional classifications (`region` and `incomelevel`): World Bank
- Regional authority index (`n_RAI`): Hooghe et al. (2021)[1], using 2018 values
- Division of power index components and horizontal accountability index(`v2x_feduni`, `v2elrgpwr`, `v2ellocpwr`, `v2x_horacc_osp`): Varieties of Democracy, available at `https://www.v-dem.net/`, using 2020 values
- Rule of Law Index (`ROL`, `Absence`, `Open`): World Justice Project, available at `https://worldjusticeproject.org/rule-of-law-index/`, using 2020 values
- Corruption Perceptions Index `CPIscore2020`: Transparency International, available at `https://www.transparency.org/en/cpi/2024`, using 2020 values

`audit_weights.csv`

`audit_weights.csv` (located in `data/`) contains the weighting of the ten audited index items on the ITA index. See Table C8 for a list of variables.

`audit_loo.Rdata`

`audit_loo.Rdata` (located in `data/`) contains alternate measures of the audited ITA matrix outcomes that leave out one of the ten index component. These are computed from the audit-measured data described in Table C8. See Table C9 for a list of variables.

[1] Hooghe, Elisabeth, Gary Marks, Arjan H Schakel, Sara Niedzwiecki, Sandra Chapman-Osterkatz, and Sarah Shair-Rosenfield. 2021. "Regional authority index (RAI) v. 3."

Variable	Description	Complete obs.	Minimum	Maximum
`unique_id`	Unique identification number for each entity in the experimental sample.	12,053		
`block_ids`	Identification codes for the blocks of 18 entities used for random assignment. Note that the first letter (T, N, or P) corresponds to entity type, where T = public sector; N = private sector; and P = political party or social movement. The second letter (S or N) indicates receipt of the ITA index data in 2019 (pre-treatment), where S = yes and N = no.	12,053		
`dpto`	Department where entity is registered.	12,053		
`cod_dane_dpto`	Numeric code for department where entity is registered.	12,049		
`categoria_suj_obligado`	Category of entity by sector. *Tradicionales* refers to public sector entities; *No tradicionales* refers to private sector entities; and *Partidos políticos, movimientos políticos o grupo significativo de personas* refers to political parties and social movements.	12,053		
`orden`	Classification of public sector entities as national (*nacional*) or local/department-level (*territorial*). This does not apply to private-sector entities (though 8 non-traditional entities are mis-classified in the raw data and 391 public sector entities are not labelled).	12,053		
`naturaleza_juridica`	Legal classification of public sector entities and political parties. This does not apply to private sector entities.	12,053		
`entidad_fuente`	National government entity responsible for distributing call for ITA data in the absence of PGN communication.	12,053		
`elected`	Indicator that takes the value `1` if the principal(s) of a public sector entity are elected. This includes executive and legislative offices at the national, departmental, and local level.	12,053	0	1
`clasificacion_organica`	Classification of entity type among public sector entities. This does not apply to private sector entities or parties.			

nivel	Classification of public sector entitites under *Ley 489 de 1998* which defines entities with hierarchical dependence on the national government as "centralized" (*central*) and those with greater administrative autonomy as "decentralized" (*decentralizado*).	12,053		
en_dian	Indicates whether an entity registered in a database of the Dirección de Impuestos y Aduanas Nacionales, the national tax authority. "Si" indicates "yes" while "No" indicates "no."	12,053		
ita2019	An indicator that takes the value of 1 if an entity submitted the ITA matrix data in 2019 (the year before our experiment).	12,053	0	1
bin_type	Indicates whether an entity belongs to the set of public sector entities that we study in the main analysis. Note that this is equivalent to the "*tradicionales*" classification in categoria_suj_obligado, but collapses the private sector and political party categories to "Other."	12,053		
message	Treatment indicator for the direct communication treatment	12,053	0	1
oversight_past	Past oversight nudge treatment condition	12,053	0	1
future_audit	Future audit nudge treatment condition	12,053	0	1
training	Training resource nudge treatment condition	12,053	0	1
reminder	Reminder nudge treatment condition	12,053	0	1
complete	Indicator for submission of the ITA matrix data in 2020 (post-treatment) per data shared directly by the PGN	12,053		
score	Score reported for the ITA index among entities that submitted data in 2020.	5,496	0	100
no_firstemail_sent	Indicator for the first email to the entity bouncing (a measure of compliance). A "1" indicates that the email bounced. Note that this variable is only measured for entities assigned to the direct communication treatment.	10,714	0	1

Table C1: Variables in the file exp_data.Rdata.

Variable	Description	Complete obs.	Minimum	Maximum
`unique_id`	Unique identification number for each entity in the experimental sample.	2,400		
`block_ids`	Identification codes for the blocks of 18 entities used for random assignment. Note that the first letter (T, N, or P) corresponds to entity type, where T = public sector; N = private sector; and P = political party or social movement. The second letter (S or N) indicates receipt of the ITA index data in 2019 (pre-treatment), where S = yes and N = no.	2,400		
`cod_dane_dpto`	Numeric code for department where entity is registered.	2,400		
`orden`	Classification of public sector entities as national (*nacional*) or local/department-level (*territorial*). This does not apply to private-sector entities (though 8 non-traditional entities are mis-classified in the raw data and 391 public sector entities are not labelled).	2,400		
`elected`	Indicator that takes the value `1` if the principal(s) of a public sector entity are elected. This includes executive and legislative offices at the national, departmental, and local level.	2,400	0	1
`national`	Indicator that takes the value `1` when the entity is a national government entity and `0` when it is a local/department-level government entity. (A numeric version of `orden`).	2,400	0	1
`wt`	Weight that is given by the inverse of the probability of selection into the audit sample. Varies according whether an entity is national or territorial.	2,400	1.19	2.69
`record`	Indicator that takes the value 1 when there is a record of an entity's completion in the publicly-available microdata.	2,400	0, 1	
`complete`	Indicator that takes the value 1 when an entity submitted the ITA data according to the PGN's dataset (shared with the authors during the experiment).	2,400	0, 1	
`pts`	Total score on audited items, weighted according to the weights used to construct the ITA index	2,400	0	32.4
`score`	Total ITA index score according to the PGN's shared dataset (shared during the experiment)	1,696	0	100

distortion	Deviation between audit-measured pts and the reported score on audited items according to the microdata. A zero was imputed when no micro data were available.	2,400	-32.4	32.4
cut_pts	Classification of audited score (pts) into four bins: pts $\in \{[0,3],(3,19.6],(19.6,32.4),32.4\}$	2,400	1	4
message	Treatment indicator for the direct communication treatment	2,400	0	1
oversight_past	Past oversight nudge treatment condition	2,400	0	1
future_audit	Future audit nudge treatment condition	2,400	0	1
training	Training resource nudge treatment condition	2,400	0	1
reminder	Reminder nudge treatment condition	2,400	0	1
idi	The National Department of Planning's 2020 Index of Institutional Performance (*Indice de Desempeño Institucional*/IDI). This is merged by fuzzy name match with the entity name. We locate records for 1190 entities.	1190	10	99.8
mdm_std	The National Department of Planning's 2020 Measurement of Municipal Performance (*Medición de Desempeño Municipal*/MDM). This is coded according to the municipality in which an entity is registered. This variable is a z-score transformation of the resultant distribution of MDM scores.	2,396	-2.55	1.98
a1_compliant	Indicator for whether entity was compliant with all components of the the first item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,377	0	1
a2_compliant	Indicator for whether entity was compliant with all components of the the second item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,379	0	1
a3_compliant	Indicator for whether entity was compliant with all components of the the third item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,367	0	1

a4_compliant	Indicator for whether entity was compliant with all components of the the fourth item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,346	0	1
a5_compliant	Indicator for whether entity was compliant with all components of the the fifth item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,365	0	1
a6_compliant	Indicator for whether entity was compliant with all components of the the sixth item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,372	0	1
a7_compliant	Indicator for whether entity was compliant with all components of the the seventh item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,363	0	1
a8_compliant	Indicator for whether entity was compliant with all components of the the eighth item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,366	0	1
a9_compliant	Indicator for whether entity was compliant with all components of the the ninth item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,376	0	1
a10_compliant	Indicator for whether entity was compliant with all components of the the tenth item in the audit. If this component was not required for the entity in question, it is missing (NA).	2,358	0	1

Table C2: Variables in the file in audit_processed.Rdata.

Variable	Description	Complete obs.	Minimum	Maximum
year	Year of sanction/disqualification from public office	52,982	2003	2023
alpha_id	Code identifying the individual who was sanctioned. (Note that this was randomly generated in order to anonymize cedula/identification numbers in the the raw data.)	52,982		
CALIDAD_PERSONA	Legal classification of sanctioned official	52,982		
CARGO	Position/title of sanctioned official	52,982		

Table C3: Variables in the dataframe siri.

Variable	Description	Complete obs.	Minimum	Maximum
unique_id	Identification of entity that corresponds to unique_id in the experimental and audit data.	6,556		
n_since_2013	Number of sanctions/disciplinary proceedings in entity between 2013 and 2019.	6,556	0	3546
n_officials_2013	Number of officials sanctioned/disqualified between 2013 and 2019.	6,556	0	1,483
n_ever	Number of sanctions/disciplinary proceedings in entity recorded in SIRI (i.e., 2003-2019)	6,556	0	3,620
n_officials_ever	Number of officials sanctioned/disqualified recorded in SIRI (2003-2019.	6,556	0	1,504
n_2019	Number of sanctions/disciplinary proceedings in entity in 2019.	6,556	0	176
n_officials_2019	Number of officials sanctioned/disqualified in 2019.	6,556	0	88

Table C4: Variables in the dataframe pub_sector_siri.

Variable	Description	Complete obs.	Minimum	Maximum
unique_id	Identification of entity that corresponds to unique_id in the experimental and audit data.	6,556		
type	Type of entity (i.e., *alcaldía*). Defined for the 2,259 entities with (an) elected principal(s).	2,259		
n	Number of investigations within an entity during the 2016-2019 period. Missing values indicate no investigation reported or an entity outside the class with elected principal(s).	1,528	1	64
n_implicated	Number of individuals implicated in investigations during the 2016-2019 period. Missing values indicate no investigation reported or an entity outside the class with elected principal(s).	1,528	1	31
n_past_prelim	Number of investigations that proceeded to a disciplinary investigation during the 2016-2019 period. Missing values indicate no investigation reported or an entity outside the class with elected principal(s).	1,528	1	20
n2019	Number of investigations within an entity in 2019. Missing values indicate no investigation reported (during 2016-2019) or an entity outside the class with elected principal(s).	1,528	0	35
n_implicated2019	Number of individuals implicated in investigations in 2019. Missing values indicate no investigation reported (during 2016-2019) or an entity outside the class with elected principal(s).	1,528	0	30
n_past_prelim2019	Number of investigations that proceeded to a disciplinary investigation in 2019. Missing values indicate no investigation reported (during 2016-2019) or an entity outside the class with elected principal(s).	1,528	0	19

Table C5: Variables in the dataframe pd_merged.

Variable	Description	Complete obs.	Minimum	Maximum
CountryCode	World Bank country code	190		
country_name	Country name	190		
region	Region in which country is located	190		
incomelevel	Income group of country (lower income [LIC], lower middle income [LMC], upper middle income [UMC], or high income [HIC]) according to the World Bank	184		
incomelevel	Income group of country (in words) according to the World Bank	184		
n_RAI	Measure of the extent of authority exercised by regional governments across dimensions such as policy scope, fiscal autonomy, and representation.	95	0	37.67
v2x_feduni	Captures the vertical dispersion of power in a country. The composite index ranges from 0 to 1 and is based on expert-coded assessments of institutional features.	179	0	.997
v2x_elrgpwr	Subcomponent of v2x_feduni that measures the relative power of regional-level government institutions.	137	-2.55	3.09
v2x_ellocpwr	Subcomponent of v2x_feduni that measures the relative power of municipal-level government institutions.	169	-2.64	2.95
v2x_horacc_osp	Measures the degree to which state institutions (e.g., courts, audit agencies, ombudsman) are capable of and willing to oversee and sanction government misconduct. It ranges from 0 (no accountability) to 1 (strong accountability).	179	0.03	0.99
ROL	Evaluates the extent to which laws are clear, publicized, stable, and applied evenly. This index incorporates factors such as constraints on govern- ment power, absence of corruption, and openness of government. It ranges from 0 to 1.	128	0.27	0.90
Absence	Subcomponent of ROL that measures the extent to which public officials in the executive, judicial, legislative, and security branches use public office for private gain.	128	0.16	0.94
Open	Subcomponent of ROL that measures the openness of government, defined by the extent to which it shares information, empowers people with tools to hold the government accountable, and fosters citizen participation in public policy deliberations	128	0.22	0.88
CPIscore2020	Captures expert perceptions of public sector corruption on a scale from 0 (highly corrupt) to 100 (very clean).	180	12	88

Table C6: Variables in the dataframe indices.

Variable	Description	Complete obs.	Minimum	Maximum
stub	Index referring to audited item	10		
criteria1	Section of ITA matrix	10		
item	Item(s) independently audited	10		
peso_overall	Total weight attributed to section of ITA matrix	10	20	80
peso_categoria	Proportion of overall section weight (percent) devoted to category of item	10	5	100
peso_subcategoria	Proportion of category weight (percent) devoted to audited items	10	10	50
total_points	Total points of index attributed to audited items. Is calculated by peso_overall × peso_categoria/100 × peso_subcategoria/100.	10	0.6	10

Table C7: Variables in the data set audit_weights.csv.

Variable	Description	Complete obs.	Minimum	Maximum
stub	Index referring to audited item	10		
criteria1	Section of ITA matrix	10		
item	Item(s) independently audited	10		
peso_overall	Total weight attributed to section of ITA matrix	10	20	80
peso_categoria	Proportion of overall section weight (percent) devoted to category of item	10	5	100
peso_subcategoria	Proportion of category weight (percent) devoted to audited items	10	10	50
total_points	Total points of index attributed to audited items. Is calculated by peso_overall × peso_categoria/100 × peso_subcategoria/100.	10	0.6	10

Table C8: Variables in the data set audit_weights.csv.

Variable	Description	Complete obs.	Minimum	Maximum
unique_id	Identification of entity that corresponds to unique_id in the experimental and audit data.	2,400		
pts_loo_x	Audit score excluding item x	2,400	0	31.8 (varies)
dist_loo_x	Difference between audit score and score reported in microdata excluding item x	2,400	-31.8 (varies)	31.8 (varies)
rep_loo_x	Score reported in microdata excluding item x	2,400	0	31.8 (varies)

Table C9: Variables in the dataframe audit_loo.Rdata. Note that the loo_df data frame contains 10 variables per metric. They correspond to the omission of each of the 10 audited items described in Table C8. We index these variables by x to reduce repetition in the description of the contents of this dataframe.