Which Comparison Counts? Information, Benchmarking, and Electoral Accountability

Adam D Roberts

September 23, 2026

Click for latest version

Abstract

Field experiments have mixed findings on how and to what extent information changes voting behavior. Many factors may explain these mixed results, but one that remains under-explored is how benchmarking, or the presentation of performance information relative to other governments, can change interpretation. In a pre-registered survey fielded in Mexico with both descriptive and experimental components, I examine respondents' preferences for benchmarked information. I find that respondents prefer benchmark municipalities that are geographically proximate, in the same state, larger than their own, and governed by the same party coalition as their own municipality. When randomly assigned to treatment conditions that present robbery statistics for their municipality and vary whether partisan benchmarks are shown, I find that only same-coalition comparisons change respondents' intentions to vote for an incumbent, consistent with the theoretical case in which voters evaluate politicians based on competence. Beliefs about the incumbent's handling of crime shift in the same direction.

Introduction

Understanding the way voters interpret and process information is essential for both theoretical and empirical research that examines electoral accountability and government responsiveness. A large body of research in the democracies of developing countries examines whether information about politicians' performance affects electoral outcomes, but findings are extremely mixed (Ferraz and Finan 2008; Chong et al. 2015; Boas and Hidalgo 2011). There are many potential reasons for these mixed findings that have been explored, but less attention has been paid to how benchmarking of information itself may affect voting behavior. Benchmarks are very commonly used to make information interpretable in informational campaigns (Grossman and Slough 2022; Keefer and Khemani 2005), but limited theoretical or empirical study has been done to understand how the composition of the comparison group matters.

The Metaketa I project (Dunning et al. 2019), a coordinated set of pre-registered experiments, illustrates this lack of attention to benchmarking. Table 1 summarizes the information and the benchmark used in the four experiments from Metaketa I that shared performance information about politicians. Even within this deliberately harmonized design, the choice of benchmark varies considerably across studies. Only the pre-analysis plan for the experiment in Mexico explicitly considers how providing a benchmark differs from providing non-benchmarked information, and in all four studies the reason for choosing the specified benchmark is not made explicit. This is less a shortcoming of any individual study than a reflection of how little attention the concept has received.

To address the lack of attention on benchmarks, I present results from a pre-registered survey in Mexico to understand how voters in developing countries use information when deciding which candidate or political party to vote for. The survey has two components. First, I design a comparison municipality choice exercise to examine which municipalities respondents prefer as benchmarks. In the survey, respondents chose their preferred benchmark

Table 1: Information and benchmarks used in the Metaketa I studies (Dunning et al. 2019).

Study	Information provided	Benchmark
Benin	Legislative performance	Other legislators in the region and country
Mexico	Municipal unauthorized or misallocated spending	Municipalities in the same state governed by a different party
Uganda II	Budget irregularities	Other districts within the country
Brazil	Municipal accounts rejected by independent audit	Other within-state municipalities

municipalities from a candidate menu of 15 municipalities constructed to show municipalities that are likely to be relevant to respondents. I find that respondents prefer benchmark municipalities that are geographically proximate, from the same state, highly populated, and governed by the same coalition as the respondent's municipality. The preference for same-coalition municipalities is in part a result of respondents' actual knowledge of which parties govern nearby municipalities, rather than unmeasured dimensions of similarity between municipalities: respondents were more knowledgeable of the governing coalition of the municipalities they had chosen as benchmarks.

Second, I run an experiment in which respondents are randomly assigned to receive information about non-violent crime rates in treatment conditions that vary whether municipalities shown as benchmarks are labeled as from the same party coalition (MORENA/PT/PVEM vs. PAN/PRI/PRD vs. Movimiento Ciudadano) as their municipal government, from a different coalition, or unlabeled. I use non-violent crime because it is highly relevant to Mexican voters, because municipal presidents largely control police forces that address it, and because it is much more tractable at the municipal level than violent crime, which is often driven by nation-wide drug cartels. I find that only same-coalition comparisons change respondents' intentions to vote for an incumbent. The observed effect is robust to a variety of specifications. I show that the effect is partially explained by changes in beliefs about how well incumbents handle crime.

These results point to two major implications for research on electoral accountability

with informational treatments. First, varying the benchmark presented to voters can change how voters respond to the information, which may explain some of the mixed results in the literature. Second, same-coalition comparisons are theoretically well suited to filtering noise out of performance information, and I find that voters treat them as more relevant than opposition-coalition comparisons. Despite this, they are rarely used in informational campaigns and deserve stronger consideration in future research.

Background: benchmarking in political science

Scholars of economic voting have examined how economic performance relative to differing benchmarks affects voter behavior. These studies generally find that accounting for how voters interpret economic information is crucial for understanding the relationship between economic outcomes and voting behavior. Past work has found that interpretation of economic policies and outcomes can be altered by comparisons between sub-national governments (Besley and Case 1995), comparisons from local to national performance (Cohen 2020), and performance relative to expectations (Epp and Wlezien 2026). Observational studies examining whether economic performance relative to international benchmarks can explain domestic vote choice have mixed results (Kayser and Peress 2012; Aytaç 2018; Arel-Bundock, Blais, and Dassonneville 2021), but experimental studies that explicitly show cross-national comparisons to respondents have shown that these benchmarks strongly influence voters' perceptions of performance (Huang 2015; Hansen, Olsen, and Bech 2015).

There is reason to believe that the importance of benchmarking in the literature cited above would also hold in developing democracies across different performance areas. Voters across many developing democracies respond to information about incumbent performance when it is relevant to them (Gottlieb 2016; Adida et al. 2020; Banerjee et al. 2024). Bhandari, Larreguy, and Marshall (2023) find that Senegalese voters treated temporally benchmarked local performance outcomes as particularly informative and updated their beliefs about in-

cumbents in a Bayesian manner. Crime information, specifically, is a salient valence issue on which voters evaluate incumbents (Singer 2011), and in Latin American contexts voters have been shown to hold politicians accountable for insecurity when they can attribute responsibility for it (Ley 2017). Additionally, voters in developing democracies do not base their preferences solely on clientelistic favors (Cruz, Keefer, and Labonne 2021; Cruz et al. 2024). Across these performance areas, the effect of information hinges on whether voters can interpret it. This makes how that information is benchmarked a central question.

The most direct precedent for studying benchmarks in a developing democracy is Arias et al. (2024). They find that information about performance relative to other municipalities does not influence voters' beliefs about their own municipality's performance in Mexico. However, while they compare benchmarked to non-benchmarked information, they do not test whether the chosen benchmark was relevant to voters. This paper takes up that question by testing the relevance of different comparisons (e.g., same-coalition vs. opposition-coalition) and directly observing which comparisons voters find most relevant. The design directly addresses the lack of consensus on how to construct benchmarks, despite their widespread use (Grossman and Slough 2022; Keefer and Khemani 2005).

A model of benchmarking performance information

To understand how benchmarks may help voters interpret noisy performance information, I present a voter decision theoretic model that builds on the framework of Besley and Case (1992). The central problem facing voters is that most observable performance outcomes, like crime rates, are driven by both the incumbent's effort and factors outside any politician's control (e.g., economic shocks or regional crime trends that affect entire areas simultaneously). Because these common factors are unobserved, a voter who sees only their municipality's crime rate cannot tell whether a high crime rate reflects poor governance or "bad luck" (factors out of politicians' control that increase crime).

To illustrate this logic formally, consider the simple case where the crime rate r_h in the voter's home municipality h is the result of government competence, z_h, and an exogenous shock, θ, that is common to municipalities in the same region:

$$r_h = \theta + z_h$$

The exogenous shock θ can take a value of 0 (no shock), 1 (small shock), or 2 (large shock). Competence z_h takes a value of 0 (competent) or 1 (incompetent).

There are four possible realizations of crime rates: $r_h \in \{0, 1, 2, 3\}$. Of these four possible crime rates, only the extremes reveal competence information. The crime rate $r_h = 0$ can only be achieved by a competent government and the crime rate $r_h = 3$ can only be achieved by an incompetent government. For the crime rates in the ambiguous set $\mathcal{A} = \{1, 2\}$, the voter must rely on their prior beliefs about the prevalence of incompetence and the likelihood of a small shock.

The information voters can extract from the crime rate is enhanced when the voter observes their home crime rate against a benchmark municipality's crime rate (r_b) along with their own municipality's crime rate (r_h), where the benchmark municipality (b) experiences the same shock (θ) as the home municipality. In this case, the middling crime cases can be fully revealing if the municipal governments have different levels of competence. Consider the case where the politician in the home municipality is incompetent ($z_h = 1$) and the politician in the benchmark municipality is competent ($z_b = 0$). For any value of θ, the difference between municipalities simplifies to reveal just the differences in competence. Formally, this can be shown as $r_h - r_b = z_h - z_b = 1$, which fully reveals the home incumbent's incompetence. Thus, adding a benchmark is weakly more informative for all values of r_h.

Adding policy preferences. The baseline model above assumes that voters do not have preferences for things other than the level of crime. I will now consider cases where policy differences across candidates play a role in voter behavior. Suppose there are two policies

x: one policy has no tolerance for crime ($x = 0$), and the other policy tolerates some crime ($x = 1$) in order to dedicate resources to other policy goals. Crime in municipality $m \in \{h, b\}$ is now

$$r_m = \theta + x_m + z_m$$

where x_m is the policy of the politician governing m. A candidate's policy x is unknown to the voter unless revealed.

In this case, the possible values for the crime rate are now $r_h \in \{0, 1, 2, 3, 4\}$, and the ambiguous set expands to $\mathcal{A} = \{1, 2, 3\}$. Again, with no benchmark only the extremes are fully revealing. $r_h = 0$ means the politician is a competent, zero tolerance candidate and $r_h = 4$ means the politician is an incompetent, tolerant candidate. Any other case is ambiguous. For example, in the case of $r_h = 1$, this outcome could be observed under a competent zero tolerance candidate, an incompetent zero tolerance candidate, or a competent tolerant candidate. For $r_h = 2$, any combination of policy preference and competence is possible. Introducing policy differences introduces more ambiguity to the voter's inference problem for every candidate type, making inferences drawn from observing the crime rate even more dependent on voter priors about politicians.

A simple comparison of crime rates across municipalities no longer isolates competence, as it now reveals differences in x and z instead of isolating z. Formally, $r_h - r_b = (x_h - x_b) + (z_h - z_b)$. The comparison still improves on plain information, because it can reveal both policy and competence in some cases that would be ambiguous without a comparison. Consider the case where $(x_h, z_h) = (1, 1)$ and $(x_b, z_b) = (0, 0)$. Here, $r_h - r_b = 2$, which fully reveals the policy preference ($x_h = 1$) and competency ($z_h = 1$) of the incumbent. Writing $d = r_h - r_b \in \{-2, -1, 0, 1, 2\}$ for the observed difference, it is the extremes $d = \pm 2$ that are revealing in this way. The intermediate values are not: $d = \pm 1$ tells the voter that the two municipalities differ, but not whether the difference is one of policy or of competence, and $d = 0$ is consistent both with candidates who match on both dimensions and with candidates whose policy and competence differences offset each other.

Partisan comparisons. I will now consider that candidates come from one of two political parties, A and B, with policy positions x_A and x_B. Let $P(m) \in \{A, B\}$ denote the party governing municipality m, so that a candidate's policy is fixed by their party, $x_m = x_{P(m)}$. However, the parties need not differ in their policies, and individual candidates can vary on competence z_m within party. In other words, two candidates from the same party can differ in competence but must have the same policy preference, while two candidates from different parties can (but need not) differ on both policy and competence. Assume that the voter knows which party governs their municipality, but does not know which party governs a comparison municipality unless informed.

The addition of political parties does not change the informational context for the non-benchmarked case or for the non-partisan benchmark, because the voter will not know which party governs the comparison municipality. If the benchmark includes information about the comparison municipality's party, however, more information can be gleaned from the benchmark. If the benchmark municipality is revealed to be governed by the same party as the home municipality ($P(h) = P(b)$), then $x_h = x_b$ and $d = z_h - z_b$. The same-party comparison therefore returns the voter to the baseline model, in which the benchmark isolates competence when differences are observed.

For a comparison to an opposition party ($P(h) \neq P(b)$), the benchmark reveals a combination of policy and competence differences, $d = (x_h - x_b) + (z_h - z_b)$, just as in the case without party information. What the party label adds is the knowledge that the two candidates are drawn from different parties, and therefore that $x_h - x_b$ may be nonzero. By itself this does not resolve the confound: $d = \pm 2$ requires both components to take the same extreme value and so reveals policy and competence together, $d = \pm 1$ is consistent either with a policy difference and equal competence or with equal policy and a competence difference, and $d = 0$ is consistent both with candidates who match on both dimensions and with candidates whose differences offset.

The label becomes informative about competence when the voter holds a confident prior

about the size of the policy gap $x_h - x_b$, whatever that size may be. The simplest such case is a voter who believes the parties do not differ on crime policy: the gap is zero, $d = z_h - z_b$ outright, and the opposition comparison isolates competence exactly as the same-party comparison does. The less trivial case is a *directional* prior, in which the voter believes not only that the parties differ on crime policy but which of the two is the stricter one. Suppose the voter is confident that $x_h - x_b = \delta$ for a known $\delta \in \{-1, 1\}$. Then $d = (x_h - x_b) + (z_h - z_b) = z_h - z_b - \delta$. Because the competence difference must lie in $\{-1, 0, 1\}$, only three values of d can be observed, two of which pin down competence exactly. Take $\delta = -1$, the case in which the home municipality is governed by the stricter party. Observing $d = 0$ implies $z_h - z_b = 1$, meaning $(z_h, z_b) = (1, 0)$. In other words, the home municipality matched the benchmark's crime rate despite the advantage of a stricter policy, which is possible only if its incumbent is incompetent and the benchmark's is not. Only $d = \delta$, the difference that exactly matches the policy gap, leaves competence unresolved, and even there the voter learns that the two incumbents are equally competent, though not at what level.

Without a confident prior on the policy gap, the opposition benchmark is no more informative about competence than the unlabeled one. A voter who does not know whether the parties differ on crime policy, or who believes they differ but does not know in which direction, cannot separate $x_h - x_b$ from $z_h - z_b$ in any realization of d. Such a voter may still learn about policy rather than competence, but only noisily, since competence is equally likely in both parties and a single comparison cannot distinguish a party-level policy gap from an idiosyncratic competence gap.

These two cases, plus the case where voters assume the baseline model of no policy differences across party, illustrate how the choice of benchmark can change how informative the crime rate is to voters. If the voters do not believe parties differ on crime policy, then all three benchmarks are equally informative. If voters believe that the parties differ in policy but are interested in learning about the competence of their own government, the

same-party comparison delivers this competence information alone, so it will be the most informative. The non-partisan benchmark will be much less informative, because it does not specify which party the benchmark municipality is governed by. The opposition party benchmark identifies competence in ambiguous cases only for voters holding a confident prior about the policy gap, which for voters who believe the parties do differ means a directional prior about which party is tougher on crime. However, the opposition party benchmark is the only benchmark that explicitly compares the two parties, so the opposition comparison will be most informative to a voter who is interested in learning about the policies of parties. Table 2 presents these differences in informativeness by voter beliefs.

	Informativeness of benchmark		
Voter beliefs	Non-partisan	Opposition	Same-party
Parties alike	High	High	High
Differ, competence focus	Low	Prior-dependent	High
Differ, policy focus	None	High	None

Table 2: How informative each benchmark type is for the voter, under alternative voter beliefs about how parties differ.

Application to the survey. The model depends on the voter accepting the benchmark as an appropriate standard of comparison. If the voter does not believe that the benchmark municipality plausibly faced conditions similar to their own, there is no clean common shock to difference out. A benchmark the voter does find relevant, by contrast, provides the comparison needed to differentiate exogenous shocks from incumbent performance. Understanding which comparisons voters treat as relevant is therefore necessary for making predictions about how voters react to information. The first wave of the survey measures this directly by eliciting the benchmarks respondents would choose for themselves.

The experimental design allows me to test whether comparisons help voters attribute outcomes to incumbent performance rather than to factors outside the incumbent's control. It also tests for whether partisan benchmarks differ from non-partisan benchmarks. Specif-

ically, the difference between same- and opposition-party comparisons can clarify whether voters are focused on re-electing competent parties or learning about parties' platforms.

It is a real question of which of these two approaches applies to Mexican voters. On the one hand, political parties and coalitions in Mexican municipalities are quite distinct. There are three main coalitions: the nationally dominant left-wing MORENA coalition (which includes two smaller parties PVEM and PT); the big tent coalition consisting of PAN, PRI, and PRD; and the center-left Citizens' Movement (*Movimiento Ciudadano*, MC) party. Of all 2,469 municipal governments in 2026, only five rural municipalities were governed by a combination of parties from two of the three major coalitions (see Appendix 1). Thus, voters may already have strong priors about how parties differ on crime and therefore only be interested in implied differences in competence. On the other hand, Claudia Sheinbaum, Mexico's current president and a member of the MORENA party, has adopted a much harsher anti-crime approach than her predecessor of the same party. This may mean that voters are unsure about how coalitions differ on crime policy. Appendix 2 shows that, after controlling for population and area, crime rates are not statistically different by governing coalition. Thus, voters may be genuinely interested in learning which party is tougher on crime in 2026.

Survey Design

The survey has two main analyses. First, a descriptive analysis of the municipalities that are chosen by survey respondents as comparisons for their own municipality. Second, an experiment that varies whether partisan information is shown in benchmarked informational treatments to understand how informational benchmarks affect perceptions of incumbent politicians. Both designs were pre-registered on OSF.

The survey is fielded in two waves. Wave 1 collects pre-treatment beliefs and elicits respondents' preferences for comparison municipalities. Wave 2 recontacts the same re-

spondents approximately one week after completion of Wave 1 and assigns respondents to treatment, then collects post-treatment beliefs.

The two components address complementary halves of the argument. The descriptive choice-based analysis characterizes which benchmarks respondents treat as relevant. The experimental component then manipulates benchmark presentation, varying whether a comparison is shown at all and whether the partisan identity of the comparison municipality's governing coalition is made visible, in order to test how these features shape belief updating. The specific questions the experiment is designed to answer are stated in the Experimental Design section below.

Sample Composition

The study population is eligible voters in the 28 states holding municipal elections in 2027 for which the partisan treatments are well defined.[1] Within this frame, Wave 1 respondents were recruited against marginal quotas on sex, age, socioeconomic level (SEL), and state. These quotas were enforced at the start of fielding but relaxed as collection proceeded in order to reach the target sample size, so the achieved Wave 1 sample of 3,028 respondents differs slightly from population demographics. Wave 2 was not quota-sampled: it recontacted respondents who completed Wave 1, and the 1,927 respondents in the Wave 2 analysis sample are simply those who returned, identified the same home municipality in both waves, and passed the attention check. Wave 2's composition therefore reflects Wave 1's composition together with differential attrition.[2]

1. Four states are excluded: Durango and Veracruz hold no 2027 municipal elections, Mexico City is excluded because its borough presidents lack municipal police authority, and Oaxaca is excluded because a majority of its municipalities govern under *usos y costumbres*, which makes partisan treatments inapplicable. Additionally, respondents from municipalities that were not governed by one of the three major governing coalitions were ineligible for the survey.

2. Sex and state benchmarks are computed from INEGI 2020 census counts for the 28 states in the frame; age benchmarks are the INEGI 2020 distribution of the 18+ population in those states; SEL benchmarks are AMAI NSE shares from ENIGH 2022 (*Estimaciones NSE 2022* 2022). Because the SEL quotas set for Wave 1 were themselves not proportional to the population distribution, the comparisons below are drawn against population shares rather than against the field quotas.

On sex and age the departures are modest. Women are overrepresented by 2.5 percentage points in Wave 1 (53.6% versus 51.0%) and 2.8 points in Wave 2. Age is close to the population distribution in Wave 1, where no bracket deviates by more than 2 points, while Wave 2 skews somewhat older: the 18–24 bracket falls 2.3 points below its population share and the 45–54 and 55–64 brackets each run roughly 2 points above.

The departures on socioeconomic level and geography are much larger, and they run in the same direction. Both samples are wealthier, more urban, and more geographically central than the population. This has no bearing on the internal validity of the experiment, since treatment is randomized within the realized sample, but it does show that the poorest and most rural segments of the Mexican electorate are not particularly well represented among respondents. Appendix 3 reports the full comparison of both samples against population benchmarks on sex, age, socioeconomic level, and region.

Benchmark Municipality Selection Analysis

In Wave 1 of the survey, I seek to identify which comparison municipalities respondents view as most relevant. To do this, I design a municipality choice exercise. Respondents were given this instruction: "Municipal governments are often evaluated by comparing them to others. Which of the following municipalities would you compare [*name of respondent's municipality*] to?" Below this instruction respondents were shown a list of 15 municipalities before treatment and asked to select any they would like to use as a comparison to their own municipality. Respondents were required to select at least one municipality, and could choose all 15. The 15 municipalities are drawn from three pools: 5 randomly sampled from 10 geographically near municipalities (with a population of at least 22,000), 5 from the 20 most populous municipalities nationally, and 5 drawn at random from the full universe. Duplicates across pools are dropped and replaced by resampling. Figure 1 shows an example of how this selection tool looks to respondents. This design gives respondents a manageable

list of municipalities that does not systematically privilege any single comparison dimension (geographic proximity, city size, or randomness).

To analyze which municipalities respondents select, I estimate a Bayesian hierarchical logistic regression with random effects, or a Generalized Linear Mixed Model (GLMM). Each of the 15 municipalities shown to respondent r is modeled as an independent Bernoulli trial:

$$\Pr(Y_{ri} = 1) = \text{logit}^{-1}(\alpha_r + \mathbf{X}_{ri}\boldsymbol{\beta}) \tag{1}$$

where $Y_{ri} = 1$ if respondent r selects municipality i and $\mathbf{X}_{ri}$ is a vector of municipality-level covariates (log distance from home municipality, log population ratio of comparison relative to home municipality, same-state indicator, same-coalition indicator, an indicator for whether the candidate municipality's governing coalition matches the respondent's pre-treatment vote intention coalition, and the candidate municipality's governing coalition). Following McCartan, Brown, and Imai (2024), $\alpha_r \sim \mathcal{N}(0, \sigma_\alpha^2)$ is a respondent-level random intercept capturing heterogeneity in the total number of municipalities selected. A pool fixed effect (nearest / largest / random) is included to partial out mechanical differences in selection rates across pools.

Prior distributions for the coefficients are formed indirectly by taking a QR decomposition of the centered covariate matrix (Goodall 1993). Coefficients on the QR-decomposed centered covariates are assigned weakly informative $t_2(0, 2.5)$ priors, which implies priors for the actual coefficients that are adapted to the scale and correlation structure of the covariates. Because the 15-municipality menu is researcher-constructed, estimated coefficients are conditional on the offered menu and should be interpreted accordingly.

Results from the 3,028 respondents from Wave 1 are shown in Table 3 and Figure 2. The figure plots posterior mean percentage point changes in selection probability relative to a 50-50 baseline. Several patterns stand out. First, spatial variables tend to matter, as respondents are less likely to select municipalities that are farther away from their municipality and more likely to select those from the same state. Second, respondents are more

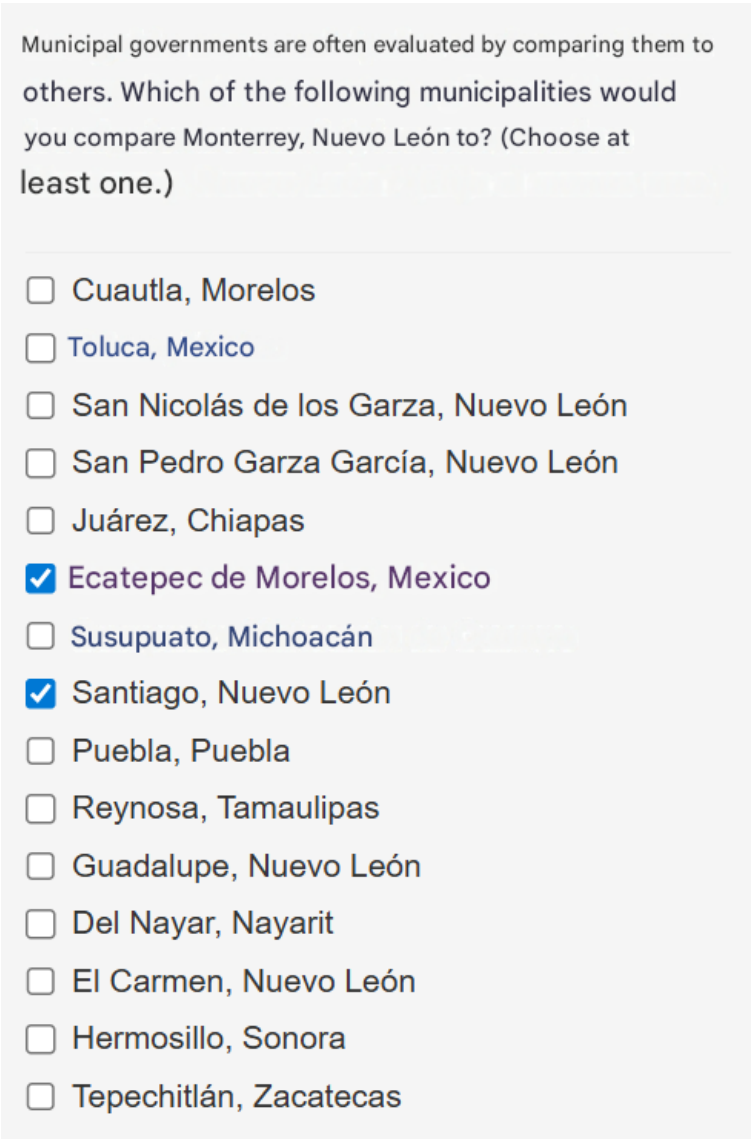

Figure 1: An example of the comparison municipality selection question shown to respondents.

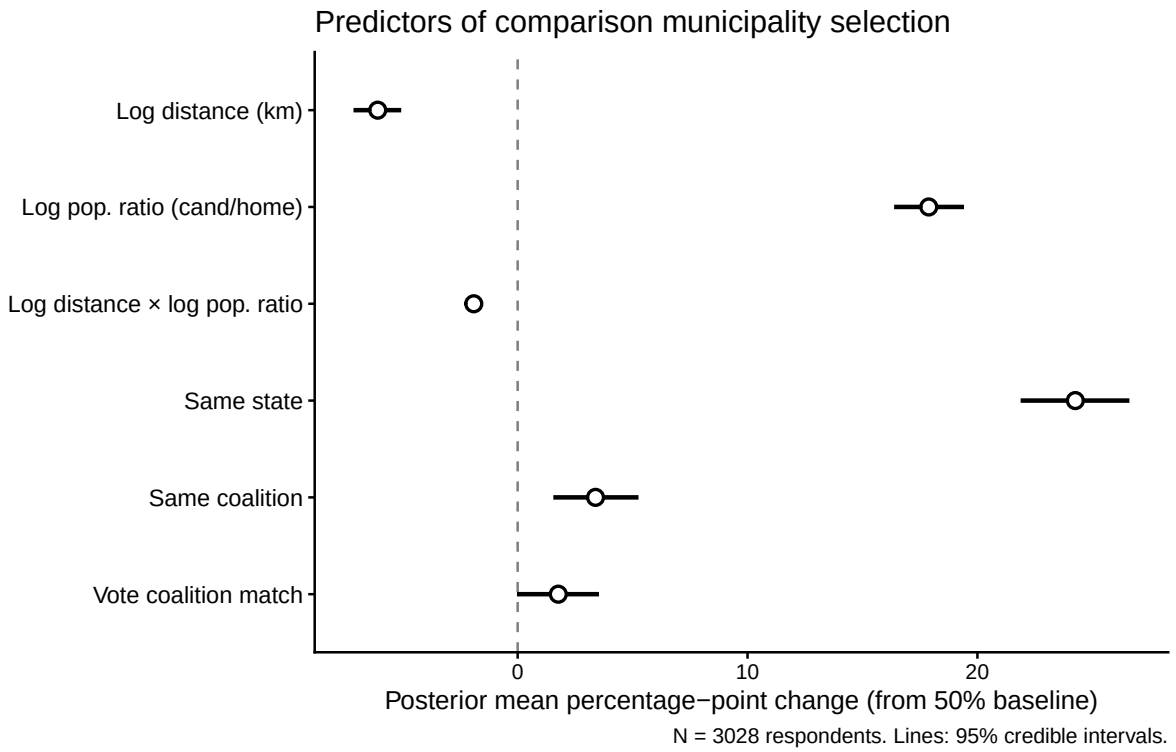

Figure 2: Analysis of the comparison municipality selections. Points are posterior means. Lines are 95% credible intervals. The reference category for the coalition variables is MC. Pool variables and home municipality controls are not shown.

likely to select municipalities that are larger than their home municipality. This effect is attenuated by distance. Third, respondents are more likely to select municipalities governed by the same coalition as their home municipality, suggesting that municipalities most relevant to respondents tend to share partisan government characteristics. Fourth, respondents are slightly more likely to select municipalities governed by the coalition that they report intending to vote for.

The preference for same-coalition municipalities is at least partially due to governing-party knowledge. I test this by examining respondents' knowledge of the governing coalition of comparison municipalities in Wave 2, where some respondents were by chance assigned a comparison municipality that they chose in the first wave. Using linear probability models with respondent fixed effects, I find that respondents were 8 percentage points more accurate at identifying the governing party of municipalities they selected as benchmarks in

Wave 1. These results suggest that knowledge of party governance did play a role in whether municipalities were selected by respondents (see Appendix 13 for full results).

Table 3: Bayesian multilevel logistic regression: predictors of benchmark municipality selection. Estimates are posterior means; Est. SD is the posterior standard deviation. Reference categories: MC (cand. and home coalition), random (pool). $N = 45{,}420$ observations, 3,028 respondents.

Parameter	**Estimate**	**Est. SD**	**l-95% CI**	**u-95% CI**	$\hat{R}$
Regression coefficients					
Intercept	−7.58	0.31	−8.20	−6.98	1.00
Log distance (km)	−0.24	0.02	−0.29	−0.20	1.00
Log pop. ratio (cand./home)	0.75	0.04	0.68	0.82	1.00
Same state	1.06	0.06	0.94	1.19	1.00
Same coalition (governing)	0.14	0.04	0.06	0.21	1.00
Vote–coalition match	0.07	0.04	−0.00	0.14	1.00
Cand. coalition: MORENA/PVEM/PT	0.02	0.06	−0.10	0.14	1.00
Cand. coalition: PAN/PRI/PRD	0.29	0.06	0.17	0.41	1.00
Cand. coalition: Other	−0.06	0.14	−0.35	0.21	1.00
Home coalition: MORENA/PVEM/PT	0.23	0.07	0.09	0.36	1.00
Home coalition: PAN/PRI/PRD	0.36	0.07	0.22	0.49	1.00
Log home pop.	0.39	0.02	0.35	0.43	1.00
Pool: nearest	1.38	0.07	1.23	1.52	1.00
Pool: largest	0.54	0.08	0.39	0.69	1.00
Log distance × log pop. ratio	−0.08	0.01	−0.09	−0.06	1.00
Random effects					
σ_α (respondent)	0.41	0.03	0.35	0.47	1.00

Experimental Design

In the second wave of the survey, I test four treatment conditions that vary how crime information for a respondent's municipality is benchmarked. The experimental setup allows me to both manipulate the information they receive and determine whether it is "news" to them by comparing it to their pre-treatment beliefs.

Designing a survey experiment that matches real electoral conditions as closely as possible is important for the external validity of the results (Incerti 2020; Breitenstein 2019). Thus, a balance must be struck between controlling for confounders and maintaining external

validity. This paper does so by using real performance information and eliciting voters' vote intentions and beliefs about political parties both pre- and post-treatment, which allows for testing of belief updating in a way that is more directly comparable to real-world conditions.

I present information on crime (specifically, robbery counts[3]) in the treatments. There are several reasons for this choice. First, crime is consistently among the top concerns of Mexican voters,[4] making it a high-salience issue. This is important, because a necessary condition for information's effect on voting is that it be relevant to voters (Bhandari, Larreguy, and Marshall 2023; Adida et al. 2020).

Second, municipal governments have genuine, if partial, responsibility for public security: municipal presidents control the funding, hiring, and structure of municipal police forces, which comprise the majority of law enforcement personnel in Mexico (Sabet 2012; Marshall 2024). Indeed, policy interventions have been shown to lower rates of robbery (Cabrera-Hernández and Zárate-Tenorio 2026). Even in cases where states have authority over municipalities (*mando único*), municipal presidents still generally control finances for public security (Víctor Cubillas 2025).[5] This makes crime a domain where voters can reasonably (partially) attribute crime outcomes to the incumbent.

Third, and most importantly for testing the theory, crime statistics are inherently noisy signals, a setting in which comparisons that filter noise are most useful. The data source is the Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP), which reports monthly robbery counts at the municipal level. I use 2025 totals to maximize proximity to the 2027 municipal elections.

Beliefs about crime are elicited before and after treatment. In Wave 1, respondents are asked how many robberies per 100,000 people they believe were reported in their home

3. I choose robberies instead of violent crime (like homicides) because the latter may be more associated with organized crime which is much harder to combat at the municipal level.

4. In a recent public opinion poll in El Economista (2025), crime was the top issue at 45.9%, while the next most popular issue (economy) was 9.0%.

5. Only 3.5% of municipalities in 2022 spent 0 pesos on local public security according to the Censo Nacional de Gobiernos Municipales y Demarcaciones Territoriales de la Ciudad de México (CNGMD), although about 15% had missing data.

municipality in 2025, answering in an open numeric field. Because a rate per 100,000 is not an especially familiar measure, the question is preceded by a reference point: in 2025, half of all Mexican municipalities recorded fewer than 79 robberies per 100,000 people and half recorded more. Without this anchor, responses might vary for reasons unrelated to beliefs about crime. The anchor is held constant across arms, so it cannot generate differences between conditions.

The second is a belief about relative standing. Immediately before treatment in Wave 2, respondents see the four comparison municipalities and are asked, for each, whether they believe it has more robberies than their own municipality, about the same, or fewer. For comparison treatment arms, these are the municipalities they will later be shown in the bar chart. For treatment arms without a comparison, the respondent is randomly assigned to municipalities from one of the three comparison groups for their municipality. Ranking the home municipality against the same four municipalities that appear in the treatment ensures the pre-treatment belief and the information shown are about the same comparison set. Both questions are repeated after treatment in identical form, which yields the post-treatment beliefs used in the belief-updating analysis.

The first treatment arm (T1) provides plain information: a robbery rate per 100,000 people for the respondent's own municipality, without any point of comparison. The second (T2) adds a bar chart comparing the respondent's municipality to a sample of similar municipalities, providing a non-partisan comparison. The third (T3) replaces that sample with municipalities governed by a political coalition (MORENA/PT/PVEM vs. PAN/PRI/PRD vs. MC) different from the coalition governing the respondent's municipality. The fourth (T4) uses municipalities governed by the same coalition as the respondent's municipality. The municipalities in the comparison arms (T2, T3, and T4) are selected using Mahalanobis distance matching on log population, log geographic area, and log distance from the respondent's home municipality.[6] This matching strategy is implemented in order to show

6. Mahalanobis distance measures how far apart two units are across several characteristics at once, rescaling each characteristic by its variability and adjusting for the correlations among them, so that no

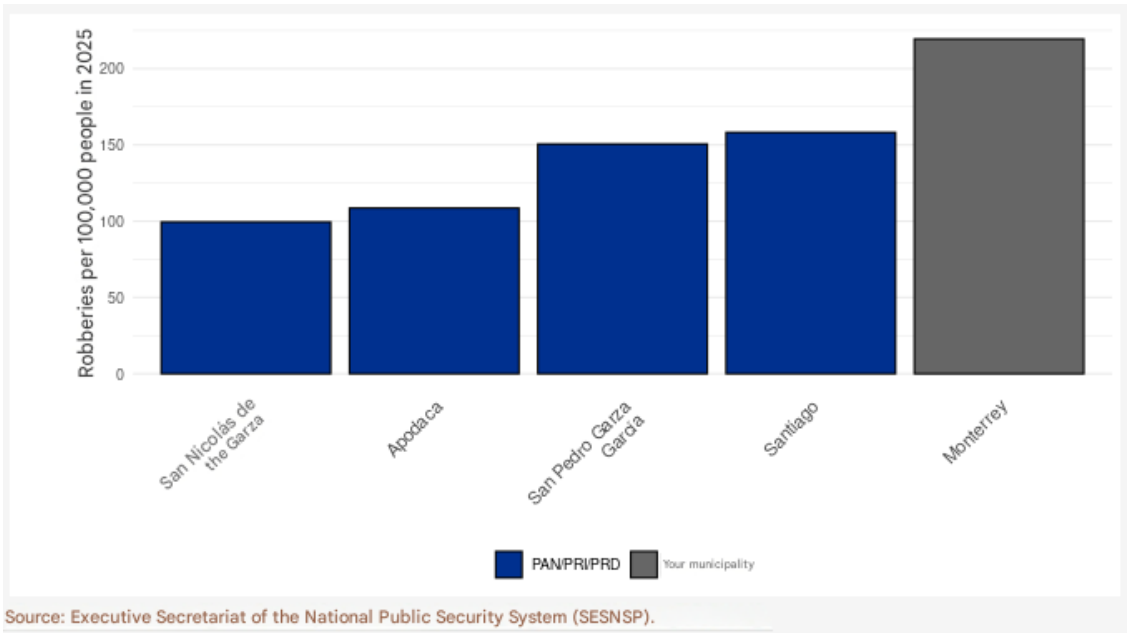

Figure 3: Example of a partisan comparison treatment graph, translated to English.

municipalities that are most relevant for comparison, and is consistent with the results from the descriptive analysis in Wave 1.

All treatment arms (T1–T4) present the following information: "Municipal presidents control police funding and, therefore, have partial influence over local crime. In [*respondent's municipality*], there were X robberies per 100,000 inhabitants in 2025." Comparison arms (T2–T4) add: "The following graph shows the robbery rate per 100,000 people in 2025 for [*respondent's municipality*] and a sample of similar municipalities. . . " In T2 (non-partisan comparison), no qualifier is added to the last sentence. In T3 (opposition-coalition comparison), ". . . that are governed by [*opposite coalition*]" is added. In T4 (same-coalition comparison), ". . . that are governed by [*same coalition*]" is added. An example of the bar graph from a partisan comparison treatment group is shown in Figure 3. A bar graph from T2 would look identical but would not label which party the comparison municipalities are from.

single characteristic dominates and correlated characteristics are not double counted. Matching on it selects, for each respondent, the municipalities that are closest to their home municipality on population, area, and distance taken together.

The main control group (Control) provides a weather placebo: factual text about rainfall variation across Mexican municipalities, followed by the respondent's home municipality's average annual precipitation with no crime content or comparison. Specifically, the text of the control group reads: "Geography, altitude, and regional climatic systems are the primary determinants of precipitation patterns—factors that, to a large extent, we cannot control. In [respondent's municipality], there were X mm of precipitation in 2025."

Control and comparison treatment arms were assigned with equal probability (0.2), while the plain information treatment arm was assigned a lower probability (0.1). An alternative control arm (Control2, assigned with probability 0.1) is used for robustness checks reported in the appendix. The plain information treatment and alternative control arms were assigned lower probability because the main focus of the paper is on how differing benchmarks change interpretation of information. Appendix 17 diagrams the full Wave 2 flow, from pre-treatment elicitation through assignment to post-treatment elicitation.

The arms are structured to address different aspects of how information is benchmarked. Specifically, the experiment in general tests whether information moves beliefs at all. The design also addresses whether differing benchmark designs have heterogeneous effects on voting behavior. These broad questions stem from the 8 pre-registered research questions in the pre-analysis plan.

Experimental Estimation Strategy

In estimating the effect of the informational treatments, heterogeneous treatment effects are of interest. Specifically, the estimation needs to account for whether the information is "good news" or "bad news" compared to baseline beliefs. To capture this, I define two complementary perception gap measures (Haaland, Roth, and Wohlfart 2023). Throughout this section i indexes respondents. The subscripts h and b from the theory correspond to respondent i's home municipality and the benchmark municipality shown, respectively.

The crime-level perception gap (CG_i) is defined as

$$CG_i = C_i^{\text{actual}} - C_i^{\text{prior}} \tag{2}$$

where C_i^{actual} is the respondent's home municipality's actual robbery rate per 100,000 inhabitants and C_i^{prior} is the respondent's pre-treatment estimate of robberies per 100,000 inhabitants in their municipality, elicited on the same scale. When eliciting the prior belief about the crime rate, respondents were informed of what the median municipality's crime rate was. When $CG_i > 0$, actual crime is higher than expected (bad news about the level of crime). In estimation, CG_i enters as a signed-log (log-modulus) transformation of this gap, $\text{sign}(CG_i) \times \ln(1 + |CG_i|)$. Respondents' priors have a long right tail, so untransformed gaps would let a small number of extreme values dominate the interaction terms.

The crime-ranking perception gap (RG_i) is defined as

$$RG_i = R_i^{\text{actual}} - R_i^{\text{prior}} \tag{3}$$

where R_i^{actual} is the respondent's home municipality's robbery rank among comparison municipalities (higher rank = more robberies = worse relative performance) and R_i^{prior} is the respondent's pre-treatment perceived rank, constructed as one plus the number of comparison municipalities they judge to have fewer robberies. Both measures run from one to five, with five indicating that the home municipality has the most robberies of the five. When $RG_i > 0$, the municipality's actual comparative standing is worse than the respondent expected.

The estimating equation is

$$y_i = \alpha_0 + \alpha_1 CG_i + \alpha_2 RG_i + \sum_{k=1}^{4} \delta_k T_{ik} + \sum_{k=1}^{4} \beta_k T_{ik} \times CG_i + \sum_{k=1}^{4} \gamma_k T_{ik} \times RG_i + \mathbf{X}_i' \boldsymbol{\lambda} + \varepsilon_i \tag{4}$$

where y_i is the post-treatment value of the outcome of interest for respondent i. $T_{ik} = 1$

if respondent i is assigned to arm k; arms are indexed $k \in \{1, 2, 3, 4\}$ corresponding to T1 (plain information), T2 (non-partisan benchmark), T3 (opposition-coalition benchmark), and T4 (same-coalition benchmark) respectively, with the non-comparison Control as the omitted baseline. $\mathbf{X}_i$ is a control for the prior, or the outcome measured pre-treatment. The key coefficients of interest are β_k, the treatment effect conditional on the crime level perception gap (good news or bad news about the level of crime), and γ_k, the treatment effect conditional on the crime ranking perception gap.

The two gap measures are not competing versions of the same quantity. Rather, they stand in for different objects in the model. CG_i captures surprise about the respondent's municipality's crime rate r_h. What varies across treatment arms is how much of that raw news the respondent can assign to the incumbent: in the same-coalition benchmark, both the common shock and the policy term difference out, so a given surprise carries more information about z_h than the same surprise does in the control, plain information, or non-partisan benchmark arms. The competence content of a treatment is thus carried by the interaction β_k rather than by CG_i on its own, which is why the theory makes predictions about β_k and not about α_1.

RG_i captures surprise about relative standing. Because the pre-treatment ranking question is asked about the same comparison municipalities the respondent is later shown, RG_i nets out what the respondent already expected the comparison to reveal. This measure is of particular interest in the opposition-coalition benchmark treatment arm, because according to the theory the opposition-coalition benchmark relies on the respondent holding a confident prior about the policy gap between the coalitions.

Outcomes estimated in the model are vote intentions and beliefs about crime handling of incumbent governments and the three major coalitions. Specifically, I use intention to vote for the incumbent post-treatment as the main outcome and belief updating for explanation of the mechanism. I verify that respondents are similar across treatment arms using equivalence tests (Hartman and Hidalgo 2018), which take substantial difference rather than equality as

the null hypothesis and therefore provide positive evidence of balance rather than a mere failure to reject it. The arms are equivalent on every pre-treatment covariate except region, which fails only marginally and because of limited power in small regional cells rather than any observed imbalance (Appendix 4).

Experimental Results

Before turning to the outcomes, I verify that the treatments changed the beliefs they were designed to change. Because both pre-treatment beliefs about crime are elicited again after treatment, I can compare each respondent's post-treatment beliefs to the truth and examine how post-treatment accuracy differs in treatment groups compared to the control group. Every treated arm improves the accuracy of respondents' beliefs about their own municipality's robbery rate, which is expected given that all four reveal the robbery rate. The comparison treatment arms significantly increase accuracy about the respondent's crime relative to comparison municipalities, while plain information leaves those beliefs essentially unchanged. See Appendix 5 for a more in-depth discussion of this manipulation check.

Vote Intention Results

I first consider the effect of information on vote intention. Vote intention is a binary variable indicating whether a respondent reported intending to vote for the current incumbent coalition (1) or not (0). In models estimating this as the outcome, I control for respondents' pre-treatment vote intention. Figure 4 presents the experimental result with the vote intention outcome. The key finding is that only same-coalition benchmarks (T4) significantly change vote intention: respondents assigned to T4 who learn their municipality has higher crime than they expected are less likely to indicate they will vote for the incumbent party in the next local elections. This effect operates only through the crime-level perception gap (left panel), where T4 shows a statistically significant negative coefficient. All other treatment arms have null effects for both perception gaps.

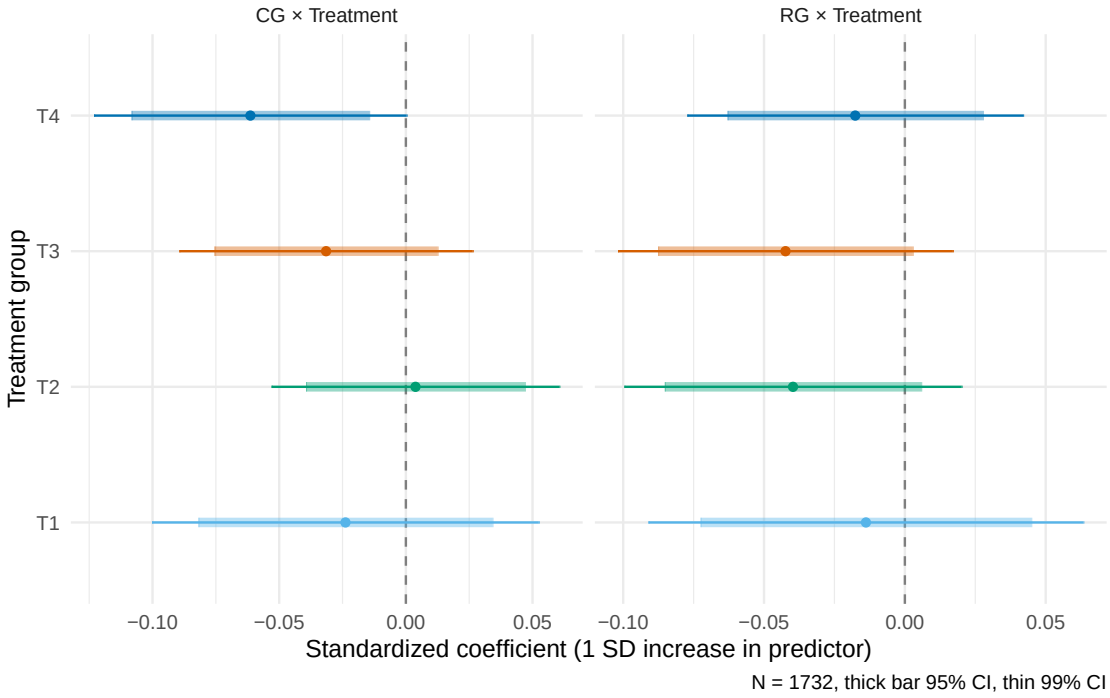

Figure 4: Coefficient estimates for the effect of each treatment on the change in vote intention for the respondent's home coalition. The left panel (CG × Treatment) plots the interaction of each treatment with the crime-level perception gap (β_k); the right panel (RG × Treatment) plots the interaction with the ranking perception gap (γ_k). Coefficients are standardized to a 1 SD increase in the respective gap measure. Points are OLS estimates; thick bars are 95% and thin bars 99% confidence intervals. The reference category for treatment is Control (weather placebo with no comparison).

Substantively, because the perception gap enters the model as a signed log, the size of the effect depends on the scale of the misperception rather than its raw magnitude. A one-standard-deviation increase in the logged perception gap is associated with a decrease of approximately 5.3 percentage points in the probability of voting for the incumbent. The median respondent has a raw perception gap of 445, which on the logged scale sits roughly one standard deviation from accurate, so effects of this size describe a typical voter rather than an extreme one. Expressed differently, each doubling of the perception gap is associated with a 0.6 percentage point decrease in the probability of voting for the incumbent.

The results are robust to alternative measures of RG (Appendix 6) and alternative methods of handling extreme values of CG (Appendix 7). Triple interactions of the treatment effects with crime importance reveal no significant moderation (Appendix 10). The same-coalition effect grows in magnitude both among respondents who placed crime among their three most important issues (Appendix 11) and when responses are reweighted to census margins (Appendix 8), although in each case the original estimate remains within the new confidence interval.

These results point to the case presented in the theory in which voters are more interested in learning about competence than policy differences, with one caveat: the theory predicts that the opposition-coalition treatment arm does reveal information about competence depending on voters' priors about how the two parties differ, whereas here the interaction that accounts for respondents' priors about relative performance ($RG \times T3$) shows no statistically significant effect. Additional evidence suggests this null result is due to noise in the full sample. The treatment does not tell respondents about which party governs their own municipality; this was done in order to not conflate information about own party governance with differences in benchmarks. When I subset the sample to only people who know which party governed their own municipality (roughly 80% of the sample), the RG interaction becomes statistically significant for the T3 treatment arm while the T4 general crime gap effect increases slightly in magnitude and maintains its statistical significance (see Appendix 9).

Thus, restricting the sample to respondents for whom the partisan treatment is most effective strengthens the evidence in favor of an interpretation that voters are interested in voting out incompetent politicians.

The estimate from the same-coalition benchmark is not an artifact of respondents being especially familiar with the comparison municipalities they were shown. If the effect were driven by familiarity, the non-partisan benchmark rather than the same-coalition benchmark should show the larger effect, because respondents who were shown the non-partisan benchmark saw comparisons they found more familiar on both available measures. The municipalities shown in the non-partisan arm are closer to the benchmarks respondents chose for themselves in the descriptive analysis than those shown in the same-coalition arm. Respondents shown the non-partisan benchmark were also more likely to be shown a municipality they had explicitly selected in Wave 1 (51% in T2 vs. 40% in T4). Appendix 12 provides more evidence that the effect is not driven by benchmark municipality familiarity alone.

Does belief updating explain vote changes?

The voting results imply that respondents are evaluating incumbents based on competence. If so, the treatment that moves vote intention should also move how respondents evaluate the incumbent's handling of crime. This section examines whether any of the treatments moves beliefs.

I estimate the specification used for vote intention, replacing the outcome with post-treatment evaluations of crime handling measured on 0–100 sliders. I look at two main outcomes: the incumbent government's rating, and the incumbent government's rating minus the average rating of the coalitions that do not govern the respondent's municipality. I also report a model using the average rating of the other coalitions. All three are estimated on one sample with an identical right-hand side, so the composite coefficient is exactly the difference of the two component coefficients.

Table 4 reports the crime-level perception gap interactions for all three outcomes and all

four arms. Reading down the third column, T4's coefficient is the largest in magnitude and the only one distinguishable from zero: a 1 SD increase in the gap lowers the composite by 3.2 points on its −100 to 100 scale. The components move in the expected directions: the incumbent rating falls 2.3 points and the other-coalition rating rises 0.9. However, neither is individually distinguishable from zero. I therefore do not claim separately that respondents downgrade the incumbent or that they upgrade the alternatives. What the estimates support is a shift in the two ratings taken jointly.

Table 4: Crime-level perception gap (CG) × arm interactions for three belief outcomes measured on 0–100 sliders. The third column is the first minus the second. Coefficients are standardized to a 1 SD increase in CG; p-values from HC2 robust standard errors in brackets. The reference category for treatment is Control.

Term	**Incumbent**	**Other coalitions**	**Incumbent − other**
CG × T1	-1.88	-0.28	-1.60
	[0.35]	[0.87]	[0.42]
CG × T2	-0.36	0.20	-0.56
	[0.83]	[0.87]	[0.70]
CG × T3	0.23	0.03	0.21
	[0.88]	[0.98]	[0.89]
CG × T4	-2.28	0.93	-3.21
	[0.19]	[0.49]	[0.03]
N	1,717	1,717	1,717
R^2	0.43	0.41	0.33

A Wald test on the composite outcome finds the T4 coefficient significantly different from T3 (a difference of 3.4 points, $p = 0.03$) and marginally different from T2 ($p = 0.07$); the difference from T1 is smaller and not distinguishable from zero ($p = 0.42$). The pattern adds to the evidence that the same-coalition benchmark moves the incumbent's rating relative to the other coalitions more than the other benchmarks do, and is consistent with the voting results.

The reason the composite outcome is statistically significant and the direct evaluation of the incumbent is not may be due to the survey instrument, which asked for ratings of the incumbent government and all of the three main coalitions on the same page, under one shared

pair of endpoint labels and with every slider initialized at the midpoint (see Appendix 18). Thus, it is possible that responses to the treatment impacted relative evaluations of the incumbent by (largely) pushing evaluations of the incumbent down and (slightly) pushing evaluations of opposition coalitions up. The difference in magnitude between the two model coefficients (-2.3 for incumbent evaluations vs. 0.9 for opposition coalitions) is evidence for this mechanism. However, I cannot rule out that the difference in significance may be just a statistical property of the composite's smaller variance than the incumbent rating alone. Therefore, the results from the belief models should be read as consistent with a mechanism that belief updating drives changes in vote intention.

The ranking perception gap does not produce a comparable result. Estimated against Control, which shows no comparison at all, the ranking-gap interactions are statistically significant in T2 and T4. However, none of the treatment coefficients are distinguishable from each other, and a placebo comparison treatment also shows a statistically significant result (Appendix 15). I therefore treat the ranking-gap belief results as null and base the analysis on the crime-level gap.

Whether the belief carries the vote is a further question, and one these data cannot answer. A pre-registered exploratory mediation analysis, reported in full in Appendix 16, splits the T4 effect on incumbent vote intention into the portion transmitted through the incumbent-versus-other-coalitions rating and the portion reaching the vote by other routes. The mediated portion has the expected sign, but its confidence interval covers zero across nearly the whole range. Additionally, the direct effect, meaning the portion of the vote intention outcome unexplained by the belief measure, is large. This could be because the crime-handling rating is a narrow proxy for what the treatment may have moved. If respondents read a same-coalition comparison as evidence about the incumbent's general competence, as the theory supposes, much of that update runs through beliefs this survey never measured and appears here as a direct effect.

Discussion

In this paper, I show that the way information is benchmarked can have a meaningful effect on how voters update their evaluations of politicians. In a two-wave survey fielded in Mexico, I find that crime information benchmarked by municipalities governed by the same political coalition changes respondents' evaluations of incumbent politicians. The theory presented suggests this is because same-party benchmarks provide a credible point of comparison that allows voters to attribute outcomes to incumbent performance rather than partisan differences. This has implications for both the design of informational interventions and our understanding of voter behavior in contexts where information is difficult for voters to interpret.

When designing informational campaigns, policymakers and researchers should consider not just the content of the information but also how it is benchmarked. Information changes votes only when it tells voters something they did not already know and arrives in a form they can use to assign blame. Benchmarks should be designed to address both of these conditions by providing comparisons that isolate what is meaningful to voters. How to determine which benchmarks are most appropriate may differ from context to context, so researchers should not assume that the results in this paper apply to different political contexts. Prior to fielding informational treatments, researchers should determine which factors are relevant to voters in order to maximize the credibility and salience of benchmarked information.

More broadly, these findings contribute to the literature on political information and voter behavior by highlighting the importance of credible comparisons for making information interpretable and indicative of government performance. In contexts where voters have limited information or face hard-to-interpret signals, providing clear and relevant benchmarks can help them make more informed choices. Whether media coverage, political campaigns, or civic education programs present information in a way that allows voters to make meaningful comparisons may be just as important as the information itself.

References

Adida, Claire, Jessica Gottlieb, Eric Kramon, and Gwyneth McClendon. 2020. "When Does Information Influence Voters? The Joint Importance of Salience and Coordination" [in en]. *Comparative Political Studies* 53, no. 6 (May): 851–891. https://doi.org/10.1177/0010414019879945.

Estimaciones NSE 2022. 2022.

Arel-Bundock, Vincent, André Blais, and Ruth Dassonneville. 2021. "Do Voters Benchmark Economic Performance?" [In en]. *British Journal of Political Science* 51, no. 1 (January): 437–449. https://doi.org/10.1017/S0007123418000236.

Arias, Eric, Horacio Larreguy, John Marshall, and Pablo Querubin. 2024. "Does the Content and Mode of Delivery of Information Matter for Electoral Accountability? Evidence from a Field Experiment in Mexico" [in en]. *Latin American Economic Review* 34:1–41. https://doi.org/10.60758/laer.v34i.335.

Aytaç, Selim Erdem. 2018. "Relative Economic Performance and the Incumbent Vote: A Reference Point Theory." *The Journal of Politics* 80, no. 1 (January): 16–29. https://doi.org/10.1086/693908.

Banerjee, Abhijit, Nils Enevoldsen, Rohini Pande, and Michael Walton. 2024. "Public Information Is an Incentive for Politicians: Experimental Evidence from Delhi Elections" [in en]. *American Economic Journal: Applied Economics* 16, no. 3 (July): 323–353. https://doi.org/10.1257/app.20220088.

Besley, Timothy, and Anne Case. 1992. *Incumbent Behavior: Vote Seeking, Tax Setting and Yardstick Competition.* Working Paper, March. https://doi.org/10.3386/w4041.

———. 1995. "Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition." *The American Economic Review* 85 (1): 25–45.

Bhandari, Abhit, Horacio Larreguy, and John Marshall. 2023. "Able and Mostly Willing: An Empirical Anatomy of Information's Effect on Voter-Driven Accountability in Senegal" [in en]. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12591, *American Journal of Political Science* 67 (4): 1040–1066. https://doi.org/10.1111/ajps.12591.

Boas, Taylor C., and F. Daniel Hidalgo. 2011. "Controlling the Airwaves: Incumbency Advantage and Community Radio in Brazil" [in en]. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2011.00532.x, *American Journal of Political Science* 55 (4): 869–885. https://doi.org/10.1111/j.1540-5907.2011.00532.x.

Breitenstein, Sofia. 2019. "Choosing the crook: A conjoint experiment on voting for corrupt politicians" [in EN]. *Research & Politics* 6, no. 1 (January): 2053168019832230. https://doi.org/10.1177/2053168019832230.

Cabrera-Hernández, Francisco, and Bárbara A. Zárate-Tenorio. 2026. "From classrooms to crime prevention: The incapacitation effect of extending the school day in Mexico." *World Development* 203 (July): 107368. https://doi.org/10.1016/j.worlddev.2026.107368.

Chong, Alberto, Ana L. De La O, Dean Karlan, and Leonard Wantchekon. 2015. "Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification." *The Journal of Politics* 77, no. 1 (January): 55–71. https://doi.org/10.1086/678766.

Cohen, Jeffrey E. 2020. "Relative Unemployment, Political Information, and the Job Approval Ratings of State Governors and Legislatures" [in EN]. *State Politics & Policy Quarterly* 20, no. 4 (December): 437–461. https://doi.org/10.1177/1532440020905800.

Cruz, Cesi, Philip Keefer, and Julien Labonne. 2021. "Buying Informed Voters: New Effects of Information on Voters and Candidates." *The Economic Journal* 131, no. 635 (April): 1105–1134. https://doi.org/10.1093/ej/ueaa112.

Cruz, Cesi, Philip Keefer, Julien Labonne, and Francesco Trebbi. 2024. "Making Policies Matter: Voter Responses to Campaign Promises." *The Economic Journal* 134, no. 661 (July): 1875–1913. https://doi.org/10.1093/ej/ueae008.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I* [in en]. Google-Books-ID: SxSgDwAAQBAJ. Cambridge University Press, July.

El Economista. 2025. *Claudia Sheinbaum Aprobacion Ciudadana Agosto 2025* [in en]. Technical report. August.

Epp, Derek A., and Christopher Wlezien. 2026. "Benchmarking in public updating of economic perceptions" [in en]. *Political Science Research and Methods* (July): 1–19. https://doi.org/10.1017/psrm.2026.10113.

Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes*." *The Quarterly Journal of Economics* 123, no. 2 (May): 703–745. https://doi.org/10.1162/qjec.2008.123.2.703.

Goodall, Colin R. 1993. "13 Computation using the QR decomposition" [in en-US]. In *Handbook of Statistics 9: Computational Statistics,* vol. 9. Elsevier, January. https://doi.org/10.1016/S0169-7161(05)80137-3.

Gottlieb, Jessica. 2016. "Greater Expectations: A Field Experiment to Improve Accountability in Mali" [in en]. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12186, *American Journal of Political Science* 60 (1): 143–157. https://doi.org/10.1111/ajps.12186.

Grossman, Guy, and Tara Slough. 2022. "Government Responsiveness in Developing Countries" [in en]. *Annual Review of Political Science* 25, no. Volume 25, 2022 (May): 131–153. https://doi.org/10.1146/annurev-polisci-051120-112501.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments" [in en]. *Journal of Economic Literature* 61, no. 1 (March): 3–40. https://doi.org/10.1257/jel.20211658.

Hansen, Kasper M., Asmus L. Olsen, and Mickael Bech. 2015. "Cross-National Yardstick Comparisons: A Choice Experiment on a Forgotten Voter Heuristic" [in en]. *Political Behavior* 37, no. 4 (December): 767–789. https://doi.org/10.1007/s11109-014-9288-y.

Hartman, Erin, and F. Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests" [in en]. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12387, *American Journal of Political Science* 62 (4): 1000–1013. https://doi.org/10.1111/ajps.12387.

Huang, Haifeng. 2015. "International Knowledge and Domestic Evaluations in a Changing Society: The Case of China" [in en]. *American Political Science Review* 109, no. 3 (August): 613–634. https://doi.org/10.1017/S000305541500026X.

Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design" [in en]. *American Political Science Review* 114, no. 3 (August): 761–774. https://doi.org/10.1017/S000305542000012X.

Kayser, Mark Andreas, and Michael Peress. 2012. "Benchmarking across Borders: Electoral Accountability and the Necessity of Comparison" [in en]. *American Political Science Review* 106, no. 3 (August): 661–684. https://doi.org/10.1017/S0003055412000275.

Keefer, Philip, and Stuti Khemani. 2005. "Democracy, Public Expenditures, and the Poor: Understanding Political Incentives for Providing Public Services." *The World Bank Research Observer* 20, no. 1 (March): 1–27. https://doi.org/10.1093/wbro/lki002.

Ley, Sandra. 2017. "Electoral Accountability in the Midst of Criminal Violence: Evidence from Mexico" [in en]. *Latin American Politics and Society* 59 (1): 3–27. https://doi.org/10.1111/laps.12008.

Marshall, John. 2024. *Tuning In, Voting Out: News Consumption Cycles, Homicides, and Electoral Accountability in Mexico* [in en]. SSRN Scholarly Paper. Rochester, NY, February. https://doi.org/10.2139/ssrn.5127675.

McCartan, Cory, Jacob R. Brown, and Kosuke Imai. 2024. "Measuring and Modeling Neighborhoods" [in en]. *American Political Science Review* 118, no. 4 (November): 1966–1985. https://doi.org/10.1017/S0003055423001429.

Sabet, Daniel M. 2012. *Police Reform in Mexico: Informal Politics and the Challenge of Institutional Change* [in en]. May.

Singer, Matthew M. 2011. "Who Says "It's the Economy"? Cross-National and Cross-Individual Variation in the Salience of Economic Performance" [in EN]. *Comparative Political Studies* 44, no. 3 (March): 284–312. https://doi.org/10.1177/0010414010384371.

Víctor Cubillas. 2025. "¿Qué es el Mando Único y cuáles son sus facultades?" [In en]. *Tribuna de San Luis* (April).

Appendix for Which Comparison Counts? Information, Benchmarking, and Electoral Accountability

Adam D Roberts

1 Coalition Networks

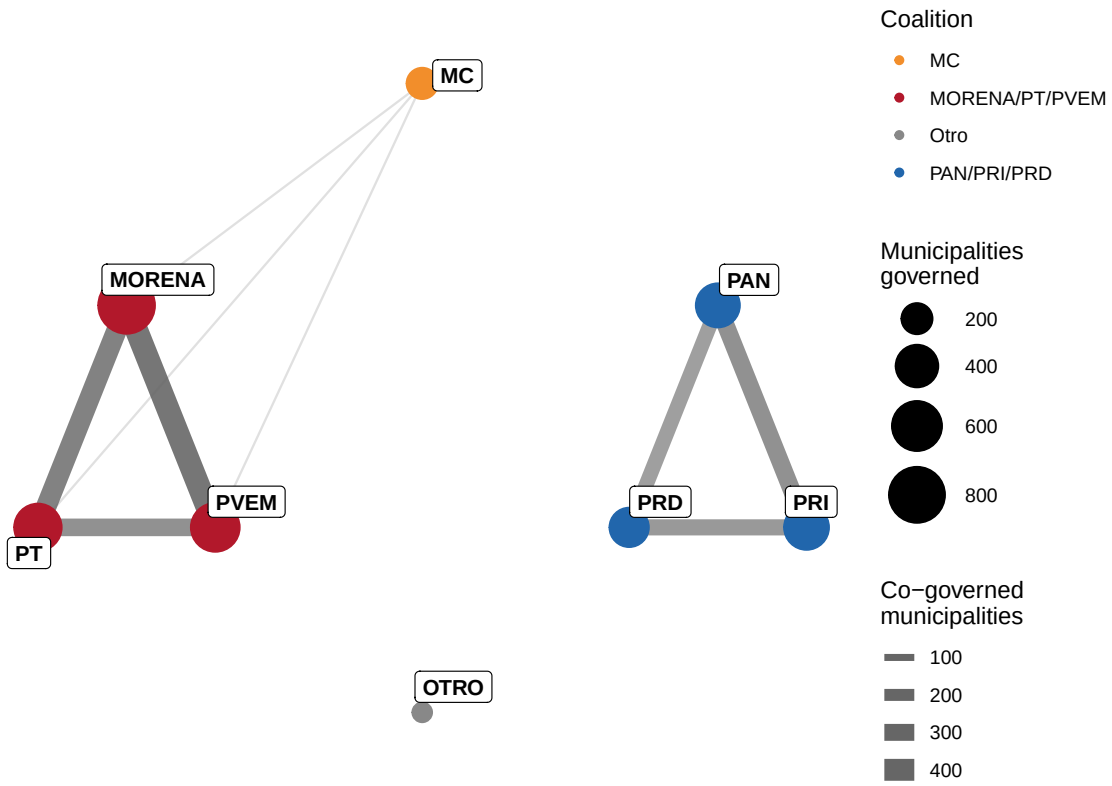

Figure 1: Party co-governance network among Mexican municipalities. Nodes represent parties; edge weight is the number of municipalities where two parties jointly govern, and node size is proportional to the total number of municipalities governed.

2 Crime and Party Coalitions Governing Municipalities

Table 1 reports OLS estimates of the association between governing party coalition and municipal robbery rates (per 100,000 residents). Across all three specifications, MC-governed municipalities are statistically indistinguishable from the PAN/PRI/PRD reference category. In the bivariate model (1), MORENA/PVEM/PT-governed municipalities are associated with roughly 31 additional robberies per 100,000 residents; however, this association disappears once population and geographic area are controlled for in models (2) and (3), suggesting it is largely driven by compositional differences rather than party governance per se.

Dependent variable:	Robberies per 100,000 residents		
Model:	(1)	(2)	(3)
Governing coalition (ref. PAN/PRI/PRD)			
MC	4.789	0.8073	13.14
	(17.24)	(14.76)	(13.54)
MORENA/PVEM/PT	31.07***	9.192	9.722
	(10.92)	(9.384)	(8.694)
Municipality controls			
Log population		76.89***	66.14***
		(3.184)	(3.275)
Log area (km^2)		-33.10***	-40.45***
		(3.111)	(3.918)
Constant	146.0***	-412.2***	
	(8.823)	(32.44)	
Fixed effects			
State			Yes
Fit statistics			
Municipalities	1,650	1,650	1,650
R^2	0.00548	0.27303	0.43602
Within R^2			0.20840

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 1: OLS regressions of robbery rate (per 100,000 residents) on governing party coalition. Reference category is PAN/PRI/PRD. Model (2) adds log population and log geographic area; model (3) additionally includes state fixed effects.

3 Sample Composition Relative to the Target Population

This section reports the composition of both survey samples against the population of the study frame. The frame is eligible voters in the 28 states holding municipal elections in 2027 for which the partisan treatments are well defined. Four states are excluded: Durango and Veracruz hold no municipal elections in 2027; Mexico City is excluded because its borough presidents lack municipal police authority, so the incumbent-performance logic of the treatments does not apply; and Oaxaca is excluded because a majority of its municipalities govern under *usos y costumbres* rather than through partisan elections, which makes the partisan comparison treatments inapplicable.

The two waves were assembled differently, and this matters for interpreting the tables below. Wave 1 respondents were recruited by the panel provider against marginal quotas—enforced independently per variable rather than cross-tabulated—on sex, age, socioeconomic level (SEL), and state. Those quotas were enforced at the start of fielding but relaxed as collection proceeded in order to reach the target sample size, so the achieved wave 1 sample of 3,028 respondents leans toward whichever cells the panel could fill most readily. Wave 2 was not quota-sampled at all. It recontacted wave 1 completers, and the 1,927 respondents in the wave 2 analysis sample are those who returned, identified the same home municipality in both waves, and passed the attention check. Wave 2's composition is therefore wave 1's composition filtered by differential attrition and by those two analysis restrictions, not the result of a separate sampling target.

Benchmarks come from three sources. Sex and state shares are computed from INEGI 2020 census counts restricted to the 28 states in the frame. Age shares are the INEGI 2020 distribution of the 18+ population in those states. SEL shares are AMAI NSE levels from ENIGH 2022 (*Estimaciones NSE 2022* 2022). The comparisons are drawn against these population shares rather than against the quotas set in the field, because the wave 1 SEL

quotas were themselves not proportional to the population distribution: they allocated 9.0% of the sample to the AB stratum against a population share of 7.3%, and 20.6% to the combined D/E stratum against a population share of 34.1%. Benchmarking against the quotas would therefore understate how far the achieved samples depart from the electorate.

Sex and Age

Table 2 reports sex and age. Departures on both dimensions are modest. Women are overrepresented by 2.5 percentage points in wave 1 and 2.8 points in wave 2. Age tracks the population closely in wave 1, where no bracket deviates by more than 1.9 points and the only visible gap is a shortfall in the 65+ bracket. Wave 2 skews moderately older: the youngest bracket falls 2.3 points below its population share while the 45–54 and 55–64 brackets run 1.7 and 1.9 points above, consistent with younger respondents being less likely to return for the second wave.

Table 2: Sample composition by sex and age. Population shares are INEGI 2020 counts for the 28 states in the study frame (sex) and the INEGI 2020 distribution of the 18+ population in those states (age). Differences are sample share minus population share, in percentage points.

		Wave 1 (N = 3,028)		Wave 2 (N = 1,927)	
	Population %	N	% (diff.)	N	% (diff.)
Sex					
Male	49.0	1,406	46.4 (−2.5)	889	46.1 (−2.8)
Female	51.0	1,622	53.6 (+2.5)	1,038	53.9 (+2.8)
Age					
18–24	17.3	539	17.8 (+0.5)	289	15.0 (−2.3)
25–34	22.5	698	23.1 (+0.5)	411	21.3 (−1.2)
35–44	20.2	653	21.6 (+1.3)	413	21.4 (+1.2)
45–54	17.0	507	16.7 (−0.3)	361	18.7 (+1.7)
55–64	11.7	349	11.5 (−0.2)	262	13.6 (+1.9)
65+	11.2	282	9.3 (−1.9)	191	9.9 (−1.3)

Socioeconomic Level

Socioeconomic level is the dimension on which the samples depart most sharply from the electorate, as shown in Table 3. The two waves are reported in separate panels because the SEL categories were recorded differently: wave 1 collapsed the D+, D, and E levels into a single bottom category at the point of collection, whereas wave 2 retained D+ separately and merged only D and E.

In wave 1, every stratum from C− upward is overrepresented—by 4.4 points for AB and 6.3 points for C+—while the combined bottom stratum, which accounts for 49.0% of the population, supplies only 29.2% of respondents. Wave 2 amplifies this pattern rather than correcting it. The AB stratum reaches 14.1% of the sample against a population share of 7.3%, nearly double its share of the electorate, and C+ reaches 21.2% against 12.0%. The lowest (D/E) stratum falls to 8.5% against a population share of 34.1%, roughly a quarter of its share of the electorate and the single largest deviation anywhere in the sample. Notably, the D+ stratum in wave 2 is very close to its population share (+1.2 points), so the shortfall is concentrated almost entirely in the poorest households, which are both least likely to be reachable through an online panel and least likely to return for a second wave.

State

Table 4 reports the distribution across the 28 states in the frame. The State of México is the most overrepresented, by 3.1 points in wave 1 and 5.0 in wave 2, and Nuevo León, Jalisco, and Querétaro are also above their population shares in both waves. The undersampled states are disproportionately poor and rural: Chiapas falls 2.3 points short in wave 1 and 2.8 in wave 2, Guerrero 1.6 and 1.8 points short, with Michoacán, Sinaloa, San Luis Potosí, and Zacatecas also below their population shares in both waves. Chiapas and Guerrero are reached at roughly half their population share in wave 1 and slightly less than half in wave 2.

Table 3: Sample composition by socioeconomic level (AMAI NSE). Population shares are AMAI NSE estimates from ENIGH 2022 (*Estimaciones NSE 2022* 2022), aggregated to match each wave's recorded categories. Differences are sample share minus population share, in percentage points.

Stratum	Population %	N	Sample % (diff.)
Panel A: Wave 1 (N = 3,027)			
AB	7.3	354	11.7 (+4.4)
C+	12.0	555	18.3 (+6.3)
C	15.3	608	20.1 (+4.8)
C−	16.4	625	20.6 (+4.2)
D+/D/E	49.0	885	29.2 (−19.8)
Panel B: Wave 2 (N = 1,926)			
AB	7.3	271	14.1 (+6.8)
C+	12.0	408	21.2 (+9.2)
C	15.3	395	20.5 (+5.2)
C−	16.4	377	19.6 (+3.2)
D+	14.9	311	16.1 (+1.2)
D/E	34.1	164	8.5 (−25.6)

Because the socioeconomic and geographic patterns are correlated—the undersampled states are also those where the D and E strata are most concentrated—the two dimensions reinforce rather than offset one another, and the samples tilt toward wealthier, more urban, and more central respondents. This has no bearing on the internal validity of the experiment, since treatment is randomized within the realized sample and the arms are balanced on pre-treatment covariates (Table 5). It does mean that the estimates are least informative about the poorest and most rural segments of the Mexican electorate, which is a substantive limitation here rather than a generic caveat: these are precisely the segments where local crime reporting is thinnest and where the informational environment the treatments simulate may differ most from what the sample experiences.

Table 4: Sample composition by state. Population shares are total-population shares from INEGI 2020, restricted to the 28 states in the study frame. Differences are sample share minus population share, in percentage points; entries marked ±0.0 round to zero.

		Wave 1		Wave 2	
State	Population %	N	% (diff.)	N	% (diff.)
Aguascalientes	1.4	50	1.7 (+0.3)	43	2.2 (+0.8)
Baja California	3.7	105	3.5 (−0.2)	58	3.0 (−0.7)
Baja California Sur	0.8	24	0.8 (±0.0)	13	0.7 (−0.1)
Campeche	0.9	35	1.2 (+0.3)	21	1.1 (+0.2)
Coahuila	3.1	97	3.2 (+0.1)	64	3.3 (+0.3)
Colima	0.7	22	0.7 (±0.0)	13	0.7 (±0.0)
Chiapas	5.4	93	3.1 (−2.3)	50	2.6 (−2.8)
Chihuahua	3.6	114	3.8 (+0.1)	70	3.6 (±0.0)
Guanajuato	6.0	175	5.8 (−0.2)	123	6.4 (+0.4)
Guerrero	3.4	55	1.8 (−1.6)	31	1.6 (−1.8)
Hidalgo	3.0	90	3.0 (±0.0)	51	2.6 (−0.4)
Jalisco	8.1	264	8.7 (+0.6)	173	9.0 (+0.9)
México	16.5	596	19.7 (+3.1)	414	21.5 (+5.0)
Michoacán	4.6	104	3.4 (−1.2)	63	3.3 (−1.3)
Morelos	1.9	70	2.3 (+0.4)	43	2.2 (+0.3)
Nayarit	1.2	31	1.0 (−0.2)	17	0.9 (−0.3)
Nuevo León	5.6	188	6.2 (+0.6)	129	6.7 (+1.1)
Puebla	6.4	215	7.1 (+0.7)	129	6.7 (+0.3)
Querétaro	2.3	87	2.9 (+0.6)	60	3.1 (+0.8)
Quintana Roo	1.8	67	2.2 (+0.4)	34	1.8 (±0.0)
San Luis Potosí	2.7	67	2.2 (−0.5)	40	2.1 (−0.7)
Sinaloa	2.9	66	2.2 (−0.8)	37	1.9 (−1.0)
Sonora	2.9	83	2.7 (−0.1)	52	2.7 (−0.2)
Tabasco	2.3	76	2.5 (+0.2)	48	2.5 (+0.2)
Tamaulipas	3.4	92	3.0 (−0.4)	54	2.8 (−0.6)
Tlaxcala	1.3	46	1.5 (+0.2)	29	1.5 (+0.2)
Yucatán	2.3	88	2.9 (+0.6)	51	2.6 (+0.4)
Zacatecas	1.6	28	0.9 (−0.7)	16	0.8 (−0.7)

4 Equivalence-Based Balance Tests

Table 5 assesses covariate balance across the six treatment arms using an equivalence-testing approach (Hartman and Hidalgo 2018), on the same estimation sample used in the main update analyses (respondents whose home municipality is identical across waves and who passed the attention check; $N \approx 1{,}927$). Conventional balance tests invert the burden of proof: they take equality of covariate means as the null hypothesis, so a non-significant result—which merely reflects low power—is treated as evidence of balance. Equivalence tests reverse this. The null hypothesis is that the arms differ *substantially*, and rejecting it provides positive evidence that any imbalance is smaller than a pre-specified, practically negligible threshold. For each covariate we fit a one-way ANOVA of the covariate on treatment arm and test, via Lakens (2017), whether the between-arm effect size (partial η^2) falls below an equivalence bound corresponding to a standardized mean difference of $d = 0.20$ standard deviations ($\eta^2 = 0.0099$), a conventional "small effect" benchmark. The omnibus form compares all six arms simultaneously, avoiding the multiple-comparison inflation of pairwise tests. Categorical covariates (region, socioeconomic level, sex, and pre-treatment vote-intention coalition) are entered as indicator variables and collapsed in the table to their least-balanced category; because that category is the hardest case, its passing the equivalence test implies every category does.

The arms are statistically equivalent on nearly every pre-treatment covariate: age, prior crime beliefs (crime rank, crime gap, home robbery rate), pre-treatment crime-handling evaluations, crime importance, and—aside from region—all demographic and partisan characteristics clear the equivalence bound ($p_{\text{equ}} < 0.05$). Region is the only pre-treatment covariate not to pass, and only marginally: four of its twenty-nine categories fail the equivalence test, yet even for these the observed between-arm effect sizes ($\eta^2 \approx 0.007$) fall *below* the 0.0099 bound. Their failure reflects insufficient power to positively certify equivalence in small regional cells rather than evidence of imbalance. The remaining exception, comparison-

party knowledge, is not a true pre-treatment covariate—it depends on which comparison municipalities a respondent happened to be shown, which is itself a function of treatment assignment—and is reported only for completeness.

Table 5: Equivalence balance tests across the six treatment arms (equivalence bound $d = 0.20$ SD, $\eta^2 = 0.0099$; $\alpha = 0.05$). Categorical covariates are collapsed to their least-balanced category.

Covariate	N	η_p^2	F	p_{equ}	Equivalent
Region (29 cat.)	1927	0.0071	2.736	0.1170	No
Socioeconomic level (7 cat.)	1926	0.0033	1.268	0.0067	Yes
Sex (2 cat.)	1927	0.0039	1.492	0.0126	Yes
Vote intention coalition (pre) (4 cat.)	1909	0.0035	1.351	0.0090	Yes
Age	1927	0.0038	1.454	0.0114	Yes
Comparison party knowledge	1927	0.0333	13.232	0.9996	No
Crime gap	1908	0.0050	1.905	0.0337	Yes
Crime handling (pre)	1927	0.0018	0.693	0.0006	Yes
Crime importance	1927	0.0041	1.569	0.0153	Yes
Crime rank (prior)	1927	0.0011	0.434	0.0001	Yes
Home robbery rate	1927	0.0022	0.837	0.0013	Yes

5 Manipulation Check

The treatments are informative only if respondents actually absorb what they are shown. Because both pre-treatment beliefs are elicited again after treatment in identical form, this can be checked directly: each respondent's post-treatment beliefs can be compared against the truth and against their own priors. I therefore measure accuracy before and after treatment and regress the change on treatment assignment, with the control arm as the omitted baseline.

Accuracy is measured separately for each belief, matching the two perception gaps used in the analysis. Crime-rate accuracy is the negative absolute log ratio of a respondent's robbery estimate to the actual rate, so it rises as the estimate approaches the truth. Relative-ranking accuracy is the Kendall τ_b concordance between a respondent's ordering of the comparison municipalities and the true ordering, so it rises as the perceived comparison approaches the real one. Both are reported as post-treatment minus pre-treatment changes, and both are oriented so that higher is better.

Figure 2 shows the result, and the pattern is the one the design requires. On crime-rate accuracy, every arm improves on control, including plain information: 0.31 for T1, 0.28 for T2, and 0.22 for both T3 and T4, with all four 95% intervals excluding zero. This is expected, since every treated arm states the home municipality's robbery rate.

The second panel is the one that validates the comparison manipulation specifically. On relative-ranking accuracy, plain information does essentially nothing: T1 is 0.07 with an interval of $[-0.05, 0.20]$. The three comparison arms all improve: 0.20 for T2, 0.19 for T3, and 0.18 for T4, with intervals excluding zero in each case. Telling respondents their own crime rate leaves their beliefs about relative standing unchanged; showing them a comparison moves those beliefs toward the truth. The manipulation therefore operates on the dimension it is meant to operate on, and the comparison arms are not simply a more elaborate way of delivering the same information as T1.

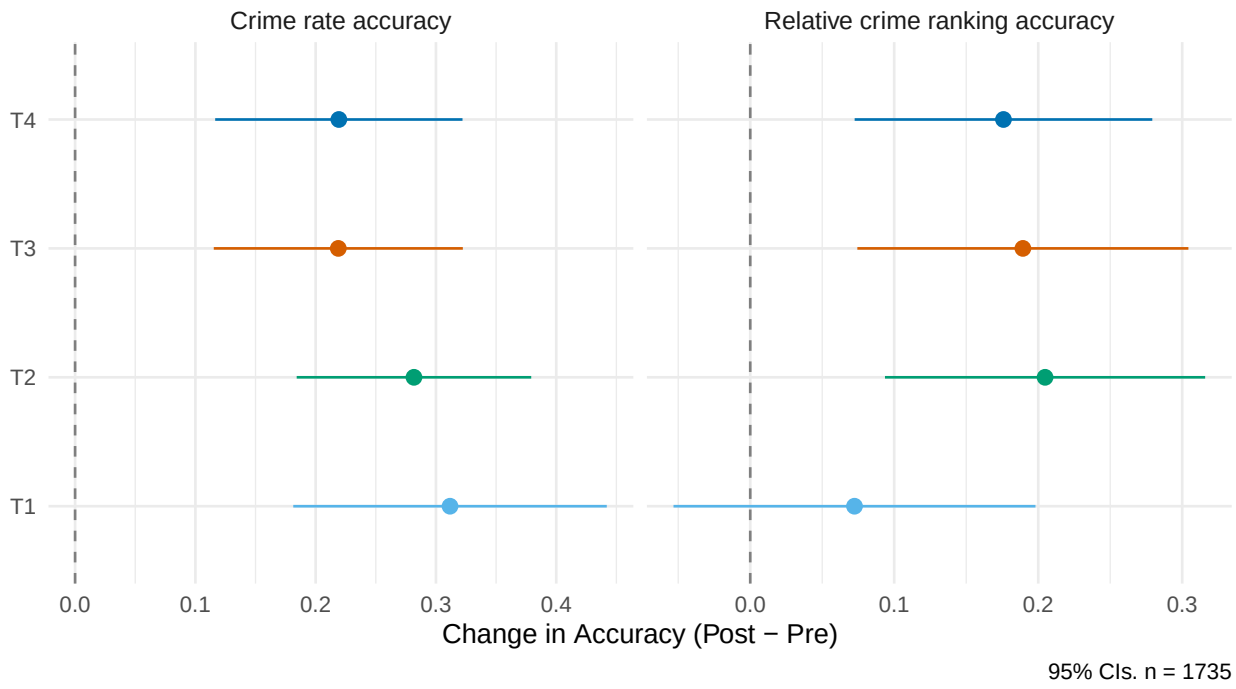

Figure 2: Change in belief accuracy from before to after treatment, by arm, relative to the control group. The left panel measures accuracy about the home municipality's robbery rate; the right measures accuracy about its ranking among the comparison municipalities. Both are oriented so that higher values indicate more accurate beliefs. Bars are 95% confidence intervals. $N = 1{,}735$.

6 An Alternative Rank-Gap Measure

The rank gap compares how many of the five municipalities shown to a respondent actually have lower robbery rates than their own with how many the respondent expected to. The main measure counts a comparison municipality as having fewer robberies whenever its rate falls below the home rate by any margin, which treats near-ties as genuine differences even though a respondent would be unlikely to perceive them that way. As a robustness check I rebuild the measure with a margin: a comparison municipality counts only if its robbery rate is at least 25% below the home rate. This is the same threshold used in the belief-updating analysis.

The stricter rule lowers the average actual rank of the home municipality, from 2.23 to 1.63, and so shifts the gap for most respondents. Figure 3 plots the treatment interaction coefficients under both measures, standardized so a point is the effect of a one-standard-deviation increase in the relevant gap. The crime-gap interactions are nearly unchanged: the same-coalition arm (T4) remains the only one distinguishable from zero, moving from -0.060 ($p = 0.013$) under the original measure to -0.062 ($p = 0.010$) under the 25% threshold. The rank-gap interactions are indistinguishable from zero under both measures. The result therefore does not depend on how finely the rank comparison is drawn.

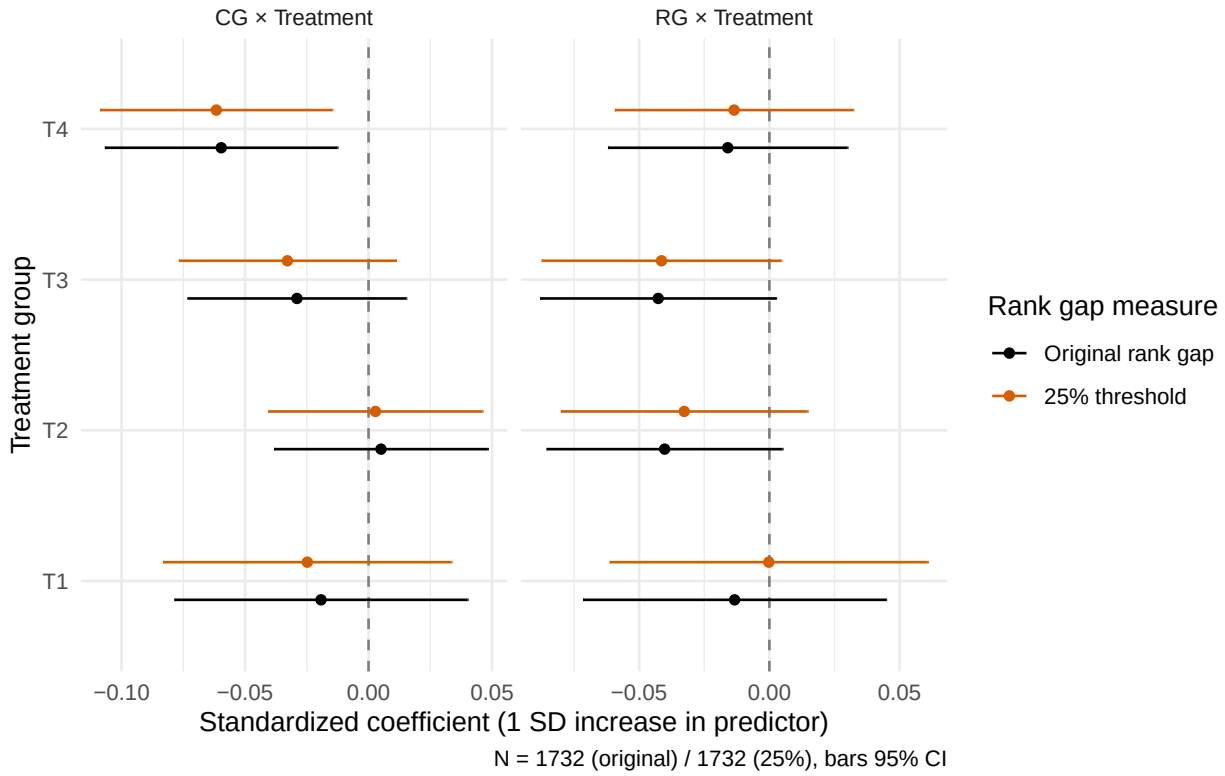

Figure 3: Treatment interaction coefficients for post-treatment incumbent vote intention under the original rank-gap measure and under the 25% threshold measure. Points are standardized coefficients (one SD of the relevant gap); bars are 95% confidence intervals. $N = 1{,}732$ under both measures.

7 Handling of Extreme Crime Gaps

The crime gap is the distance between a respondent's prior robbery estimate and the actual robbery count, and a small number of respondents give estimates far larger than any municipality's true rate. How those respondents are handled could in principle drive the vote result, so I refit the main model under alternative treatments of the gap and compare the interaction coefficients directly.

The main specification uses the logged gap, which compresses the whole scale so that extreme estimates carry no more leverage than a doubling anywhere else. A second specification uses the capped level gap, which leaves the measure untransformed but bounds extreme respondents' crime gap to two times the highest robbery rate observed. The outcome, sample, and controls are identical across the two.

A third specification takes the most direct route and simply drops the extreme respondents: those whose robbery estimate exceeds ten times the largest observed home rate, a cutoff of roughly 16,200 robberies per 100,000 residents. This removes 128 of the 1,732 respondents in the estimation sample, leaving 1,604. Where the first two specifications keep those respondents but limit their leverage, this one asks whether the result survives their removal altogether.

Figure 4 plots the treatment interaction coefficients under all three, standardized to a one-standard-deviation increase in each model's own gap predictor. The same-coalition arm (T4) is the only crime-gap interaction distinguishable from zero, and it is distinguishable from zero under every specification: -0.061 under the log gap, -0.055 under the capped level gap, and -0.056 once the extreme respondents are dropped, with 95% intervals of $[-0.108, -0.014]$, $[-0.103, -0.008]$, and $[-0.100, -0.011]$ respectively. The remaining arms are indistinguishable from zero throughout, as are all of the rank-gap interactions. The result therefore does not rest on how extreme crime gaps are handled: bounding their leverage and removing them entirely produce the same conclusion, and dropping the 128 most extreme

respondents leaves the T4 estimate essentially where the capped specification puts it.

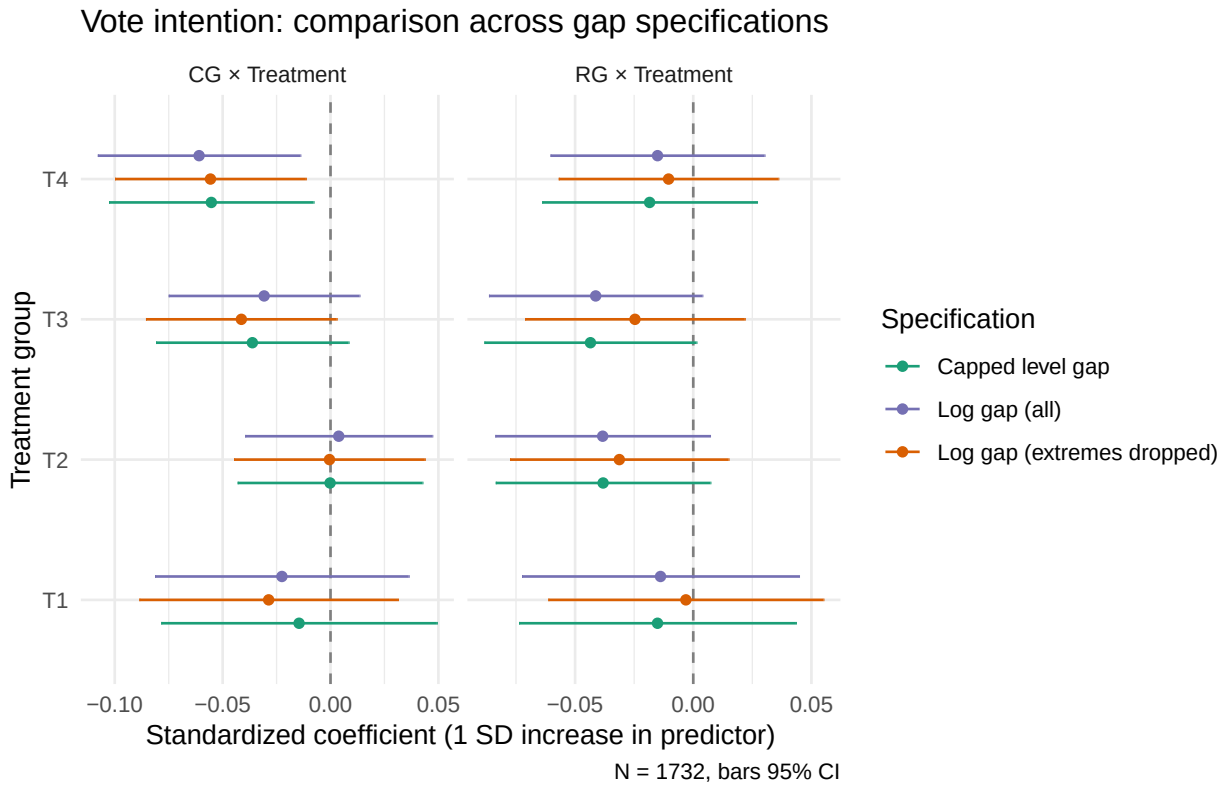

Figure 4: Treatment interaction coefficients for post-treatment incumbent vote intention under three treatments of the crime gap: the capped level gap, the log gap, and the log gap with extreme estimates dropped. Points are standardized coefficients (one SD of each model's own gap predictor); bars are 95% confidence intervals. $N = 1{,}732$ for the capped and log specifications and $N = 1{,}604$ for the extremes-dropped specification.

8 Census-Weighted Estimates

As Appendix 3 shows, the achieved sample departs from the electorate on socioeconomic level and geography. Because treatment is randomized within the realized sample, this does not threaten the internal validity of the experiment, but it does raise the question of whether the estimated effects would look the same in a sample that matched the population. To address this I reweight the sample and refit the main model.

Weights are constructed by raking, or iterative proportional fitting, to four marginal distributions (sex, age bracket, socioeconomic level, and state) using the same population benchmarks reported in Appendix 3. Raking matches each margin in turn and cycles until the weights stop moving, which is what makes it feasible here: post-stratifying on the full cross of the four variables would require weighting more cells than there are respondents, leaving most of them empty. Weights are normalized to mean one and trimmed to $[0.25, 4.01]$ so that respondents in thin cells cannot dominate the fit. The trimmed weights reproduce the target margins closely, with a maximum deviation of 0.3 percentage points on socioeconomic level and less on the other three. All 1,735 respondents are weightable. Kish's effective sample size is 811, or 47% of the 1,735 respondents actually used. The weighted estimates should therefore be expected to carry intervals roughly 1.5 times wider than the unweighted ones, before any change in the point estimates.

Figure 5 plots the treatment interaction coefficients under both estimators, scaled by the same unweighted standard deviations so the two series sit on a common axis. Both are fit on the same $N = 1{,}732$ rows. The same-coalition arm (T4) remains the only crime-gap interaction distinguishable from zero, and weighting moves it away from zero rather than toward it: from -0.061 ($p = 0.010$) unweighted to -0.096 ($p = 0.010$) weighted, with the interval widening from $[-0.108, -0.014]$ to $[-0.168, -0.023]$ as the loss of effective sample size predicts. The remaining arms are indistinguishable from zero under both estimators, as are all of the rank-gap interactions. The population-average effect is thus, if anything,

larger than the sample-average effect, though the two are well within each other's confidence intervals.

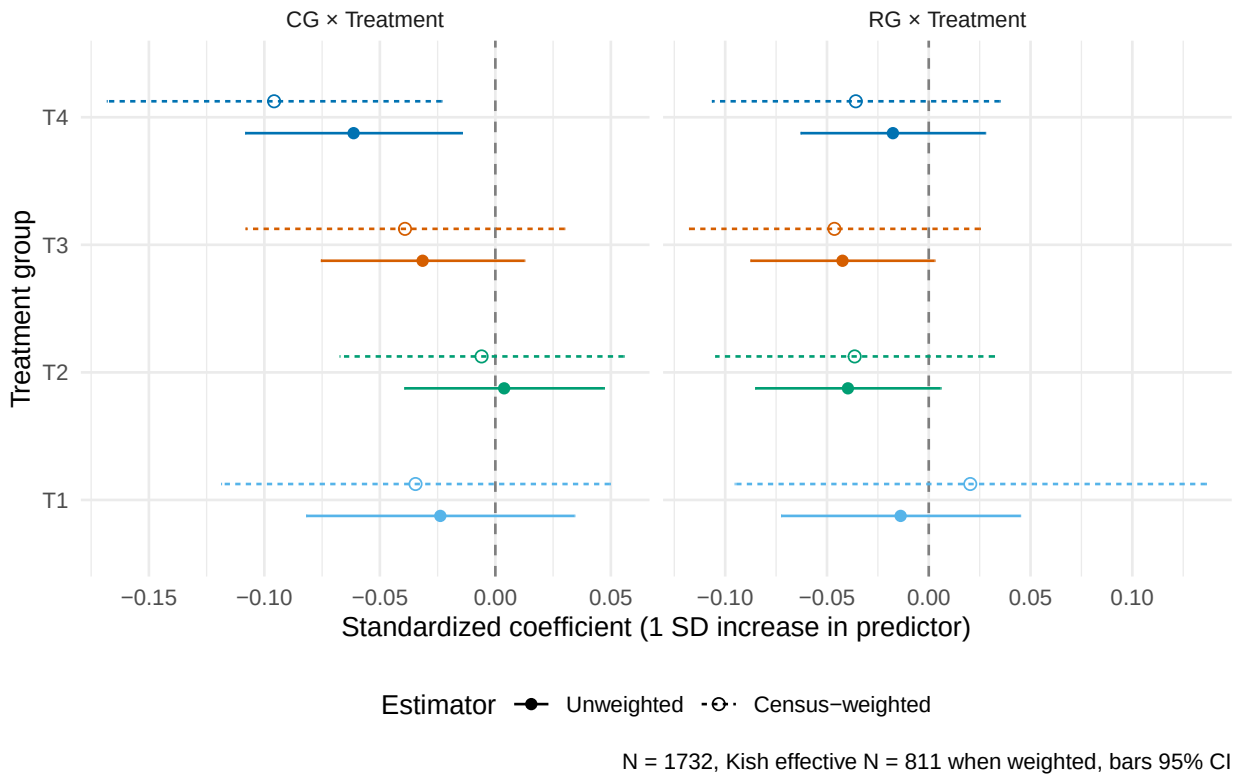

Figure 5: Treatment interaction coefficients for post-treatment incumbent vote intention, unweighted and with census raking weights. Points are standardized coefficients (one SD of the unweighted predictor); bars are 95% confidence intervals. $N = 1{,}732$ for both; Kish effective sample size under weighting is 811.

9 Respondents Who Know Their Governing Coalition

The same-coalition treatment can only operate through partisan identity for respondents who know which coalition governs their own municipality. A respondent who does not cannot recognize the comparison as same-coalition at all, and for them T4 is simply a comparison treatment with an unfamiliar party label attached. If the T4 result were an artifact of confusion—respondents reacting to the presence of party labels rather than to their content—it should weaken among respondents who have the knowledge the mechanism requires. I therefore refit the main model on the subset who correctly named their home municipality's governing coalition. Appendix 13 reports how accurate respondents are at this task; here the measure is used only to define the subgroup.

The subgroup is $N = 1{,}376$ of the 1,732 respondents in the estimation sample. Figure 6 overlays the two sets of interaction coefficients, both scaled by the full-sample standard deviations so they sit on a common axis. The T4 crime-gap interaction strengthens rather than weakens: from -0.061 ($p = 0.010$) in the full sample to -0.073 ($p = 0.007$) among respondents who know their governing coalition, and it remains the only crime-gap interaction distinguishable from zero in either sample. This runs against the confusion account, and is what the partisan-identity mechanism predicts.

The subgroup also speaks to the caveat raised in the main text. The theory predicts that the opposite-coalition arm (T3) reveals information about competence conditional on voters' priors about how the two parties differ, which should show up as a nonzero $RG \times T3$ interaction. In the full sample it does not ($p = 0.066$). Among respondents who know their governing coalition it does: -0.072, with a 95% interval of $[-0.124, -0.020]$ and $p = 0.006$. Since the treatments never tell respondents which party governs their own municipality—a deliberate choice, so that information about own-party governance is not conflated with the benchmark manipulation—the full-sample estimate pools respondents for whom the partisan contrast is legible with those for whom it is not. That the interaction emerges precisely where

the contrast can register is the pattern the theory implies, and it suggests the full-sample null reflects noise rather than an absent effect.

Neither result is an artifact of testing several coefficients at once. Applying Benjamini-Hochberg false discovery rate control to the interaction family—the three $CG\times$ arm and three $RG\times$ arm terms for T2–T4, the same family convention used elsewhere in this appendix—leaves both significant, with $q = 0.020$ for $CG \times T4$ and $q = 0.020$ for $RG \times T3$. Widening the family to all eight interactions shown in the figure raises both to $q = 0.027$, still below conventional thresholds. No other interaction in either family comes close: the next smallest q-value is 0.12.

The subgroup is defined by a pre-treatment measure, so the comparison is internally valid. It remains a subgroup analysis rather than a randomized contrast, and the estimates for respondents who know their governing coalition need not carry over to those who do not.

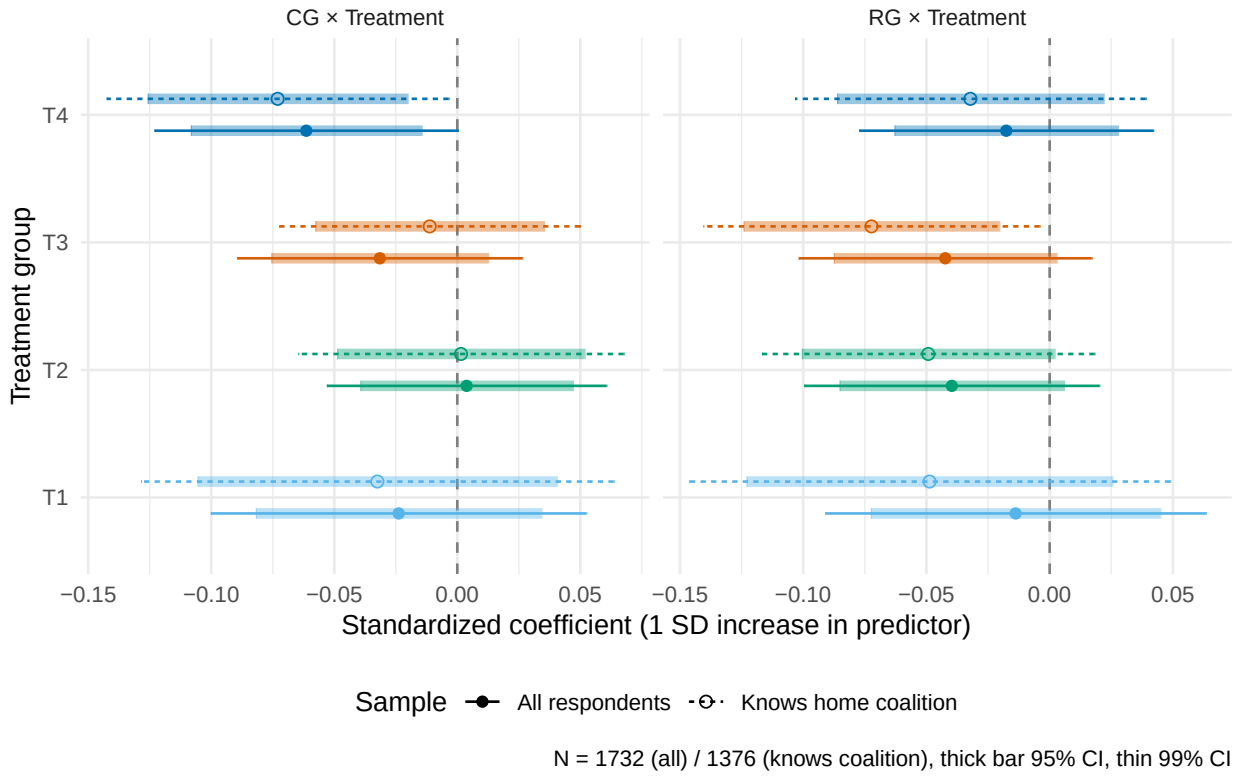

Figure 6: Treatment interaction coefficients for post-treatment incumbent vote intention, in the full estimation sample and among respondents who correctly named their home municipality's governing coalition. Points are standardized coefficients (one SD of the full-sample predictor); thick bars are 95% and thin bars 99% confidence intervals. $N = 1{,}732$ and $N = 1{,}376$ respectively.

10 Crime-Importance Moderation of Treatment Effects

Do the informational treatments move incumbent vote intention differently depending on how personally important respondents consider crime? To test this, I estimate a linear probability model of post-treatment incumbent vote intention on the logged crime-rate gap and the rank gap, each fully interacted with treatment arm and with respondents' self-reported crime importance (an integer 1–5 scale), controlling for prior incumbent vote. The model is fit on the same estimation sample as the main vote-update analysis (respondents whose home municipality is identical across waves and who passed the attention check; $N = 1{,}732$), with HC2 robust standard errors.

Figure 7 plots the implied treatment effect of each arm relative to control across the five levels of crime importance, holding both belief gaps at their sample medians. Because crime importance enters linearly, each arm's effect traces a straight line. Every point's 95% confidence interval includes zero, and no arm's importance slope is statistically distinguishable from zero (all $p \geq 0.20$).

Formal tests confirm the absence of robust moderation. A joint Wald test that all twelve treatment-by-importance interaction terms equal zero is only marginal ($\chi^2 = 21.88$, df $= 12$, $p = 0.039$), and this omnibus signal is carried by the two-way arm-by-importance terms rather than by the theoretically substantive three-way terms. A joint test restricted to the eight gap $\times$ arm $\times$ importance coefficients, which capture whether crime importance changes *how* voters respond to the crime information, is far from significant ($\chi^2 = 9.07$, df $= 8$, $p = 0.34$). Figure 8 relaxes the linearity assumption by entering crime importance as a set of indicators; the level-by-level estimates are noisier but tell the same story, with no level significantly different from zero for any arm. Given the modest cell sizes once the sample is divided across arms and five importance levels, these results are best read as an absence of *detectable* moderation—consistent with low power for a triple interaction—rather than positive evidence of none.

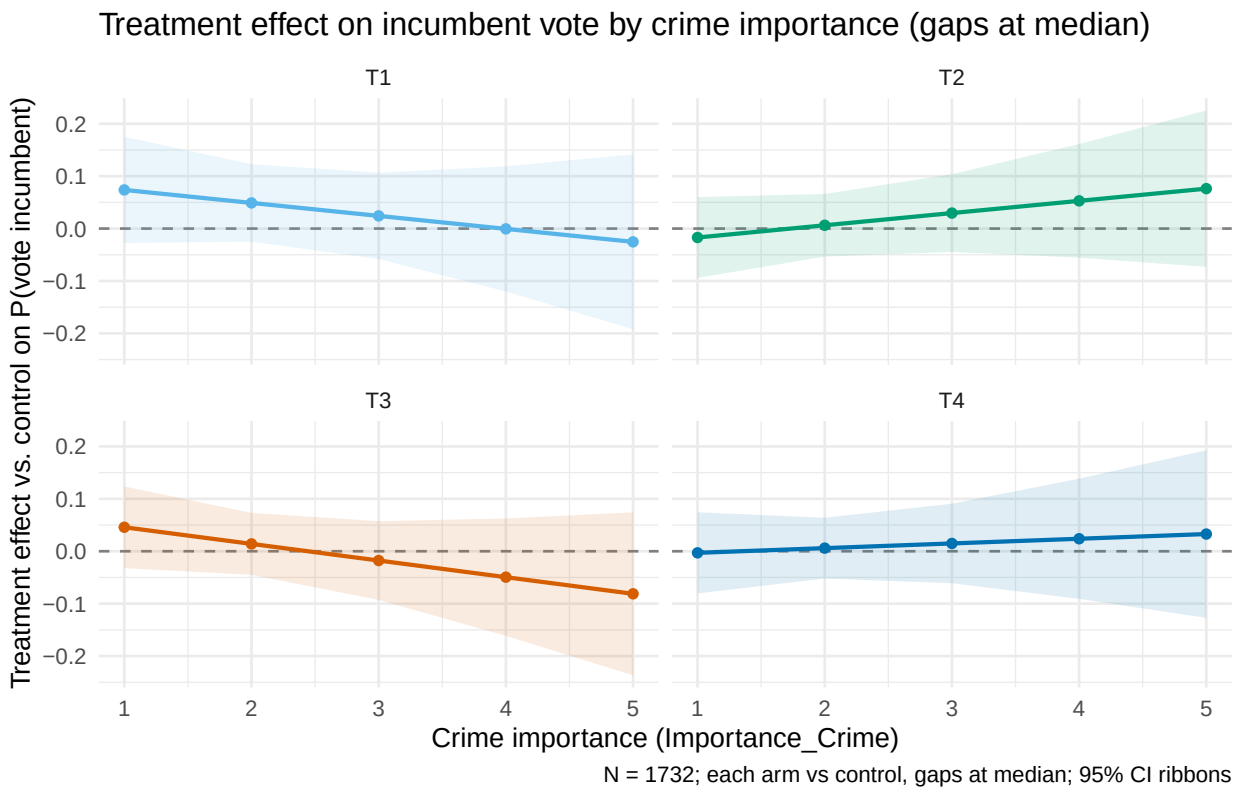

Figure 7: Treatment effect on post-treatment incumbent vote intention (relative to control) across levels of self-reported crime importance, by arm, with the crime-rate gap and rank gap held at their sample medians. Crime importance enters the model linearly, so each arm's effect is a straight line. Bands are 95% confidence intervals.

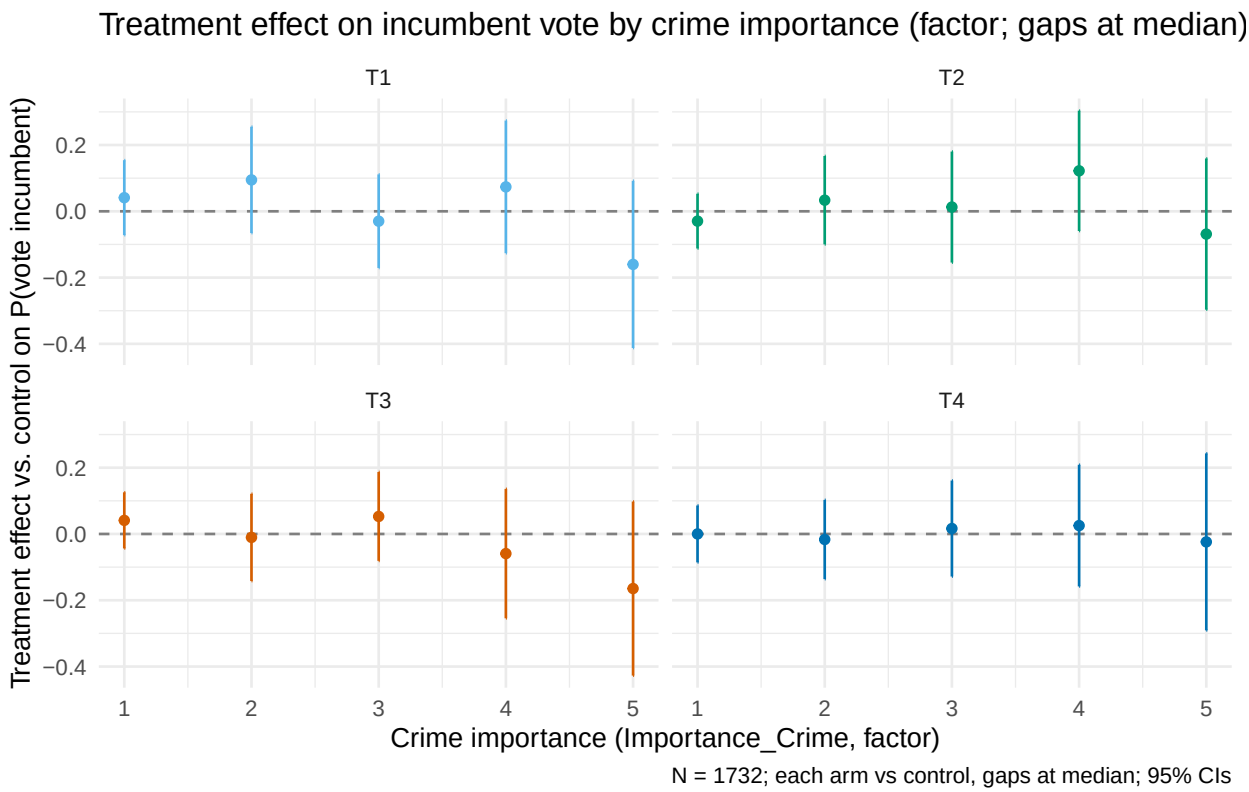

Figure 8: Robustness of Figure 7 with crime importance entered as a factor, allowing each level to differ freely. Points are the treatment effect relative to control at each importance level, with the belief gaps held at their medians; bars are 95% confidence intervals.

11 Vote Updating When Crime Is a Top Issue

The moderation tests above use the full estimation sample and let crime importance enter the model as a covariate. As a subgroup counterpart, I refit the main log-crime-gap vote model on the respondents who rank crime among their three most important issues. Because the wave 1 issue ranking records the position of crime among five issues, this is the subset for whom crime is plausibly decisive rather than incidental, and it is the subgroup for which the informational treatments should bite hardest.

Table 6 reports the eight gap-by-arm interaction terms for the $N = 1{,}432$ respondents in this subset. The same-coalition arm (T4) is the only one whose crime-gap interaction is distinguishable from zero ($\hat{\beta} = -0.013$, SE $= 0.004$, $p = 0.002$): respondents in T4 who learn that their municipality is doing worse than they believed shift away from the incumbent. The other three crime-gap interactions are an order of magnitude smaller or wrong-signed and none approaches conventional significance ($p \geq 0.39$), and the four rank-gap interactions are uniformly indistinguishable from zero ($p \geq 0.16$). The T4 estimate is somewhat larger in this subset than in the full sample, where the same coefficient is -0.010 (SE $= 0.004$, $p = 0.013$), but the two are well within each other's confidence intervals. The pattern therefore matches the full-sample result rather than departing from it, which is what one would expect if crime importance is not a strong moderator.

This subgroup split is on a pre-treatment measure, so the comparison is internally valid, but restricting the sample costs power on precisely the interaction terms of interest. These estimates should be read as consistent with the main analysis, not as independent confirmation of it.

Table 6: Treatment interactions among respondents who rank crime in their top 3 issues. Entries are linear-probability coefficients for post-treatment incumbent vote intention, with HC2 robust standard errors in parentheses and 95% confidence intervals in brackets. The model also includes each arm's main effect, both gap main effects, pre-treatment coalition, actual robbery rank, and prior incumbent vote; these are omitted here. Control is the reference arm. $N = 1{,}432$, $R^2 = 0.496$. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Arm	**Estimate**	**SE**	**95% CI**
Crime-rate gap (log) × arm			
T1	−0.0040	(0.0053)	[−0.0145, 0.0064]
T2	0.0034	(0.0040)	[−0.0044, 0.0112]
T3	−0.0031	(0.0042)	[−0.0112, 0.0051]
T4	−0.0134**	(0.0043)	[−0.0218, −0.0050]
Rank gap × arm			
T1	−0.0161	(0.0188)	[−0.0530, 0.0207]
T2	−0.0193	(0.0148)	[−0.0482, 0.0097]
T3	−0.0221	(0.0155)	[−0.0525, 0.0084]
T4	−0.0048	(0.0153)	[−0.0349, 0.0253]

12 Salience Moderation of the Same-Coalition Effect

A natural alternative reading of the T4 result is that same-coalition comparisons work not because they are politically informative but because they happen to be the comparisons respondents find most relevant. The main text addresses this in levels: respondents in T2 were shown comparisons they were, on average, *more* inclined to select for themselves than those shown in T4, so familiarity predicts the opposite of the observed ordering. This appendix tests the implication that a familiarity account makes about slopes rather than levels—if salience is what carries the effect, the effect should be larger for respondents shown more salient comparisons.

I score each comparison municipality with the wave 1 benchmark-selection model reported in the main text, yielding the model-implied log-odds that a respondent would have chosen it as a benchmark, and collapse the (up to four) comparisons each respondent saw into a single salience index. The primary index is the mean, matching the quantity reported in the main text; because a respondent shown one highly comparable municipality and three irrelevant ones has a low mean but a high maximum, I also report the maximum, which is the right summary if what moves a respondent is a single compelling benchmark rather than the average quality of the set.

The quantity of interest is the cross-partial $\partial^2 \Pr(\text{incumbent vote})/\partial CG\,\partial \text{Salience}$: how much a respondent's salience changes the slope on the crime-level perception gap. I estimate it with a linear probability model of post-treatment incumbent vote intention on treatment arm fully interacted with both perception gaps and with salience, controlling for prior incumbent vote, with HC2 standard errors. Each comparison is fit on its own two arms rather than on one pooled model, so no slope is constrained to be common across arms the comparison does not involve; the two contrasts therefore have slightly different samples. Two contrasts are reported: T4 against control, which asks whether salience moderates the treatment effect at all, and T4 against T2, which holds the presence of a comparison fixed and so isolates

whether the partisan content of the benchmark, rather than its salience, is doing the work.

Table 7 reports the results. No contrast is distinguishable from zero under either index: the T4-versus-control difference is +0.32 percentage points per standard deviation of salience ($p = 0.42$) with the mean index and +0.51 ($p = 0.22$) with the maximum, while the T4-versus-T2 differences are −0.33 ($p = 0.45$) and −0.20 ($p = 0.64$). The signs reverse between the two contrasts and the two aggregations agree throughout, which is the pattern expected of noise around zero rather than of a moderating effect the design is failing to resolve.

These intervals are, however, too wide to convert into positive evidence. The upper bound of the T4-versus-control interval under the mean index, 1.10 percentage points per standard deviation, is nearly the magnitude of the $T4 \times CG$ coefficient itself in the same model (−1.16 percentage points), so the data cannot rule out salience moderation as large as the effect it would be explaining. As with the crime-importance moderation tests above, this is best read as an absence of *detectable* moderation—the design has limited power for a triple interaction estimated on two arms at a time—rather than as positive evidence of none. The level-based argument in the main text, which rests on a design fact rather than on the precision of an interaction estimate, remains the stronger of the two.

One feature of the table warrants comment. The salience slope within the control arm is itself negative and, in isolation, statistically distinguishable from zero (−0.51 and −0.63 percentage points per standard deviation under the two indices). This is not a placebo failure. Control respondents receive no crime information, so neither their perception gap nor the salience of the municipality names they were shown has any treatment content; both are pre-treatment respondent characteristics, and neither is randomly assigned. The salience index is constructed partly from home-municipality population, state, and governing coalition, so an association between it, prior misperception, and vote intention among untreated respondents is a descriptive fact about who lives where rather than evidence about the experiment. What randomization licenses is the *difference* across arms, which is the column the test is read from.

Neither the index nor these tests were pre-registered; the index is constructed after the fact from the wave 1 model, and this analysis should be read as exploratory.

Table 7: Salience moderation of the crime-gap slope. Each cell is the cross-partial $\partial^2 \Pr(\text{incumbent vote})/\partial CG\,\partial \text{Salience}$, in percentage points per one standard deviation of the salience index. Each row is a separate linear probability model fit on its own two arms, with HC2 standard errors. The "Difference" column is the triple-interaction coefficient and is the test of whether the two arms differ in how salience moderates the response to crime information.

Comparison	Ref. arm	T4	Difference	95% CI	p	N
Panel A: Mean salience index						
T4 vs. control	-0.51	-0.19	0.32	[-0.46, 1.10]	0.42	769
T4 vs. T2	0.14	-0.19	-0.33	[-1.17, 0.52]	0.45	773
Panel B: Max salience index						
T4 vs. control	-0.63	-0.12	0.51	[-0.30, 1.31]	0.22	769
T4 vs. T2	0.07	-0.13	-0.20	[-1.06, 0.66]	0.64	773

13 Governing Coalition Knowledge

I assess whether respondents can correctly identify the governing coalition of comparison municipalities shown during treatment. For each comparison municipality, respondents selected one of three coalitions (MORENA/PVEM/PT, PAN/PRI/PRD, or MC) or "don't know." I estimate linear probability models with respondent fixed effects, identifying effects purely from within-respondent variation across the four comparison municipalities each respondent evaluates.

Table 8 reports three specifications. Column (1) tests whether respondents are more accurate for municipalities they selected as benchmarks in Wave 1. Column (2) tests whether respondents are more accurate for municipalities governed by the coalition they voted for (pre-treatment vote intention). Column (3) tests whether respondents are more accurate for municipalities governed by the same coalition as their home municipality. The smaller sample size in column (2) reflects respondents who did not report a pre-treatment vote intention that maps to one of the three major coalitions.

The home-coalition match effect (column 3) is the largest and most precisely estimated: from a baseline accuracy of 29%, respondents are 22 percentage points more likely to correctly identify the governing coalition when the comparison municipality is governed by the same coalition as their home municipality. The W1 benchmark effect (column 1) is smaller but statistically significant: from a baseline of 39%, self-selected municipalities are about 8 percentage points more likely to be correctly identified. Surprisingly, the vote-coalition match effect (column 2) is negative and not statistically significant, suggesting that respondents do not have better knowledge of municipalities governed by their preferred party. Overall, geographic and political proximity to one's home municipality—rather than partisan alignment with one's own preferences—appears to drive knowledge of local governing coalitions.

Table 8: Governing coalition knowledge: predictors of accuracy

	(1)	(2)	(3)
W1 benchmark	0.079***		
	(0.018)		
Vote-coalition match		-0.011	
		(0.038)	
Home-coalition match			0.221***
			(0.030)
Baseline accuracy	0.390	0.398	0.288
Respondent FE	Yes	Yes	Yes
Observations	7,293	6,979	7,283
Respondents	1,827	1,724	1,827

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

14 Power Analysis

I conduct a simulation-based power analysis (Arnold et al. 2011). I simulate the estimating equation with an outcome ranging from 0 to 100, centered at 50 with within-unit standard deviation 20 ($Y \sim (50, 20^2)$; values outside $[0, 100]$ are truncated). The ranking perception gap measure is drawn as $RG_i \sim \mathcal{N}(0, 1.5)$, rounded to the nearest integer and clamped to $[-4, 4]$. The coefficient of interest is γ_k; I define the standardized interaction effect size as $\delta = \gamma_k \cdot \text{SD}(RG_i)/\text{SD}(Y)$. Also of interest is β_1 for testing whether plain information affects beliefs, where CG_i is calculated similarly. I fix the sample size at $N = 2{,}000$ Wave 2 retained respondents (approximately 400 respondents per comparison treatment arm) and run 1,000 replications across a continuous range of effect sizes. I report power under both an uncorrected threshold ($\alpha = 0.05$) and a Bonferroni-corrected threshold ($\alpha = 0.05/5 = 0.01$) to account for the primary interaction tests.

Figures 9 and 10 display simulated power as a function of effect size under each threshold, with the minimum detectable effect (MDE) at 80% power annotated for each curve. Under either assumption, I am well-powered to pick up small (< 0.3 Cohen's d) effects. These effect sizes may be conservative relative to the meta-analytic evidence in Incerti (2020).

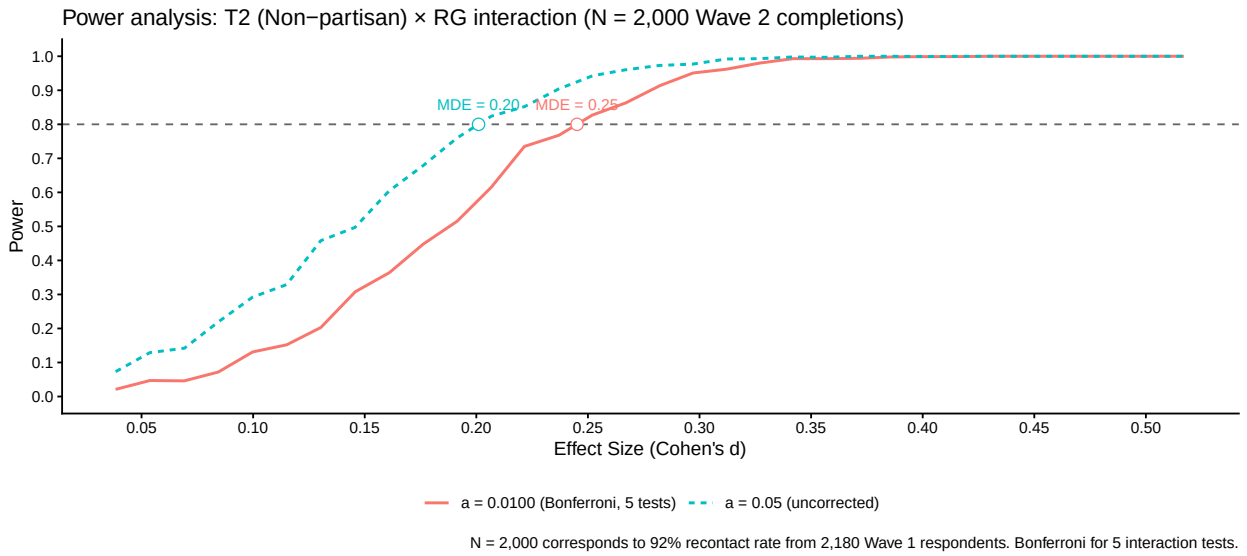

Figure 9: Simulated power as a function of interaction effect size (δ) at $N = 2{,}000$ Wave 2 completions. Circles mark the minimum detectable effect at 80% power under each significance threshold. Displays power for γ_k, $k \in \{2, 3, 4\}$.

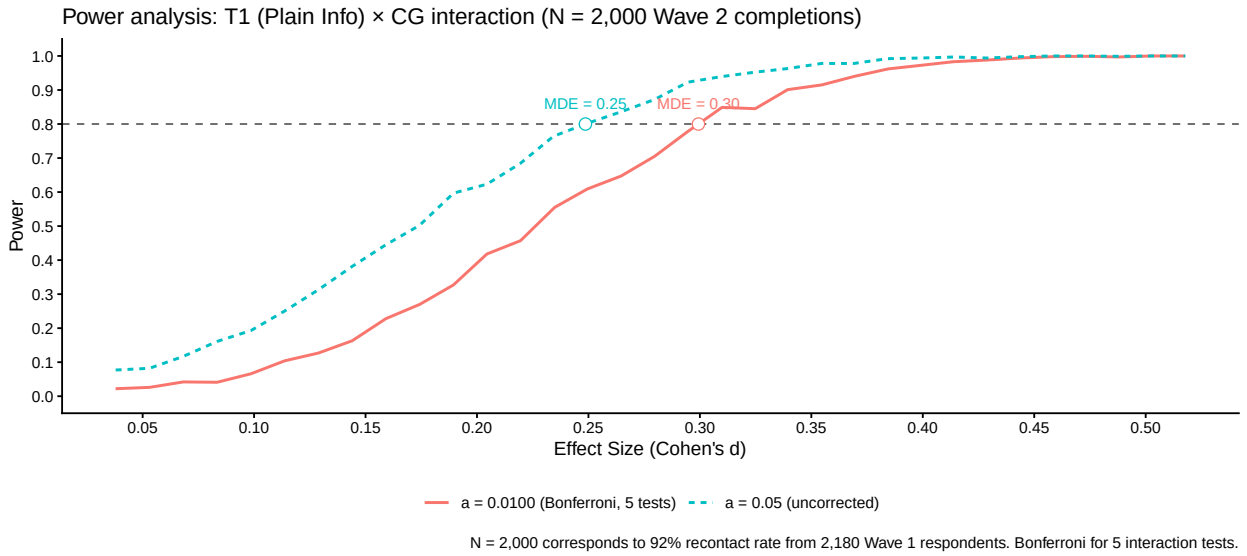

Figure 10: Simulated power as a function of interaction effect size (δ) at $N = 2{,}000$ Wave 2 completions. Circles mark the minimum detectable effect at 80% power under each significance threshold. Displays power for β_1.

15 Belief Results with Control2 as the Baseline

The belief models in the main text take Control as the omitted baseline. Control shows placebo text about rainfall in the respondent's own municipality and no comparison of any kind. That is the right baseline for the crime-level gap, which is a statement about the respondent's own municipality, but it is a poor one for the ranking gap. RG_i is by construction a comparison quantity—the difference between the standing the respondent is shown and the standing they expected—so an RG_i interaction estimated against Control combines the content of the comparison with the mere fact of having been shown one. Control2 separates the two: it pairs the same placebo rainfall text with a comparison bar chart, so benchmarking against it nets out the presence of a chart and leaves only what the chart said.

That the distinction matters is not hypothetical. Restricted to the two placebo arms, Control2 has a ranking-gap slope of its own that is significantly negative on both belief outcomes (-2.67, $p = 0.007$ for the incumbent rating; -2.45, $p = 0.015$ for the composite), while its crime-gap slope is indistinguishable from zero on both ($p = 0.93$ and $p = 0.31$). Respondents shown a rainfall chart move their evaluations of the incumbent with the rank that chart displayed, even though the chart carries no information about crime or about the incumbent's handling of it. This is precisely the "saw a chart" component that a Control baseline leaves inside the RG_i interactions.

Table 9 re-estimates the belief models with Control2 as the omitted baseline and Control dropped, alongside the published Control-baseline estimates. The specification is otherwise identical. For Control2 respondents the displayed rank is rebuilt from the precipitation figures they actually saw rather than from robbery rates; RG_i's pre-treatment component is left alone, since the pre-treatment ranking question asks about crime and the rainfall chart does not speak to it. Dropping Control rather than Control2 reduces the sample from 1,735 to 1,533.

Panel B is the result that motivates the appendix. Against Control, the ranking-gap

Table 9: Perception-gap × arm interactions for two belief outcomes, estimated against Control (placebo text, no comparison) and against Control2 (placebo text with a comparison chart). Coefficients are standardized to a 1 SD increase in the respective gap; t-statistics from HC2 robust standard errors in brackets. $N = 1{,}735$ against Control and $N = 1{,}533$ against Control2.

	Incumbent		**Incumbent − other**	
Term	vs. Control	vs. Control2	vs. Control	vs. Control2
Panel A: crime-level gap (CG)				
$CG\times$ T1	−1.93	−1.55	−1.61	0.25
	[−0.97]	[−0.71]	[−0.81]	[0.11]
$CG\times$ T2	−0.45	−0.03	−0.68	1.20
	[−0.28]	[−0.02]	[−0.48]	[0.67]
$CG\times$ T3	0.20	0.62	0.19	2.06
	[0.13]	[0.35]	[0.12]	[1.10]
$CG\times$ T4	−2.21	−1.82	−2.99	−1.14
	[−1.28]	[−0.95]	[−2.01]	[−0.62]
Panel B: ranking gap (RG)				
$RG\times$ T1	−2.40	0.05	−1.76	0.79
	[−1.44]	[0.03]	[−0.97]	[0.34]
$RG\times$ T2	−3.72	−1.32	−3.16	−0.62
	[−2.82]	[−0.72]	[−2.41]	[−0.31]
$RG\times$ T3	−1.57	0.86	−2.41	0.11
	[−1.08]	[0.45]	[−1.53]	[0.05]
$RG\times$ T4	−3.14	−0.75	−2.93	−0.41
	[−2.25]	[−0.40]	[−2.16]	[−0.20]

interaction is significantly negative in T2 and T4 on both outcomes. Against Control2, all four arms retain between 14 and 36 percent of their magnitude on the two outcomes where they had been significant, and none of the eight interactions is distinguishable from zero. The ranking-gap belief result is therefore better read as the effect of being shown a comparison than as the effect of what the comparison contained, and I treat it as null in the main text.

Panel A does not survive this baseline swap cleanly either, and I state that rather than claim it as corroboration. The composite $CG\times$ T4 interaction falls from -2.99 to -1.14 and is no longer distinguishable from zero. Both terms move against it: the point estimate shrinks and the standard error rises from 1.49 to 1.83 as the sample falls by 202 respondents. Control2 is assigned with probability 0.1 and Control with probability 0.2, so the arm that anchors this specification is the smallest in the design and the comparison is underpowered by construction. The crime-gap result reported in the main text rests on the Control-baseline specification, and this table should be read as showing that the ranking-gap and crime-level-gap results differ in kind—the first is eliminated by the placebo-chart baseline, the second is attenuated and imprecisely estimated under it—rather than as an independent test that the crime-gap result passes.

The Control-baseline columns above are re-estimated within this comparison and use 1,735 respondents rather than the 1,717 in the main text, because respondents with missing pre-treatment coalition preference are retained here as an "Other" category instead of being dropped. The resulting estimates differ marginally from the corresponding column of the main-text table.

16 Mediation Analysis

The results in the main text establish two things separately: same-coalition comparisons move vote intention, and they move beliefs about the incumbent relative to the other coalitions. They do not establish that the first happens *because of* the second. A treatment can shift a belief and shift a vote without the belief being what carries the vote. Mediation analysis splits the total effect of the same-coalition benchmark on incumbent vote intention into two pieces: the causal mediation effect and the direct effect. The causal mediation effect is the portion of the main effect that operates through a specified intermediate variable, and the direct effect is the portion of the effect that reaches the vote by other routes. Comparing their magnitudes says how much of the treatment's work the proposed mechanism is actually doing. This analysis was pre-registered as exploratory, and the design lacks power to detect small mediation effects, which is why it is reported here rather than in the main text.

The mediator I use is the incumbent-versus-other-coalitions rating, the only belief measure reported in the main text that responds detectably to T4. I estimate the decomposition using the two-equation approach of Imai, Keele, and Yamamoto (2010): one model for the mediator and one for vote intention, with treatment assignment, both perception gaps and their interactions with arm, and pre-treatment controls entering each. Because the treatments deliver good or bad news depending on what a respondent believed beforehand, a single average over respondents would obscure the pattern by averaging a positive and a negative effect. I therefore trace both quantities across the range of the crime-level perception gap, holding the ranking gap at its mean, with confidence intervals from a quasi-Bayesian approximation.

The mediator equation regresses the rating M_i on treatment assignment, the perception gaps and their interactions with arm, and the pre-treatment ratings from which the post-

treatment measure is built:

$$M_i = \alpha_0 + \alpha_1 R_i^{\text{inc}} + \alpha_2 R_i^{\text{opp}} + \alpha_3 CG_i + \alpha_4 RG_i + \sum_{k=1}^{4} \delta_k T_{ik} + \sum_{k=1}^{4} \beta_k T_{ik} \times CG_i + \sum_{k=1}^{4} \gamma_k T_{ik} \times RG_i + \mathbf{X}_i' \boldsymbol{\lambda} + \varepsilon_i \tag{1}$$

The outcome equation adds the mediator to that same specification, with incumbent vote intention V_i as the outcome, and interacts the mediator with arm:

$$V_i = \pi_0 + \tau M_i + \sum_{k=1}^{4} \theta_k T_{ik} \times M_i + \pi_1 R_i^{\text{inc}} + \pi_2 R_i^{\text{opp}} + \pi_3 CG_i + \pi_4 RG_i + \sum_{k=1}^{4} \rho_k T_{ik} + \sum_{k=1}^{4} \phi_k T_{ik} \times CG_i + \sum_{k=1}^{4} \psi_k T_{ik} \times RG_i \tag{2}$$

The θ_k terms are necessary rather than optional. Constraining the mediator to a single coefficient τ common to all arms forces $\text{ACME}(g) = \tau\,(\delta_4 + \beta_4 g)$, so the mediated curve becomes the mediator equation's T4 coefficient rescaled by a constant: its slope, and the location of its zero crossing, would be algebraic consequences of the functional form rather than features of the data. The restriction is also substantively inappropriate for this design, since the argument under test is that the mapping from beliefs to vote intention depends on which benchmark the respondent received. With θ_k estimated, the average causal mediation effect differs by condition; the reported quantity is the average of the effect under T4 and under Control, which in practice are close (for example 0.037 and 0.034 at the good-news end of the plotted range). where R_i^{inc} and R_i^{opp} are respondent i's pre-treatment ratings of the incumbent and of the other coalitions, CG_i and RG_i are the crime-level and ranking perception gaps, $\mathbf{X}_i$ is pre-treatment coalition preference, and arms are indexed $k \in \{1, 2, 3, 4\}$ for T1–T4 with Control as the omitted baseline. Control2 is excluded from this analysis, and the sample is restricted to respondents whose home municipality is unchanged across waves and who passed the attention check.

Let $M_i(t)$ denote the mediator under arm t, and $V_i(t, m)$ the vote intention under arm t when the mediator is set to m. For the T4-versus-Control contrast, the two quantities of

interest are

$$\text{ACME}(g) = \mathbb{E}\big[V_i\big(\text{T4}, M_i(\text{T4})\big) - V_i\big(\text{T4}, M_i(\text{Control})\big) \;\big|\; CG_i = g\big] \tag{3}$$

$$\text{ADE}(g) = \mathbb{E}\big[V_i\big(\text{T4}, M_i(\text{Control})\big) - V_i\big(\text{Control}, M_i(\text{Control})\big) \;\big|\; CG_i = g\big] \tag{4}$$

and their sum is the total effect of T4 at $CG_i = g$. The average causal mediation effect holds the arm fixed at T4 and varies only the rating, isolating what the belief carries; the average direct effect holds the rating fixed and varies only the arm, capturing everything else. Both are evaluated at each value g across the range of CG_i, with RG_i held at its mean, which is what produces the two curves in the main text. Confidence intervals come from the quasi-Bayesian approximation of Imai, Keele, and Yamamoto (2010), drawing 1,000 simulations of the model parameters from their estimated sampling distribution and recomputing both quantities at each draw.

Identification of these quantities requires sequential ignorability: treatment assignment is independent of the potential outcomes and potential mediator values, and, conditional on treatment and the covariates above, the observed mediator is independent of the potential outcomes. Randomization secures the first condition by construction. The second is an assumption about the mediator, which is not randomized, and the design does not enforce it—any unobserved characteristic that shapes both a respondent's relative rating of the incumbent and their vote intention would violate it. The estimates should accordingly be read as a decomposition that is informative about magnitudes under that assumption rather than as a fully experimental result.

Figure 11 reports the decomposition. Both components slope downward as the news worsens, as expected: information raises incumbent support when respondents learn crime is better than they had assumed and erodes it when they learn the opposite. At the good-news end of the plotted range (a gap of roughly 2,650 robberies fewer than the respondent expected), the mediated component is 0.035, with a 95% confidence interval of $[-0.0001,\ 0.075]$

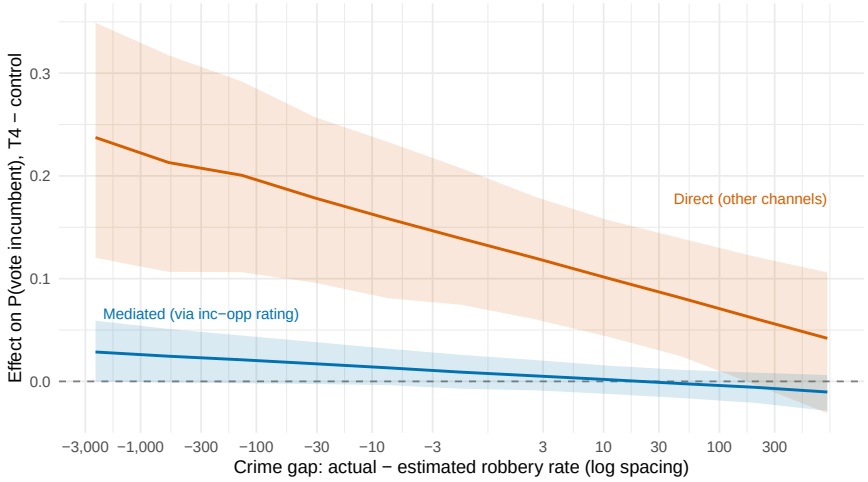

Figure 11: The T4 (same-coalition comparison) effect on the probability of voting for the incumbent coalition, relative to Control, decomposed into the portion transmitted through the incumbent-versus-other-coalitions rating ("Mediated") and the portion running through every other channel ("Direct"). Both are estimated across the range of the crime-level perception gap, with other variables held at their means. The horizontal axis is the signed log of the gap between the actual and the estimated robbery rate, labelled at nominal values. Estimates are quasi-Bayesian approximations (1,000 simulations) from pooled linear mediator and outcome models, each controlling for both pre-treatment ratings, the perception gaps interacted with arm, and pre-treatment coalition preference; bands are 95% confidence intervals.

that only just includes zero. At the bad-news end the mediated component falls to -0.01, and is not statistically distinguishable from zero with a 95% confidence interval of $[-0.03, 0.01]$. The direct component is larger at all points in the range.

The results provide suggestive evidence that same-coalition benchmarks do move beliefs about the incumbent relative to the alternatives, and those beliefs do carry some of the effect on vote intention in the expected direction. However, this particular mediation pathway is only a small part. Respondents may also be updating about the incumbent's general competence rather than on crime specifically. Additionally, note that the mediated component's confidence interval covers zero across nearly the whole range, so its magnitude is better described as small and imprecisely estimated than as reliably positive.

17 Wave 2 Survey Flow

The survey flow is illustrated in Figure 12, designed to mirror other survey experiments with informational treatments (Armantier et al. 2016; She 2024; Sawler 2025). Using a pre-generated random assignment vector stratified by the governing coalition of the respondent's home municipality (MORENA/PT/PVEM, PAN/PRI/PRD, or MC), respondents are assigned to the Control, T2, T3, and T4 arms with a 0.2 probability. The Control2 (Chart Placebo) and T1 arms are each assigned with a 0.1 probability. T1 is assigned a lower probability for two reasons. First, the most common convention for informational treatments is to provide a benchmark, so preference was given to improve power for comparison-treatments. Second, a similar study to this paper's design showed suggestive evidence that non-benchmarked information alone has similar or larger effect compared to benchmarked information (Arias et al. 2024). Control2 serves as a robustness check to determine if the act of comparison itself contributes significantly to the treatment effect.

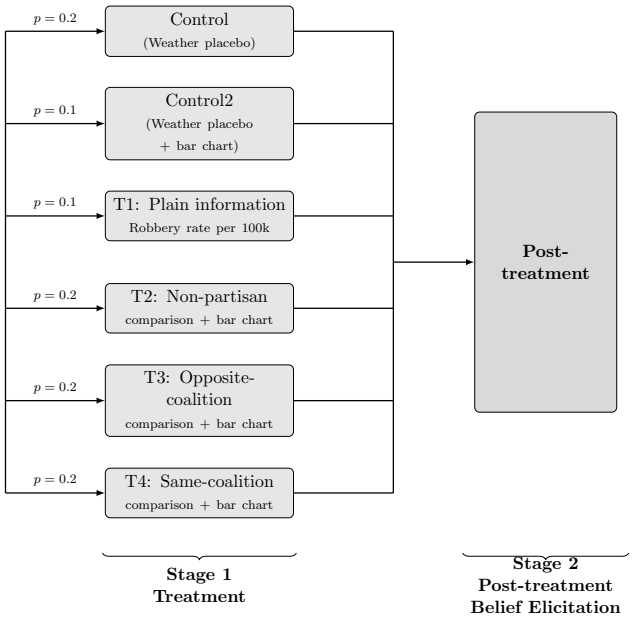

Figure 12: Treatment assignment in wave 2 of the survey

18 Survey Instrument

Questions are reproduced in English translation; respondents saw them in Spanish. Text filled in from respondent data appears in *[brackets]*, and notes on the response format appear in [gray brackets]. Page numbers refer to the screens of the survey application, and all questions listed under a given page were displayed together on that screen.

Wave 1

Consent (page 0)

The information sheet is displayed as an embedded PDF, with a link to an online copy for respondents whose device cannot render it. Respondents on small screens are advised to complete the survey in landscape orientation.

Please read this information sheet before continuing.
Button: "I understand, continue →"

Home municipality (page 1)

Q1. Enter your address to find your home municipality.
[Free-text entry, geocoded on submission. Only the resulting municipality name and INEGI code (CVEGEO) are stored; the address text is not. Respondents confirm the identified municipality, and a searchable dropdown of all municipalities is offered as a fallback. Respondents whose home municipality is not governed by one of the three main coalitions are screened out here.]

Practice ranking and attention check (page 3)

Please put this list of news sources in the following order: 1. Radio, 2. Print newspapers, 3. Social media.

[Drag-and-drop ranker with three items presented in randomized order and three numbered slots. All three items must be placed before advancing. Serves both to familiarize respondents with the interface used later and as an attention check; the submitted order is recorded.]

Benchmark municipality selection (page 4)

Q2. Municipal governments are often evaluated by comparing them to others. Which of the following municipalities would you compare *[home municipality]* to? (Choose at least one.)
[Checkbox list of 15 candidate municipalities, drawn as 5 sampled from the 10 geographically nearest municipalities, 5 sampled from the 20 largest municipalities nationally, and 5 sampled at random from the remainder, then shuffled. Both the candidate list and the selections are recorded.]

Q3. Add other comparison municipalities (optional).

[Type-ahead multi-select over all municipalities.]

Issue importance and vote intention (page 6)

Q4. How important are the following issues to you? Rank them from most to least important.
[Drag-and-drop ranker with five slots (1 = most important, 5 = least). Item order is randomized per respondent. All five must be placed before advancing.]

- Security / crime
- Economy / inflation
- Unemployment / low wages
- Corruption
- Education and health services

Q5. If you had to vote, which party or parties would you vote for in the 2027 municipal elections?
[Checkboxes with party logos. Multiple parties from the same coalition may be selected; selecting a party from a different coalition clears the earlier selections. Pre-treatment; repeated post-treatment in Wave 2.]

- PAN (Partido Acción Nacional)
- PRI (Partido Revolucionario Institucional)
- PRD (Partido de la Revolución Democrática)
- PVEM (Partido Verde Ecologista de México)
- PT (Partido del Trabajo)
- MC (Movimiento Ciudadano)
- MORENA
- Other [text field]

Robbery estimate and crime-handling ratings (page 9)

In 2025, ***half of all municipalities in Mexico had fewer than 79 robberies per 100,000 people, and the other half had more.***

Q6. How many robberies per 100,000 people do you think were reported in *[home municipality]* in 2025?
[Numeric entry, integer ≥ 0. Pre-treatment; repeated post-treatment in Wave 2. Supplies the pre-treatment component of CG_i.]

The following questions will ask you to evaluate how certain municipalities and parties "handle" non-violent crime such as robbery. "Handling crime" here refers to government efforts to prevent crime, enforce the law, and ensure public safety.

Q7. How well do you think the government of *[home municipality]* handles crime?
[Slider, 0–100, initialized at 50. Endpoints labeled "handles crime extremely poorly" and "handles crime extremely well." Pre-treatment; repeated post-treatment in Wave 2.]

Q8. On average, how well do you think the municipal governments of the following parties and coalitions handle crime?
[A single question stem followed by three sliders on one screen, all 0–100, all initialized at 50, under one shared pair of endpoint labels ("0 = handles crime extremely poorly," "100 = handles crime extremely well") printed once above the three. Coalition order is fixed. Pre-treatment; repeated post-treatment in Wave 2.]

- Sigamos Haciendo Historia (MORENA/PT/PVEM)
- Fuerza y Corazón por México (PAN/PRI/PRD)
- MC (Movimiento Ciudadano)

End of Wave 1. Respondents are re-contacted by Netquest to complete Wave 2.

Wave 2

Attention check, governance beliefs, and pre-treatment ranking (page 7)

Q9. We want to make sure you are reading each question carefully. To confirm that you are reading carefully, please select "Somewhat agree" below.
[Five-point agree–disagree scale. Correct response: "Somewhat agree."]

Q10. Which party or coalition do you believe currently governs each of the following municipalities?
[Radio-button grid, one selection per row. Rows are the home municipality and the four comparison municipalities the respondent will later be shown; the comparison set is determined by treatment arm and is selected by Mahalanobis distance on log population, log area, and log distance. Respondents in the Control and T1 arms, who see no comparison chart, are assigned one of the three comparison types at random for this grid.]

Municipality	MORENA/ PT/PVEM	PAN/ PRI/PRD	MC	Other	Do not know
[Home]	◯	◯	◯	◯	◯
[Comp. 1]	◯	◯	◯	◯	◯
[Comp. 2]	◯	◯	◯	◯	◯
[Comp. 3]	◯	◯	◯	◯	◯
[Comp. 4]	◯	◯	◯	◯	◯

Q11. Compared to ***[home municipality]***, how do you think the number of robberies in these municipalities compares?
[Radio-button grid, one selection per row, over the same four comparison municipalities. Pre-treatment; repeated post-treatment on page 11. Supplies RG_i.]

Municipality	More	About the same	Fewer
[Comp. 1]	◯	◯	◯
[Comp. 2]	◯	◯	◯
[Comp. 3]	◯	◯	◯
[Comp. 4]	◯	◯	◯

Treatment (page 9)

Respondents are assigned to one of six arms by block randomization from a pre-generated assignment vector stratified on the coalition governing the home municipality (Appendix 17); the Control, T2, T3, and T4 arms are assigned with probability 0.2 each and the Control2 and T1 arms with probability 0.1 each. A minimum reading timer is enforced before the "next" button becomes active. No responses are collected on this page.

Control (weather placebo)

Geography, altitude, and regional climate systems are the main determinants of rainfall patterns, factors that we largely cannot control.

In *[home municipality]* there were *[X]* mm of annual precipitation in 2025.

Source: WorldClim v2.1 (Fick & Hijmans, 2017).

Control2 (weather placebo with comparison chart)

[Control text, followed by a bar chart of annual precipitation in *[home municipality]* and four same-state comparison municipalities selected without regard to governing party.]

Source: WorldClim v2.1 (Fick & Hijmans, 2017).

T1 (plain information)

Municipal presidents control police funding and therefore have partial influence over local crime.

In *[home municipality]* there were *[X]* robberies per 100,000 people in 2025.

Source: Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP).

T2 (non-partisan comparison)

[T1 text, then:]

The following graph shows the robbery rate per 100,000 people in 2025 for *[home*

municipality] and a sample of similar municipalities.

[Bar chart: home municipality in blue, four same-state comparison municipalities in orange, selected without regard to governing party.]

Source: SESNSP.

T3 (opposite-coalition comparison)

[T1 text, then:]

The following graph shows the robbery rate per 100,000 people in 2025 for *[home municipality]* and a sample of similar municipalities governed by *[opposite coalition]*.

[Bar chart, with the governing coalition named for each comparison bar.]

Source: SESNSP.

T4 (same-coalition comparison)

[T1 text, then:]

The following graph shows the robbery rate per 100,000 people in 2025 for *[home municipality]* and a sample of similar municipalities governed by *[same coalition]*.

[Bar chart, with the governing coalition named for each comparison bar.]

Source: SESNSP.

Post-treatment robbery estimate and ranking (page 11)

In 2025, ***half of all municipalities in Mexico had fewer than 79 robberies per 100,000 people, and the other half had more.***

Q12. How many robberies per 100,000 people do you think were reported in *[home municipality]* in 2025?

[Numeric entry. Repeated from Wave 1. Used for the manipulation check in Appendix 5.]

Q13. Compared to ***[home municipality]***, how do you think the number of robberies in these municipalities compares?

[Same three-option grid and same comparison municipalities as page 7. Repeated post-treatment.]

Post-treatment ratings and vote intention (page 13)

The following questions will ask you to evaluate how certain municipalities and parties "handle" non-violent crime such as robbery. "Handling crime" here refers to government efforts to prevent crime, enforce the law, and ensure public safety.

Q14. How well do you think the government of *[home municipality]* handles crime?

[Slider, 0–100. Repeated from Wave 1.]

Q15. On average, how well do you think the municipal governments of the following parties and coalitions handle crime?

[As in Wave 1: one stem, three sliders on the same screen, one shared pair of endpoint labels, fixed coalition order. These three ratings, together with the home-municipality rating immediately above them, are the components of the belief outcomes reported in the main text. Repeated from Wave 1.]

- Sigamos Haciendo Historia (MORENA/PT/PVEM)
- Fuerza y Corazón por México (PAN/PRI/PRD)
- MC (Movimiento Ciudadano)

Q16. If you had to vote, which party or parties would you vote for in the 2027 municipal elections?
[Same checkbox format as Wave 1. Repeated from Wave 1. Primary outcome.]

Respondents submit the survey and are shown a thank-you screen.

References

Estimaciones NSE 2022. 2022.

Arias, Eric, Horacio Larreguy, John Marshall, and Pablo Querubin. 2024. "Does the Content and Mode of Delivery of Information Matter for Electoral Accountability? Evidence from a Field Experiment in Mexico" [in en]. *Latin American Economic Review* 34:1–41. https://doi.org/10.60758/laer.v34i.335.

Armantier, Olivier, Scott Nelson, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar. 2016. "The Price Is Right: Updating Inflation Expectations in a Randomized Price Information Experiment." *The Review of Economics and Statistics* 98, no. 3 (July): 503–523. https://doi.org/10.1162/REST_a_00499.

Arnold, Benjamin F., Daniel R. Hogan, John M. Colford, and Alan E. Hubbard. 2011. "Simulation methods to estimate design power: an overview for applied research" [in en]. *BMC Medical Research Methodology* 11, no. 1 (June): 94. https://doi.org/10.1186/1471-2288-11-94.

Hartman, Erin, and F. Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests" [in en]. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12387, *American Journal of Political Science* 62 (4): 1000–1013. https://doi.org/10.1111/ajps.12387.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects" [in en]. *Statistical Science* 25, no. 1 (February): 51–71. https://doi.org/10.1214/10-STS321.

Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design" [in en]. *American Political Science Review* 114, no. 3 (August): 761–774. https://doi.org/10.1017/S000305542000012X.

Lakens, Daniël. 2017. "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses" [in EN]. *Social Psychological and Personality Science* 8, no. 4 (May): 355–362. https://doi.org/10.1177/1948550617697177.

Sawler, Dot. 2025. "Changing Partisan Minds, Not Hearts: Information Changes Beliefs, But Not Affective Polarization." *Working Paper* (December).

She, Hongyi. 2024. *Learning About Trade* [in en]. SSRN Scholarly Paper. Rochester, NY, April. https://doi.org/10.2139/ssrn.4803318.