Information Architecture for Corruption Messaging: Evidence from an Adaptive Experiment

Felipe Torres-Raposo[1] Raymond Duch[2]

[1]London School of Economics and Political Science
[2]Nuffield College, University of Oxford

September 5, 2025

The presentation and framing of information are central to many information experiments. In the context of corruption, policymakers and NGOs face the challenge of effectively informing citizens about malfeasance in their local governments. This task involves identifying messaging strategies that are sufficiently compelling to capture public attention. To address this, we conducted three online experiments: (1) a repeated-measures factorial experiment, (2) an adaptive experiment that sequentially allocates subjects based on the relative performance of each treatment arm, and (3) an AI-enhanced adaptive experiment. Across these studies, we evaluated various information strategies commonly employed in corruption-focused research. In Studies 1 and 2, we found no single messaging strategy that consistently outperformed others. However, loss-framed messages tended to be slightly more persuasive. We also observed that simpler metrics of corruption were often as persuasive as more sophisticated ones. Finally, we found no substantial differences between spatial comparisons (across municipalities) and temporal comparisons (over time) in terms of persuasive impact.

1 Motivation

There is extensive evidence suggests that regardless of national context, respondents express negative attitudes towards corruption, particularly in the public sector (Lagunes and Seim, 2021). There is also some consensus among scholars that corruption is a salient issue for the average individual (Polk et al., 2017; Engler, 2020). Thus, it would be reasonable to expect that citizens would pay attention and care about news or information about corruption that may affect their private or social welfare (Harm Rienks, 2023).

Early research has pointed out that information about malfeasance generated from government audit reports affects the political preferences of the average voter. Ferraz and Finan (2011) find that media coverage of corruption audits reduces support for the incumbent. Similarly, de Figueiredo, Hidalgo and Kasahara (2012) report finding mixed results in changing voters' political preferences when they are informed about the incumbent's performance more generally (Pande, 2011; Pande et al., 2011). More recent research has found null the effects of malfeasance information on voting behavior (Arias et al., 2022; Chong et al., 2015; Boas, Hidalgo and Melo, 2019). This accrued evidence liking information with accountability raises several questions about the information chain through which citizens learned about their elected officials' wrongdoing.

More recent research has leveraged the information produced from audits to causally identify the effects of corrupt information on voting behavior and belief updating using large-scale field experiments. A closely related study is Arias et al. (2022), where the authors examined how voters updated their beliefs once informed of malfeasance in mayoral spending. The authors compiled data from audits conducted by the Federal Auditor's Office in Mexico, which reports the share of spending of federal funds diverted into unauthorized projects, as well as the share of these funds for which the municipality is unable to provide accounts. Similarly, Chong et al. (2015) evaluate the impact of audit-based information, using a similar treatment to Arias et al. (2018), and find that corruption information did not shift citizens' beliefs about corruption levels or support for the incumbent, but it did depress turnout.[1] Boas, Hidalgo and Melo (2019) also conducted an experiment using data from audits conducted by the *State Accounts Courts of Pernambuco* in Brazil to study significant disparities between voting behavior obtained from

[1]Similar studies to Arias et al. (2022) were conducted in Mexico by Enrique et al. (2019) and Larreguy, Marshall and Snyder (2020).

survey and field experiments. [2]

As outlined in the previous literature, most experimental work on corruption relies on messaging strategies that convey quantitative information about malfeasance. However, it remains challenging to determine whether null effects are, in part, artifacts of how this information is assembled and presented to voters.[3] This puzzle raises several important questions: What types of messages are most likely to influence public opinion? Which components of corruption-related content are most persuasive or accessible to citizens? Addressing these questions requires identifying and systematically testing a range of possible message candidates. This paper contributes to this effort by thoroughly evaluating several information strategies commonly used in the literature while also introducing previously unexplored narrative approaches to malfeasance messaging.

This paper investigates which corruption-related information is most compelling to the average citizen to the average citizen drawing on evidence from three experimental designs: 1) a repeated-measures factorial design, 2) An adaptive experiment, and 3) a conventional experiment augmented with AI-powered qualitative interviews. Rather than focusing on political outcomes, we assess respondents' evaluations of various messaging strategies along dimensions such as clarity, persuasiveness, and trustworthiness. We complemented these evaluation measures with a behavioral outcome: respondents' willingness to seek information about corruption in their local constituency. We further examine participants' attentiveness and comprehension to evaluate the extent to which each message format facilitates understanding and engagement.

Our findings suggest that information strategies that highlight the foregone social benefits lost due to corruption are slightly more persuasive than those focusing on its economic costs. Additionally, when comparing different benchmarking strategies, we find that tracking malfeasance over time within a municipality is somewhat more effective than spatial comparisons, where malfeasance is compared across municipalities within the same region. We do not observe significant differences across treatments in terms of their impact on information-seeking behavior. However, treatments that perform poorly in encouraging information-seeking behavior tend to be positively evaluated on their persuasiveness measures.

[2] Jablonski et al (2021) also compiled information from audits carried out by enumerators on primary schools, roads, health care, and water access quality in Ugandan villages. The author created indices of "service quality" for each village and informed voters whether their village performed better or worse than other villages within the same district. The authors find no effects of their treatments on changing citizens' beliefs about the incumbent's integrity and effort, as well as voting outcomes.

[3] Incerti (2019) provides a comprehensive analysis of why information treatments on corruption may yield null effects.

This paper contributes to the literature that links malfeasance information and corruption by providing insights into what features of malfeasance information are relevant/informative to individuals. It also contributes to the research within the field of behavioral science on the effects of framing. This research has broadly studied whether conveying the same information in a gain/loss frame significantly affects individuals' behavior of interest. In this study, we give empirical evidence to demonstrate whether individuals find more compelling information that incorporates several frames versus standard "default" information messaging strategies.

The second contribution of this paper is to the nascent literature that implements adaptive experiments within the fields of political science and economics. This study introduces and incorporates the concepts of dynamic stochastic optimization and reinforcement learning, predominantly used within the computer science discipline, into an online experiment aimed at identifying the best-performing information strategies while exploring underperforming ones.

This paper continues as follows. Section 2 summarizes the different information strategies used in corruption surveys and field experiments. Section 3 explains the experimental approaches implemented in this study and the selection of outcomes. Section 4 provides more details about estimation methods and adjustments required to draw inferences from adaptive experiments. Section 5 summarizes the study's empirical findings using several estimation strategies. Finally, in Section 6, we summarized this study's main findings and limitations.

2 Information Treatments

Most of corruption information experiments extract critical information from audit reports or close facsimiles to generate their information treatments. These dossiers, for example, usually contain information about the number of irregularities, the severity of the irregularities, or the amount of malfeasance found. Table 1 synthesizes the message content or format of the malfeasance information treatments used in these experiments. We identify ten specific dimensions organized into two general categories: metrics and benchmarking. As Table 1 suggests, these studies have little consistency regarding messaging strategies, content, and metrics.

Table 1: Summary information treatments

		Metrics reported							Benchmark		
Study	Medium	Approve/rejection accounts	Number irregularities	Convitions/ Criminal charges	Malfeasance percentage	Total Malfeasance	Gains or foregone losses	Rent-seeking or Accepting bribes	Spatial	Temporal	Challenger
Banerjee, Hanna and Mullainathan (2012)	Report cards			x					x		
Buntaine and Daniels (2020)	SMS					x			x		
Enrique et al. (2019)	Videos				x	x	x				
KB thesis	Mainstream media/ Political campaigns		x		x						
Botero et al. (2015)	Telephone			x				x			x
Weitz-Shapiro and Winters (2017)				x							
Ferraz and Finan (2011)	Mainstream media/ Political campaings		x			x					
Ferraz and Finan (2008)	Mainstream media/ Political campaigns		x			x					
Arias et al. (2022)	Leaflets				x	x			x		
Arias et al. (2022)	Leaflets				x	x	x		x		
Boas, Hidalgo and Melo (2019)	Fliers/Enumeratos	x			x				x		
Boas, Hidalgo and Melo (2019)- Survey		x									
Figueroa et al. (2012)	Leaftlets			x							
Breitenstein (2019)				x				x			x
Chong et al. (2015)	Leaflet					x					
Winters and Weitz-Shapiro (2013)				x				x			
Winters & Weitz-Shapiro 2016								x			
Winters & Weitz-Shapiro 2017								x			
Winters and Weitz-Shapiro (2020)								x			
Avenburg (2019)		x					x	x			x
Klašnja, Lupu and Tucker (2021)								x			
Klašnja, Lupu and Tucker (2020)								x			x
Franchino and Zucchini (2015)				x							x
Agerberg (2020)	Leaflets										x
Mares (2003)				x				x			x
Eggers, Vivyan and Wagner (2018)								x			x
Vera (2020)								x			
Bobonis, Cámara Fuertes and Schwabe (2016)			x					x			
Humphreys and Jeremy (2012)								x			
Larreguy, Marshall and Snyder (2020)	Radio/Television				x						

An essential piece of information that these government audit reports generate is the financial losses associated with corruption. Here is where the literature on "choice architecture" and financial decision-making could be relevant and help explain how citizens would respond to losses related to wrongdoing (Thaler, 2016). This literature can provide valuable insights into how the average citizen draws conclusions about their governments' levels of corruption. This literature also underlines how framing this financial information can promote welfare-enhancing decisions.

Recent contributions from "choice architecture" literature suggest that enhancing the presentation of information associated with financial choices improves consumer welfare. The U.S. CARD Act of 2009 (Soll, Keeney and Larrick, 2013) is a classic illustration of this and, more recently, Carpenter et al. (2020) have demonstrated that an enhanced information presentation design for pre-paid cards can generate substantial consumer benefits. Similar welfare gains can be found in information provided to individuals about savings and retirements option. (Cronqvist, Thaler and Yu, 2018) has found that choice architecture has a long-term impact on pension savings. (Liebman and Luttmer, 2015) identifies that even relatively minor information enhancements can significantly impact planning for Social Security. As Benartzi and Thaler (2007) note, there are no neutral choice architecture designs, and even small details could matter in the design of saving vehicles.

Average voters, like the average consumer, are expected to digest malfeasance information in their government and use this to inform their vote choice. Thus, testing different contents and frames of malfeasance information would help us tease apart which dimensions are more welfare-enhancing for the average citizen. From what we have gathered from the literature (see Table 1), the information about malfeasance can be broken into two dimensions: corruption metrics and how these metrics are bench-marked.

Metrics Several studies have used different metrics to inform citizens about elected officials' corrupt behavior. The table 1 reports seven distinct measures of malfeasance. For example, Boas, Hidalgo and Melo (2019) in Mexico provided a simple metric on whether the audit agency *approves or rejects* the municipal accounts. Similarly, Pande et al. (2011) used a newspaper report card showing election candidates' criminal convictions. These studies illustrate the wide range of metrics used in corruption information experiments.

A second group of metrics informs citizens of the monetary costs associated with irregular

activities. A case in point is the Buntaine and Daniels (2020) Uganda experiment, in which the authors include the total cost of malfeasance as one of their treatments. Likewise, Arias et al. and Buntaine and Daniels express the cost of malfeasance as a percentage of the municipal budget. Thus, this metric is arguably much richer than the former, as it provides information about the magnitude of corruption.

A third metric is to express the cost of malfeasance as foregone expenditures on a particular public good. For example, Arias et al. (2019) 's treatment expresses the magnitude of malfeasance in terms of foregone spending on social programs. Similarly, Larreguy, Marshall and Snyder (2020)'s experiment reports the amount of funds spent on unauthorized projects. All three metrics convey different dimensions and features of wrongdoing.

Benchmarks. Table 1 also highlights field and survey experiments that incorporate a benchmark as part of their information strategy. The rationale is that voters are better able to assess whether malfeasance in their local government is high or low when provided with a relevant reference point. These benchmarks signal to individuals the extent to which government institutions or elected officials deviate either (1) from other municipalities (spatial benchmarks) or (2) from their own levels of malfeasance in previous years (temporal benchmarks). These benchmarking strategies build on a substantial body of theoretical and empirical literature that underscores the role of retrospection and comparative evaluations in shaping vote choice (Ferejohn, 1986; Fiorina, 2006; Besley, 2006; Besley and Case, 1995; Healy and Malhotra, 2013).

Recent efforts to assess the impact of malfeasance messaging on political accountability have increasingly emphasized the role of benchmarking. The studies presented in Table 1 adopt different benchmarking strategies. Lierl and Holmlund (2019) employ a temporal frame by comparing current performance to that of the previous municipal government and a spatial frame by benchmarking the subject's municipality against a national average. However, neither frame significantly affected their outcome variables: voter turnout and voter preferences. In contrast, Bhandari, Larreguy and Marshall (2019) find that temporal benchmarking (comparing an incumbent's performance to that of previously elected officials in the same district) leads to belief updating relative to their spatial benchmark.

Benchmarking can also be connected to the literature on reference points. At the core of this theory is the notion that individuals evaluate outcomes and form preferences relative to a reference point. Prospect theory, as developed by Kahneman and Tversky (1979), posits that

the utility of an outcome is driven more by changes from this reference point than by the outcome's absolute value. Scholars have proposed several candidates as potential reference points, including the status quo (Kahneman, 1992), individual expectations (Augenblick and Rabin, 2021), and social comparisons both within and across groups (Hargreaves Heap, Ramalingam and Rojo Arjona, 2017). As such, it is reasonable to expect that information treatments incorporating benchmarks may shift the salience of specific reference points, thereby influencing how individuals interpret and respond to corruption-related messages.

In our experiment, all information treatments were designed using historical data from audits undertaken by the Comptroller General Office of Chile (CGO). This data contains over 900 audit reports undertaken in the last six years. These audit reports provide detailed information on malfeasance identified in local governments, including the extent of malfeasance and the number of irregularities discovered. The purpose of using this data is to convey information about malfeasance that citizens would likely be exposed to in conventional media outlets or social media platforms.

3 Experimental Designs

To evaluate what frames and content of information metrics are more compelling, we conducted three online experiments that have three critical components: 1) The medium to deliver the information treatment is short 50-second videos.[4] 2) A repeated-measures factorial and an adaptive sampling strategy that sequentially modifies treatment assignment shares based on the performance of each arm. 4) An AI interview. 5) A behavioral information-seeking outcome that seeks to differentiate which frames and pieces of content about corruption are evaluated to be more compelling.

Repeated-Measures Factorial Design - Study 1 In this repeated measures factorial experiment, each respondent evaluated 6 randomly selected videos out of twenty-four information treatments. We collected 3,996 observations, which translate, on average, 166 evaluations per

[4] An increasing number of online and field experiments have used videos as the medium to deliver their information interventions. For example, Dunning et al. (2019) filmed interviews with 100 parliamentarian candidates in 265 villages in Uganda. These videos contained information such as the candidates' names, party affiliations, and policy priorities. Similarly, Bidwell, Casey and Glennerster (2020) randomly screened video debates across pooling stations, aiming to increase political knowledge. The evidence for choosing this medium is twofold: Firstly, from a previous survey, Chileans reported receiving political information through platforms such as WhatsApp, and they primarily received this information in video format. Secondly, Wittenberg et al. (2021) find that videos are modestly more persuasive than written material when it comes to communicating political issues to citizens.

video.[5]

The 24 different treatment combinations can be derived from Table 2 summarizes the three factors and their levels (2 x 6 x 2 = 24) in this factorial design. The first factor is *benchmarks*, which has two levels. In the *spatial* level, respondents will learn the amount of malfeasance of the hypothetical municipality compared to other regional municipalities. In contrast, the *temporal* level will compare the current amount of malfeasance (in Chilean pesos) versus the amount reported in the previous audit. Screenshots of these information treatments are shown in Figures 15 and 17.

The second factor, *metrics*, has three levels: The first is *Individual*, which expresses the cost of malfeasance as a share of every $1,000 Chilean pesos of the municipality's spending budget. The second metric is *Foregone Social Benefits*, which expresses the number of influenza vaccines that could have been bought with the total malfeasance amount. The third level is *Standard*, which reports the number of irregularities. One important point to highlight is that all treatments report the exact quantities; as a result, the only feature that changes across treatment conditions is the presentation and framing of the information.

The third dimension, Levels, has two levels: a *High* or a *Low* malfeasance. In the *High* condition, respondents were informed of high levels of corruption either in terms of resources, the number of irregularities, or their severity relative to the spatial or temporal frame. In the *Low* condition, respondents in the spatial frame received information that the municipality is within the lowest levels of the distribution of malfeasance in its region. In contrast, within the temporal frame, respondents received information indicating that the municipality has shown a small reduction in malfeasance compared to the previous year.

[5]Table 25 in the Appendix provides details of the sample per arm.

Table 2: Description of treatment factors - Study 1

	Factor: Bench-marking
Spatial	Comparison of total amount of malfeasance across municipalities in the same region.
Temporal	Comparison of total amount of malfeasance across time of the same municipality
	Factor: Metrics
Standard	Total number of irregularities
Severity	Severity of the irregularities
Resources	Total amount of malfeasance expressed in Chilean pesos
Foregone	Equivalent number of resources loss in terms of influenza vaccines
Programme	Report the total amount of malfeasance and the programmes affected
Individual	Share of $1,000 of the municipality's budget loss due to malfeasance
	Factor: Levels
High	Temporal: Substantial increase in the amount of malfeasance Spatial: Top-ranked across the region
Low	Temporal: Minor increase in the amount of malfeasance Spatial: Low-ranked across the region

Adaptive Experiment - Study 2 A common limitation in these types of studies is that researchers/policy-makers are rarely able to test the full range of dimensions they are interested in. As a result, they face a trade-off between exploring a broad set of factors versus producing well-powered, precise estimates for each treatment arm. In this scenario, adaptive experiments may ease this search and efficiently discern the best-performing treatment arm(s).[6] In this design, treatment assignment shares are updated based on observed outcomes rather than allocating experimental units with a fixed probability. It is in this continuous cycle of updating treatment assignment probabilities based on interim results that adaptive experiments can efficiently identify the most promising treatment arm(s) (Scott, 2010).[7]

Adaptive experiments rely on algorithms that balance the trade-off between exploitation and exploration, conditional on the goals of the researcher or policymaker.[8] These goals may range

[6]There has been a relatively small but increasing number of adaptive examples within economics and political science. Bahety et al. (2021) conducted a ten-batch experiment sending SMS texts aimed at encouraging compliance with social distancing and hand-washing guidelines in India during COVID-19. Their experimental design is very similar to this study, comprising 10 information arms divided into 10 waves. Esposito and Sautmann (2022) conducted a two-wave adaptive experiment testing six forms of automated calls in Kenya to encourage parents to read to their children. Caria et al. (2020) carried out an adaptive experiment to identify welfare-maximizing employment policies for Syrian refugees and local job-seekers.

[7]The recent literature on adaptive experiments draws heavily from *Reinforcement learning*, specifically addressing the exploration-exploitation trade-off, also known as the "multi-armed bandit problem." Researchers test multiple interventions but sequentially observe noisy signals about their quality (Russo, 2020). Various algorithms balance exploration and exploitation of policies.

[8]Common algorithms include equal allocation, greedy, softmax learning, upper confidence bound (UCB),

from maximizing the precision of treatment effect estimates and minimizing regret to maximizing exploration in settings where there are no strong priors about the quality of competing treatments.[9]

The *Thompson Sampling for the Bernoulli Bandit* algorithm (Thompson, 1933) stands out for balancing exploration with exploitation of the best-performing arm(s). This algorithm's heuristic explores all treatment arms to gather more information about their quality. However, as soon as it gains sufficient knowledge of which arm is the best, it assigns more units to that arm (Scott, 2010). The number of experimental units assigned to each treatment is based on the posterior probability of that arm being the best intervention. Table 1 outlines the Thompson Sampling Algorithm for batch experiments.

Based on simulations and pilot results reported in the Appendix, we implemented a modified version of the Thompson sampling algorithm proposed by Kasy and Sautmann (Kasy and Sautmann, 2021), known as *Exploration Sampling*. This algorithm effectively balances exploration/exploitation by gradually switching sampling away from those treatment arms that are unlikely to be optimal.[10] Kasy and Sautmann demonstrate that their proposed algorithm converges at the same rate as their posterior probability. When a treatment converges more quickly, the algorithm reduces or halts further assignment to that arm, allowing other, less-explored arms to "catch up." This mechanism increases expected welfare by preventing more than 50 percent of experimental units from being assigned to the seemingly best-performing treatment too early. Furthermore, this algorithm avoids prematurely halting exploration of potentially sub-optimal arms, thus preserving learning opportunities. In their framework, $q_{k,t}$ denotes the share respondents assigned to each treatment arm k, where equations 3 and 4 detailed $q_{k,t}$ its computation.

Gittins index, and top-two Thompson sampling.

[9] In computer science, regret refers to the difference between the expected reward of the best arm and the reward obtained through the experimental assignments.

[10] Thompson sampling substantially increases the probability of making a Type-1 error by roughly 10%.

Table 3: Algorithm Batch-wise Thompson Sampling

Algorithm 1: Batch-wise Thompson sampling
1: Iniatize priors such that $(\alpha_{k,1} = 1,\ \beta_{k,1})$ for $k = 1, ..., K$
2: Calculate $p_{k,t} = P[\Theta_k = \text{argmax}\{\Theta_1, ...\Theta_K\}\|(\alpha_{1,t}, \beta_{1,t}), ..., (\alpha_{K,t}\beta_{K,t})]$ for $k = 1...K*$
3: Sample n observations, assigning treatment with probabilities $(p_{1,t}, ...p_{K,t})$.
4: Update posteriors, for $k = 1, ..., K$: $\alpha_{k,t+1} = \alpha_{k,t} + \#$success observed for arm k in period t $\beta_{k,t+1} = \beta_{k,t} + \#$success observed for arm k in period t

$$q_{k,t} = S_t \cdot p_{k,t}(1 - p_{k,t}) \tag{1}$$

$$p_{k,t} = P_t\left(k = \arg\max_{k'} \theta^{k'}\right) \text{ and } S_t = \frac{1}{\sum_k p_{k,t} \cdot (1 - p_{k,t})} \tag{2}$$

$$q_{k,t} = S_t \cdot p_{k,t}(1 - p_{k,t}) \tag{3}$$

$$p_{k,t} = P_t\left(k = \arg\max_{k'} \theta^{k'}\right) \text{ and } S_t = \frac{1}{\sum_k p_{k,t} \cdot (1 - p_{k,t})} \tag{4}$$

The adaptive experiment was conducted in eleven batches. The first batch comprised 200 respondents, while each of the remaining 10 batches included 100 respondents, resulting in a total sample size of 1,200.[11] We increased the size of the first batch based on the simulations reported in Figure 11 and the results from the pilot conducted before the study. Having a larger first batch size can contribute to yielding more accurate estimates of the quality of each arm from the outset. However, this design choice involves a trade-off: treatments that begin to perform optimally in later batches may receive insufficient exposure due to earlier allocation decisions. Additional details of the first batch size and RMSE are reported in Figures 12

[11] We are considering running a final batch with 400 respondents.

to 14 in the Appendix. Furthermore, pilot results indicated that poor-performing treatment arms in the first batch were severely penalized in terms of treatment assignment shares for the second batch, resulting in treatment assignment probabilities falling below 1%. Table 18 in the Appendix compares treatment assignment shares for first batch sizes of 100 and 200 respondents, based on this pilot study.

For the first batch, we set a uniform prior, treating all treatment arms as equally likely to be an optimal arm. This means that treatment assignment shares for each arm were the same (one-sixth each). After collecting data from each batch, we computed the posterior probabilities that each arm was the best. We used Monte Carlo sampling to approximate the posterior probability values $p_{\alpha_k,t+1}$. These posterior estimates were then used to determine the exploration sampling shares, $q_{k,t}$, for the subsequent batch.[12]

In this second study, we focused on the top 6 top-performing treatments identified in the repeated-factorial design. Table 4 summarizes the content and format of the malfeasance messages used in the experiment. The experimental design incorporates two key dimensions: benchmarks and metrics. The first dimension, *benchmarks*, includes two levels. In the *spatial* condition, respondents were shown the amount of malfeasance in a hypothetical municipality relative to other municipalities in the same region. Whereas in the *temporal* condition, respondents observed the current amount of malfeasance from the most recent audit compared to the amount reported in a previous audit. Screenshots of these information treatments are displayed in Figures 15 and 17.

The second dimension, *metrics*, includes three levels. The first, *Individual*, presents the cost of malfeasance as a proportion for every $1,000 Chilean pesos the municipality spends. The second metric, *Foregone Loses*, frames the cost in terms of the number of influenza vaccines that could have been purchased with the total amount of malfeasance. The third level, *Standard*, reports the number of identified irregularities in the audit process. All treatments report the exact quantities; thus, the only variation across treatment conditions lies in the presentation and framing of the information.

[12]The model fitted is a Bayesian Bernoulli model with uninformative beta priors. However, for more sophisticated experimental settings, it would be possible to use hierarchical Bayesian models to capture the heterogeneity within different strata or groups

Table 4: Description of Treatment Factors - Study 2

	Factor: Frame
Spatial	Comparison of the total amount of malfeasance across municipalities in the same region.
Temporal	Comparison of the total amount of malfeasance across time of the same municipality
	Factor: Metrics
Individual	Share of $1,000 of the municipality's budget loss due to malfeasance
Foregone	Equivalent number of resources loss in terms of influenza vaccines
Standard	Number of irregularities found in the last audit

The treatments were delivered through a 50-second video that presented the information in a fixed sequence: 1) For the *Individual* and *Foregone* treatments, participants were first informed that corruption occurred in a municipality, followed by the total amount of malfeasance expressed in Chilean pesos. For the *Standard* treatments, it also communicated the presence of malfeasance but then reported the total number of irregularities found in the hypothetical municipality. In the second part of the video, respondents received the benchmark component of the treatment. Then, at the end of the video, respondents received the *Metrics* component of the treatment.

AI-Chatbot Enhanced Adaptive Experiment - Study 3 (Pilot) In Study 3, we built on the six treatment conditions in Study 2 by further examining the role of benchmarks in persuading respondents and facilitating learning. Table 5 summarizes the factors and levels included in the study. For the benchmark factor, we introduced two additional levels: 1) *National*, which compares the hypothetical municipality's levels of malfeasance to the national average, and 2) *Ranking*, which compares malfeasance levels across the region with municipalities ranked by their corruption levels. In this frame, the hypothetical municipality was assigned the highest rank for malfeasance. For the temporal benchmark, we added two additional levels: 1) A comparison of the total amount of malfeasance against the average amount of malfeasance found in previous audits since 2016, and 2) a comparison of the current malfeasance level to the results over the same period.

After watching the video and respondents answering a set of assessment questions, they participated in an AI-assisted qualitative interview to gain more insights into their evaluations.

The respondents interacted with the AI interviewer through a chat interface that resembles popular text messaging applications on modern smartphones. Respondents in our pilot consistently displayed high levels of engagement; on average, they gave 6-7 responses and wrote 50 words per interview.

Respondents in this experiment are expected to be assigned adaptively using the Exploration Sampling algorithm introduced in Study 2. However, instead of relying on a binary behavioral outcome, we will use the average score from the evaluation questions as the response variable. We plan to split the sample into ten batches of 400 respondents each (total n = 4,000).

Table 5: Description of Treatment - Study 3

	Benchmark: Spatial across municipalities
National	Comparison of the total amount of malfeasance to the national average
Regional	Comparison of the total amount of malfeasance to the regional average.
Ranking	Comparison of the total amount of malfeasance ranked across municipalities within the region
	Benchmark: Temporal - within same municipality
Previous	Comparison of the total amount of malfeasance with respect to previous audit
Average (since 2016)	Comparison of the total amount of malfeasance with respect to average malfeasance since 2016
Ranking (since 2016)	Comparison of the total amount of malfeasance ranked previous audits since 2016

Evaluation and behavioral outcomes In all three studies, we aimed to evaluate which frames and metrics about malfeasance are more compelling to the average citizen. Our initial approach involved asking respondents a brief series of questions immediately after viewing the videos. Participants indicated their level of agreement regarding whether the video treatments were (1) convincing, (2) trustworthy, (3) reliable, and (4) precise. These evaluation outcomes have been utilized within the field of political communication as a means of evaluating the persuasiveness of various messaging strategies. O'Keefe (2021) points out that one approach to measuring persuasiveness is by using a survey through which individuals are asked to rate different messaging formats on their *perceived message effectiveness*. These questions are usually formatted using Likert scales with end-anchors such as *persuasive/not persuasive* or *convincing*

or not convincing.

Behavioral outcome Our behavioral outcome measures whether respondents sought additional information about malfeasance in their local government. Prior studies have frequently used information-seeking behavior as a proxy for the persuasiveness of specific informational treatments. The underlying rationale is that sufficiently compelling messages are likely to prompt individuals to engage more deeply with the political content, including seeking further information.

In this study, we captured this behavioral response using a binary outcome: at the end of the survey, respondents were asked whether they wanted to conclude the questionnaire or learn more about irregularities uncovered in their local government. If respondents were interested in learning about the amount of malfeasance found in their local council, they had to click a link at the end of the survey. This design replicates real-world scenarios in which citizens decide whether to explore news stories or official reports on corruption. Figure 20 in the Appendix shows how we implemented this into the survey. Figure 21, also in the Appendix, displays screenshots of the information shown to respondents once they decided to obtain more information about corruption.

Similar experimental studies have used information-seeking behavior as a proxy of persuasiveness. Singh and Roy (2018) conducted an online survey experiment to examine how respondents modified their information-gathering behavior after being informed of results from opinion polls. In their study, respondents were allowed to browse various information boards displaying the parties' economic and environmental positions. The author measured the level of engagement by assessing whether respondents clicked on these websites and the amount of time they spent browsing. Likewise, Ryan (2012) conducted an online field experiment to evaluate which emotions spurred information gathering among citizens on several political issues. The author used Facebook ads' CTRs to different political advertisement ads to measure engagement.[13]

Other research areas within political science have similarly employed CTRs on websites or social media posts as their primary behavioral measure. Vosoughi, Roy and Aral (2018) analyzed the differential diffusion of verified information versus false news on Twitter. They find that social media users tend to click and share false tweets more than verified information. Research

[13]Within the literature of digital marketing, it has also used CTRs on Facebook ads as their primary outcome measure. Matz et al. (2017) conducted three field experiments that tested multiple persuasion appeals and found that tailored ads that incorporated individuals' psychological traits yielded significantly higher CTRs

on political participation and engagement has also extended beyond conventional outcomes, such as voting and partisanship, incorporating more "passive" or "soft" forms of political participation, such as signing petitions or subscribing to newsletters (Porten-Cheé et al., 2021).

Selecting an appropriate outcome is an essential step in adaptive experiments. as treatment assignment probabilities depend on the success rates of each treatment arm. In this experiment, treatment assignment shares were a function of the treatment's performance only for the behavioral outcome. Thus, the messaging strategies with higher information-gathering rates yield higher treatment assignment probabilities.[14]

Regarding the items included in these studies to examine treatment' arms perceived persuasiveness, Thomas, Masthoff and Oren (2019) found that several studies include questions that assess *whether the content is convincing*, *important to me* or *believable*. As pointed out before, we included four items with two rating scales, i.e., whether they *Agree* or *Disagree* on the statements listed before. While this binary scale facilitates respondents' assessment of the information treatments, it does reduce the precision with which these assessments are captured.[15] In Study 1, in addition to behavioral and perceived measures, we included two stated preference questions: (1) whether respondents would be willing to share the video with friends, family, or colleagues, and (2) whether they would be willing to post the video on any social media platform they use.

Following these questions, respondents were asked several attention and comprehension check questions. The first assessed whether respondents recalled the main topic of the video, while the second and third evaluated whether individuals remembered the benchmarking and metric components, respectively. Table 15 in the Appendix summarizes the proportion of participants who answered all three questions correctly. Table 6 provides an overview of the outcomes, attention check questions, and their scales for all three studies.

[14] A more suitable approach to measure persuasiveness may be using a composite index that combines behavioral and evaluation items. However, most advances within the literature on adaptive experiments prioritize binary outcomes (Hadad et al., 2021; Zhang, Janson and Murphy, 2020; Offer-Westort, Coppock and Green, 2021; Jack and Lorenzo, 2017; Deshpande et al., 2017) A limitation of this approach is that it may bind existing adaptive studies to use these type of outcomes.

[15] Most of the studies that the authors find either use a five-point or a seven-point Likert scale

Table 6: Summary of outcomes and attention checks

Outcomes	Study 1	Study 2	Study 3
Evaluation		Binary	Binary
Behavioral	-	Binary	5-point likert scale
Attention/Comprehension			
Topic	Yes	Yes	Yes
Content	Yes	Yes	Yes
Benchmark type			Yes
Algorithm Variable			
Evaluation	-	-	Index (Continuous)
Behavioral	-	Binary	-

4 Treatment Effects Estimation and Inference

Estimating treatment effects from adaptively collected data requires appropriate adjustments to obtain unbiased estimates (Shin, Ramdas and Rinaldo, 2019; Rafferty, Ying and Williams, 2019). This need arises as treatment assignment probabilities are dynamically updated based on observed outcomes (Nie et al., 2018; Pallmann et al., 2018). Moreover, adaptively collected data can result in non-normal and heavy-tailed sampling distributions for both treatment and control groups, potentially leading to incorrect inference (Hadad et al., 2021).

One advantage of using *Exploration sampling* is that the sampling bias generated from adaptively collected data is expected to be negligible. As Kasy and Sautmann points out, treatment assignment probabilities for sub-optimal arms are bound to be away from zero. Moreover, given that this algorithm ensures that each treatment arm exponentially converges its posterior probability, this property allows us, in theory, to use conventional t-stats to draw inferences of treatment effects. However, as Bahety et al. (2021) indicate, given that some treatment conditions may end up with fewer observations, randomization inference would be a suitable approach to deal with possible high-leveraging observations. Caria et al. (2020) follow this same approach of providing randomization-based p-values under a sharp null hypothesis of no treatment effects.[16]

Inference for adaptive experiments is a relatively new and active area of study. Zhang, Janson

[16]high leveraging observations affect the variance of the estimate. An alternative approach is to apply weighting methods, followed by bootstrapping to account for this.

and Murphy (2020) proposed a Batched OSL (BOLS), where treatment effects are estimated using OLS at the batch level and then weighted by the estimated variance of all batches. The authors prove that their BOLS estimator follows an asymptotically normal distribution. Furthermore, their simulations show that their estimation strategy significantly outperforms similar approaches in reducing a type-I error.[17]

Along with providing estimates from the previously outlined estimators, we report estimates pooled unweighted and Inverse probability of Treatment Weighting (IPTW) estimates proposed by Horvitz and Thompson (1952). In this statistical technique, each observation is weighted by the inverse probability of being assigned to its condition. This estimator would be a sufficient adjustment to yield consistent and asymptotically normal estimates.[18]

Equation 5 shows the econometric specification to estimate the difference in means across treatments. The coefficient of interest in this model is β_k, which represents the vector of average treatment effects for each treatment arm.

$$Y_i = \beta_{0t} + \sum_{k=1}^{5} \beta_j T_{ij} + \omega X_i + \epsilon_i \tag{5}$$

The outcome variable Y_i is a binary outcome of whether respondents clicked or did not click to get information regarding the levels of malfeasance within their local government. Finally, X_i is a vector of covariates, including age, gender, and education. Given that this design did not incorporate a pure control, the reference condition is the *Standard Spatial* arm. This experimental condition, under our theoretical framework, would represent the least sophisticated messaging strategy.[19]

5 Results

In this section, we discuss the performance of the different information treatments for behavioral information-seeking, evaluation items, and stated preference outcomes for all three studies. We present the results by examining basic statistics and treatment effect estimates from both

[17] Similarly, Hadad et al. 2021 has put forward an *Adaptively Weighted Augmented-Inverse Probability Weighted Estimator* (AW-AIPW), which also yields asymptotically unbiased and normal with low variance estimations.

[18] As long as there is evidence of convergence of each arm to their posterior probability of being the best (Jack and Lorenzo, 2017)

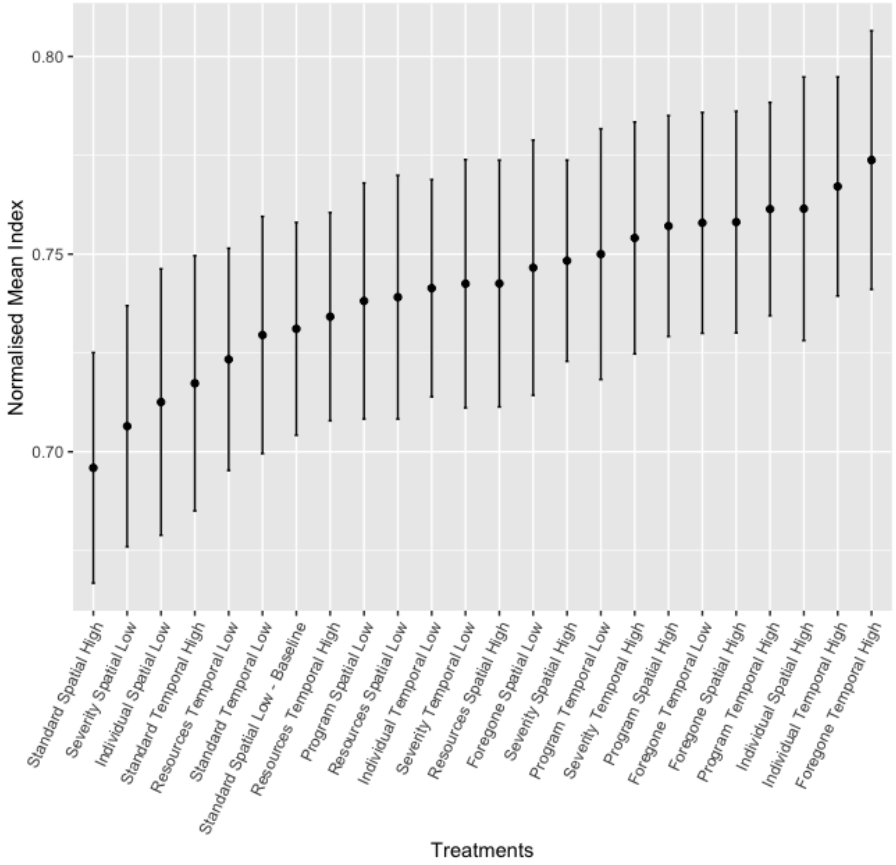
[19] Although this treatment condition is among the most promising treatment arms, this is based on my prior belief that this condition would yield the lowest CTR as it is the most rudimentary messaging strategy. Furthermore, this is the reference category set in the pre-analysis plan. The pre-analysis plan is available at https://www.socialscienceregistry.org/trials/7233

unweighted and weighted estimators, which are tailored for adaptively collected data. Based on these results, we discuss further the findings on whether we can identify systematic evidence to indicate whether the framing and content of information about malfeasance matter.

Repeated Measures Factorial Design - Study 1 As previously described, we used four commonly employed messaging evaluation items, asking respondents to assess each video treatment on the dimensions of credibility, persuasiveness, reliability, and accuracy.

Figure 1 presents the treatment effects for all 24 video treatments. The outcome is a standardized measure combining all evaluation items. The Foregone-Temporal and Individual Temporal treatments rank the highest. Regarding framing, the temporal frame performs slightly better than the spatial frame. The multivariate analyses reported in Tables 7 and 8 confirm the same patterns shown in Figure 1, even controlling for relevant covariates and order effects.

Figure 1: Reliability Index (normalized) by Treatment Arms



Note: This plot shows treatment effects for all 23 treatment arms. The reference category is the *Baseline Spatial Low*.This treatment effects include order fixed effects. Standard errors are clustered at the subject level.

These results suggest that information frames can play a crucial in persuading individuals when learning about corruption. Although these estimations are underpowered, Programme, Individual, and Foregone emerge as the most promising arms. Each of these exhibits larger treatment effects in both the temporal and spatial framing conditions. The Foregone level appears particularly promising due to the magnitude of its positive coefficient; however, these estimates lack precision and should be interpreted with caution.

The *Benchmark* model in Table 7 includes a dummy variable for the temporal treatments. This model confirms that temporal benchmarking is a slightly more persuasive messaging strategy than spatial benchmarking. However, in this case, the difference in means between these two levels is statistically indistinguishable.

In the *Metrics* model, we regressed the standardized evaluation index on five dummy variables representing the metric treatments, *Programme, Individual, Resources, Severity, and Standard*, with the *Standard metric* serving as the reference category. Consistent with previous findings, the Foregone metric emerges as the most positively evaluated messaging treatment. While all other treatments have positive coefficients, these are comparatively smaller in magnitude.

Table 8 reports estimates of treatment effects, including relevant covariates and/or order fixed effects. These results are robust to these controls, and in fact, treatment effect coefficients tend to be larger.

Table 7: Regression Results - Information Treatments

	Dependent variable: Standardized Evaluation Index		
	Benchmark/Metrics	Benchmark	Metrics
Standard Temporal	0.010		
	(0.015)		
Severity Spatial	0.015		
	(0.015)		
Severity Temporal	0.035*		
	(0.015)		
Resources Spatial	0.027+		
	(0.015)		
Resources Temporal	0.015		
	(0.014)		
Programme Spatial	0.034*		
	(0.015)		
Programme Temporal	0.042**		
	(0.015)		
Individual Spatial	0.021		
	(0.016)		
Individual Temporal	0.041**		
	(0.014)		
Foregone Spatial	0.039*		
	(0.016)		
Foregone Temporal	0.051**		
	(0.016)		
Severity			0.019+
			(0.011)
Resources			0.016
			(0.011)
Programme			0.033**
			(0.011)
Individual			0.026*
			(0.011)
Foregone			0.040***
			(0.011)
Temporal		0.009	
		(0.006)	
Order FE No	No	No	No
Covariates	No	No	No
Num.Obs.	3804	3804	3804
R2	0.006	0.001	0.004

SE are clustered at the commune level. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 8: Regression results to Information Treatments - Including Controls and Order FE

	Dependent variable: Standardized Evaluation Index					
	Benchmark/Metrics	Benchmark	Metric	Benchmark/Metrics	Benchmark	Metric
Standard Temporal	0.007			0.002		
	(0.066)			(0.066)		
Severity Spatial	0.044			0.045		
	(0.066)			(0.066)		
Severity Temporal	0.110+			0.112+		
	(0.066)			(0.065)		
Resources Spatial	0.076			0.072		
	(0.063)			(0.063)		
Resources Temporal	0.064			0.062		
	(0.061)			(0.061)		
Programme Spatial	0.136*			0.123+		
	(0.068)			(0.068)		
Programme Temporal	0.165*			0.155*		
	(0.066)			(0.066)		
Individual Spatial	0.091			0.092		
	(0.069)			(0.068)		
Individual Temporal	0.150*			0.143*		
	(0.062)			(0.062)		
Foregone Spatial	0.111			0.102		
	(0.068)			(0.068)		
Foregone Temporal	0.180*			0.173*		
	(0.071)			(0.070)		
Severity			0.070			0.074
			(0.047)			(0.047)
Resources			0.067			0.066
			(0.045)			(0.045)
Programme			0.147**			0.138**
			(0.049)			(0.049)
Individual			0.116*			0.116*
			(0.048)			(0.048)
Foregone			0.141**			0.136**
			(0.049)			(0.049)
Temporal		0.036			0.035	
		(0.027)			(0.027)	
Order FE	No	No	No	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3336	3336	3336	3336	3336	q3336
R2	0.022	0.018	0.022	0.031	0.026	0.030

Standard errors are clustered at the commune level. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Results Adaptive Experiment - Study 2 Figure 2 reports the results for the behavioral outcome (CTR), the evaluation index (EI), and the treatment assignment shares computed by *Exploration* algorithm. We can observe the low variance of the *Foregone Spatial* and *Foregone Temporal* treatment arms for both behavioral and evaluation outcomes. The CTR outcome ranges between 0.45 to 0.8 for both treatment conditions. Meanwhile, the EI is between 0.75 and 1. Conversely, we observe that treatment conditions such as *Individual Temporal* and *Individual Spatial* have significant CTR and EI index variance. Regarding the algorithm's performance,

we observe how it balances the trade-off between exploiting optimal arms and exploring poor-performing ones. First, the algorithm bounds treatment assignment shares below %50 for the best-performing ones. In contrast, for suboptimal treatment, arms continue to receive experimental units throughout each batch but with very low probabilities.

Table 16 reports the CTR for all six treatment arms and reports this same outcome by frame and metric. We observe that both the *Foregone* and *Spatial* metrics outperform the *Individual* metric. Specifically, we found that the *Foregone Spatial* and *Standard Spatial* arms yield the highest CTRs equal to 0.54. This number means that out of all respondents assigned to those treatments, 54% clicked on the link. Regarding bench-marking information, the *Spatial* framing shows to be more compelling than a *Temporal* frame, but this difference is small; of two percentage points. Table 17 summarizes the number of respondents assigned to each treatment condition.

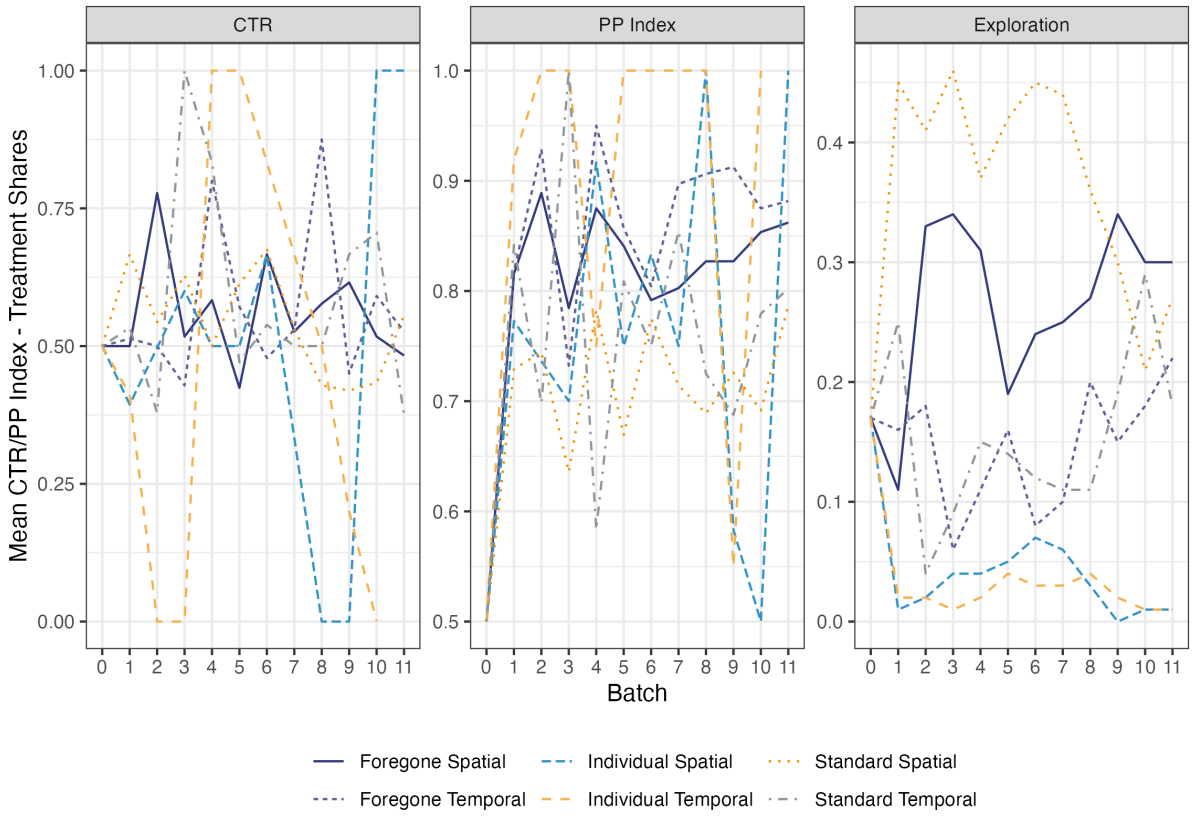


Figure 2: CTR, EI, and Assignment Shares from *Exploration Sampling*.

Note: This figure shows the CTR, EI, and treatment shares obtained from the exploration sampling for all treatments in each batch. We observe considerable variance in these metrics from one batch to another.

Once we aggregate outcomes, the results show how the persuasive proxy somewhat deviates from the behavioral outcome. Table 9 shows the proportion of people who answered 'yes' to each of the four binary items included in this battery of questions, including the evaluation index [20]. This analysis suggests that, once collapsing treatment arms by their metric, the *Individual* metric yields the highest rate of persuasiveness compared to the other two metrics. This result is somewhat puzzling, as this treatment condition attained the lowest CRT. We also notice barely any difference between temporal and spatial benchmarks for both outcomes. As for the different metrics, the results are more consistent between the behavioral and the evaluation items for the *Foregone* metric. This treatment factor is rated highly and has one of the highest click-through rates (CTR).

Table 9: Persuasiveness by Treatment, Frame and Metric

Condition	CTR	Convincing	Trustworthy	Clear	Precise	EI
Foregone Spatial	0.54	0.74	0.81	0.89	0.75	0.80
Foregone Temporal	0.53	0.79	0.81	0.90	0.77	0.82
Individual Spatial	0.41	0.78	0.74	0.85	0.82	0.80
Individual Temporal	0.39	0.88	0.80	0.98	0.92	0.89
Standard Spatial	0.54	0.62	0.67	0.77	0.59	0.66
Standard Temporal	0.52	0.68	0.75	0.88	0.73	0.76
Frame						
Spatial	0.53	0.68	0.73	0.82	0.67	0.73
Temporal	0.51	0.76	0.79	0.90	0.77	0.80
Metric						
Foregone	0.53	0.76	0.81	0.89	0.76	0.81
Individual	0.40	0.82	0.76	0.91	0.86	0.84
Standard	0.53	0.64	0.69	0.80	0.63	0.69

Note: This table provides the Click-through rate (CTR) for each arm and also reports the proportion of respondents that answer *Yes* to each of the PE items. The *Index* composite metric is the average of all four evaluation items. The table also reports these same indicators by *Frame* and by *Metric*.

The evidence is also mixed regarding the willingness to share the corruption information. Table 10 summarizes the proportion of respondents who reported willingness to share and/or post the videos. Notably, approximately 92% of those assigned to the Individual Temporal treatment indicated they would post the video on social media, despite this treatment having the lowest click-through rate (CTR). Consistent with earlier results, the Foregone Spatial treat-

[20]There is extensive literature that investigates how to construct a Composite index, but for the baseline results, we only provide this indicator. We will explore other composite indexes in the last version of the manuscript.

ment yielded some of the highest proportions of respondents willing to share or post the video containing information about local government malfeasance. These findings reinforce previous evidence that the Foregone Spatial arm is among the most effective messaging strategies.

Table 10: Willingness to Sharing and/or Posting by Treatment, Frame and Metric

Condition	CTR	Posting	Sharing	Both P & S
Foregone Spatial	0.54	0.89	0.82	0.82
Foregone Temporal	0.53	0.89	0.79	0.79
Individual Spatial	0.41	0.74	0.62	0.62
Individual Temporal	0.39	0.92	0.80	0.80
Standard Spatial	0.54	0.84	0.69	0.69
Standard Temporal	0.52	0.77	0.71	0.71
Frame				
Spatial	0.53	0.73	0.73	0.73
Temporal	0.51	0.76	0.76	0.76
Metric				
Foregone	0.53	0.81	0.81	0.81
Individual	0.40	0.69	0.69	0.69
Standard	0.53	0.69	0.69	0.69

Note: This table summarizes the proportion of respondents that answered *Yes* to the *Posting* and *Sharing* questions. It also reports the proportion of respondents that answered *Yes* to both. The table also lists these same indicators by *Frame* and by *Metric*.

Multivariate Analysis The multivariate results align with the descriptive analysis presented earlier but offer a more nuanced understanding of which factors most influence citizens' responses to corruption information. Table 11 presents pooled from the model specified in Equation 5. This estimation approach does not explicitly account for the adaptive nature of the experimental design but leverages the fact that *Exploration Sampling* ensures exponential convergence of each treatment arm's posterior probability. The baseline category in these models is *Standard Spatial*, which ranks among the best-performing arms in terms of CTR but performs relatively poorly in the evaluation items; thus, most point estimates are negative for the CTR outcome and positive for the EI.

These results indicate that both *Individual* arms fail to be sufficiently compelling in motivating respondents to seek additional information about corruption. One dimension where both *Individual* treatments notably fall short of is their perceived *Trustworthiness*. This sug-

gests that certain specific attributes of this messaging strategy undermine its credibility and prevent respondents from learning more about corruption. Another possible explanation is the lower attentiveness of respondents assigned to these treatments. Table 15 in the Appendix summarizes the results of three attention-check questions included in the study, illustrating a significant drop in the number of people who correctly answered the question about the framing used in the videos for the *Individual Spatial* arm. Thus, there could be issues with treatment compliance. A third plausible explanation is that the indicators used in this study may not fully capture all the critical dimensions needed to distinguish which information strategies are most compelling to different audiences.

Table 11: Unweigthed and Weighted Results - Behavioral and Persuasiveness Index Outcomes

	CTR		EI		CTR		EI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	0.54***	0.24	0.66***	0.74***	0.54***	0.10	0.66***	0.88***
	(0.02)	(0.12)	(0.02)	(0.09)	(0.02)	(0.10)	(0.02)	(0.05)
Standard Temporal	−0.02	−0.04	0.10***	0.10***	−0.04	−0.04	0.10***	0.10***
	(0.05)	(0.05)	(0.03)	(0.03)	(0.05)	(0.04)	(0.03)	(0.03)
Individual Spatial	−0.13	−0.12	0.14***	0.12**	−0.13	−0.13	0.12**	0.12**
	(0.07)	(0.07)	(0.04)	(0.04)	(0.07)	(0.07)	(0.04)	(0.04)
Individual Temporal	−0.15*	−0.17*	0.23***	0.23***	−0.15	−0.14	0.22***	0.22***
	(0.07)	(0.07)	(0.04)	(0.04)	(0.08)	(0.08)	(0.04)	(0.04)
Foregone Spatial	0.00	0.00	0.14***	0.14***	0.01	0.01	0.13***	0.13***
	(0.04)	(0.04)	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.02)
Foregone Temporal	−0.01	−0.02	0.16***	0.16***	−0.01	−0.01	0.16***	0.16***
	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
R2	0.01	0.04	0.05	0.06	0.01	0.04	0.05	0.06
Adj. R^2	0.00	0.03	0.04	0.05	0.00	0.03	0.04	0.05
Num. obs.	1200	1179	1200	1179	1200	1200	1200	1200

Note: This table reports the Unweighted and Inverse Probability of Treatment Weighting estimator using ordinary least squares. Columns 1 and 2 report the treatment effects for the CTR outcome without and with covariates, respectively. Columns 3 and 4 report the estimates for the Composite PE index without and with covariates, respectively. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

The results for the *Foregone Spatial* arm are more consistent across all outcome measures. While there is no statistically significant difference in the click-through rate (CTR) outcome compared to the *Standard Spatial* baseline, the *Foregone Spatial* treatment stands out in terms of the Perceived Persuasiveness (PP) Index. It yields a sizable and positive treatment effect, highlighting its potential as an effective messaging strategy.[21]

[21]The average treatment effect on the EI is approximately 0.57 standard deviations above the mean of the baseline category.

As outlined in Section 4, we applied Inverse Probability of Treatment Weighting (IPTW) to address potential sampling bias resulting from the adaptive nature of the data collection process. Table 11 presents the average treatment effects estimated using both unweighted and IPTW-adjusted specifications. The point estimates and standard errors are broadly consistent across specifications (Columns 1 to 4). However, the treatment effect for the *Individual Temporal* condition is no longer statistically significant at the conventional 5% level under the IPTW adjustment.

We perform the same regression analysis by collapsing observations based on their frame and metrics component. The results of this analysis are set out in Table 12. When examining the behavioral outcome (CTR), we find no statistically significant differences across benchmark types, but there are for the evaluation outcome. We can observe that conveying information about malfeasance using a temporal comparison yields 2 to 4 percentage points lower than the spatial benchmark. However, this difference is small and statistically indistinguishable. Concerning the persuasion index, the difference is quite substantial, which is about 0.18 standard deviations from the baseline category.

When collapsing the arms by metric, the results are mixed. *Individual* underperforms relative to the *Standard* metric in terms of CTRs but receives higher evaluations in persuasion items. In contrast, the *Foregone* metric shows a more consistent pattern: it has a positive effect on both CTRs and the EI but yields statistical significance only for the latter.

Table 12: IPTW estimates by Frame and by Metric - Clicked and PE Index Outcomes

	Frame				Metric			
	CTR		EI		CRT		EI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	0.53^{***}	0.14	0.77^{***}	0.90^{***}	0.53^{***}	0.11	0.74^{***}	0.83^{***}
	(0.02)	(0.12)	(0.01)	(0.06)	(0.02)	(0.12)	(0.01)	(0.10)
Temporal	-0.03	-0.03	0.05^{**}	0.05^{**}				
	(0.03)	(0.03)	(0.02)	(0.02)				
Individual					-0.12^{*}	-0.12^{*}	0.08^{**}	0.08^{**}
					(0.05)	(0.05)	(0.03)	(0.03)
Foregone					0.01	0.01	0.10^{***}	0.10^{***}
					(0.03)	(0.03)	(0.02)	(0.02)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.00	0.03	0.01	0.01	0.01	0.04	0.03	0.04
Adj. R^2	0.00	0.03	0.01	0.01	0.00	0.03	0.03	0.03
Num. obs.	1200	1200	1200	1200	1200	1200	1200	1200

Note: This table reports the pooled weighted estimates using IPTW with and with covariates using ordinary least squares. Columns 1 and 2 report the treatment effects for the CTR outcome. Columns 3 and 4 report the estimates for the Composite EI. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Inference for Adaptive Experiment One of the advantages of using *Exploration sampling* algorithm is that sampling averages of each treatment arm should be consistent and asymptotically unbiased (Caria et al., 2020). This property allows the researcher to conduct inference without explicitly adjusting for the adaptive nature of the design. However, to rule out any possible bias we implemented randomization inference, a nonparametric approach that does not rely on this assumption. Table 13 reports the F-tests and associated p-values for treatment, frame, and metric comparisons. Based on these results, we fail to reject the sharp null of no differences across treatments and frames, suggesting there is no unique best arm for the CTR outcome. However, when examining the results by *frame*, we do find statistically significant differences in CTRs.

Table 13: Randomisation Inference and Anova F-test

	CTR		EI	
Treatments	F-statistics	Two-tailed p-value	F-statistics	Two-tailed p-value
R. Inference	1.555	0.173	8.634	0***
Anova Test	1.555	0.169	8.633	0***
Frame				
R. Inference	0.508	0.471	4.470	0.040
Anova Test	0.508	0.475	4.470	0.034*
Metric				
R. Inference	3.701	0.042*	18.284	0***
Anova Test	3.701	0.024	18.284	0***

Note: This table provides the F-statistics obtained from conducting randomization inference. The sharp null hypothesis of this test is that each observation would yield the same outcome, regardless of the allocated treatment. Based on the results from this analysis, there is sufficient evidence to reject the null hypothesis that all treatment arms are not equal. The estimation of the F-statistic from the randomization inference is computed using the Inverse Probability of Treatment Weights. The table also shows the F-statistics obtained from an ANOVA test using ordinary least squares where the outcome is the 'clicked' or evaluation index variable and the predictor is the treatment variable.

We extended this inferential analysis using a Batch OLS hypothesis testing procedure. This inference method is particularly suitable for batch adaptive experiments. [22] This procedure incorporates randomization inference in its setup, but the computation of the observed t-statistics is a weighted combination of each t-test from each batch. As Zhang, Janson and Murphy point out, this approach closely relates to compute p-values for group sequential experiments, or confidence intervals construction for the overall effect in a meta-analysis. Here, we imposed the sharp null that there are no differences to the reference category.[23]

Table 14 reports the p-values for both outcomes by treatment using the Batch OLS inference procedure. The results are consistent with those obtained from conventional hypothesis testing. For the CTR outcome, none of the treatment arms were statistically distinguishable from the reference category. However, for the composite score, we find that the *Foregone* treatment and the *Standard Temporal* arm yield statistically significant effects, suggesting these information strategies can indeed shape people's perception of the quality of these signals.

[22] Where confidence intervals are constructed using asymptotic distribution of estimators under finite samples. Furthermore, when the difference-in-means is near zero, and there is no non-stationary control. This technique is particularly suitable for small sample sizes.

[23] In a meta-analysis, once you construct the confidence interval of the overall effect, you weigh each study as each study is independent of the others. In adaptive experiments, each observation is independent of the other observations within its batch.

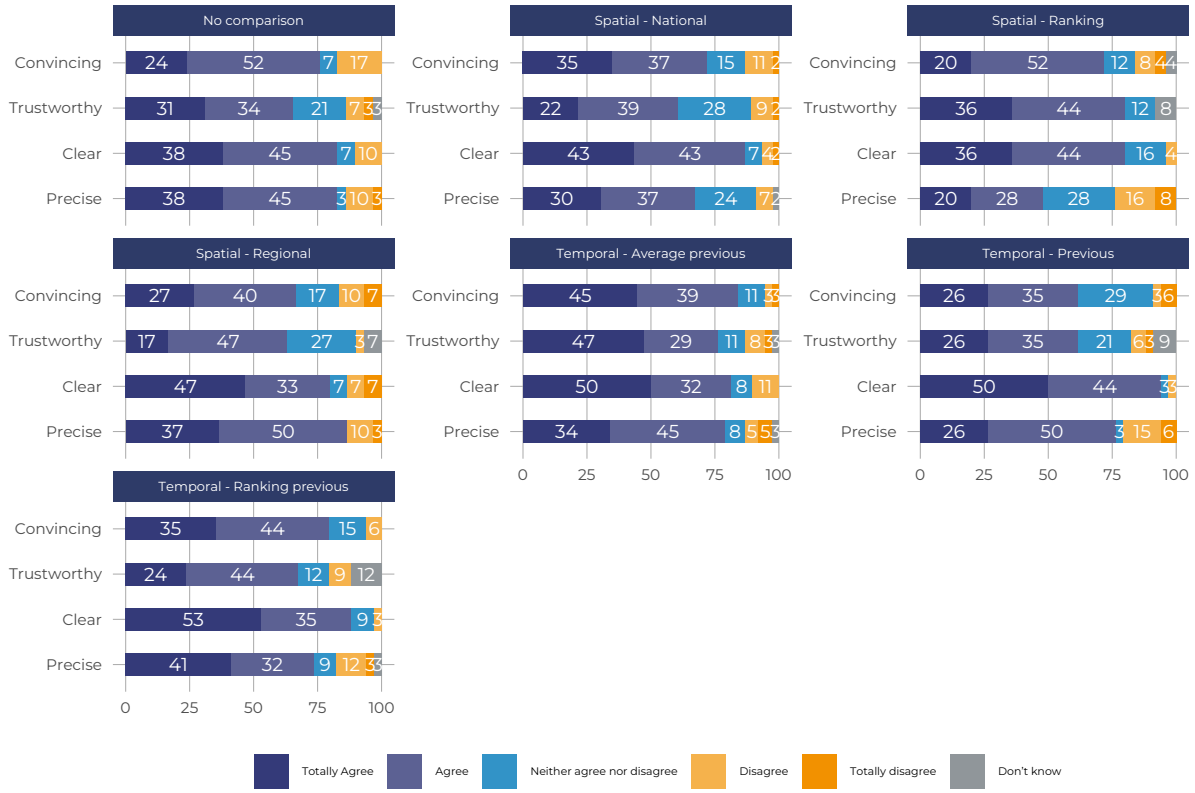
Table 14: Batched OLS Inference - Treatments

Treatment	P-values	
	CTR	EI
Standard Temporal	0.300	0.052
Individual Spatial	2.000	2.000
Individual Temporal	2.000	2.000
Foregone Spatial	0.737	0.000**
Foregone Temporal	0.815	0.000**
Frame		
Temporal	0.968	0.003*
Metric		
Individual	0.205	0.042*
Foregone	0.915	0.000**

Examining the frame and metric factors, Table 14 presents the results from the BOLS hypothesis testing procedure. Consistent with earlier findings, the Temporal framing is assessed more favorably on evaluation items, while no significant differences emerge in the behavioral outcome (CTR). Regarding the metrics factor, both the Individual and Foregone factors yield significantly higher evaluation scores, but we observe no difference for the behavioral outcome.

Results AI-enhanced (Pilot): We conducted a pilot with 236 respondents, where we asked them to evaluate 1 out of the 7 randomly assigned video treatments. Figure 3 summarizes the respondents' evaluations across multiple evaluation items. The results indicate that no single information strategy consistently outperforms across all evaluation criteria. The *Temporal - Average Previous* treatment was rated as the most convincing information treatment, whereas *Spatial - Ranking* was perceived as the most trustworthy. While, the *Temporal - Previous* treatment was evaluated as the clearest, and the *Spatial - Regional* was rated highest in terms of precision.

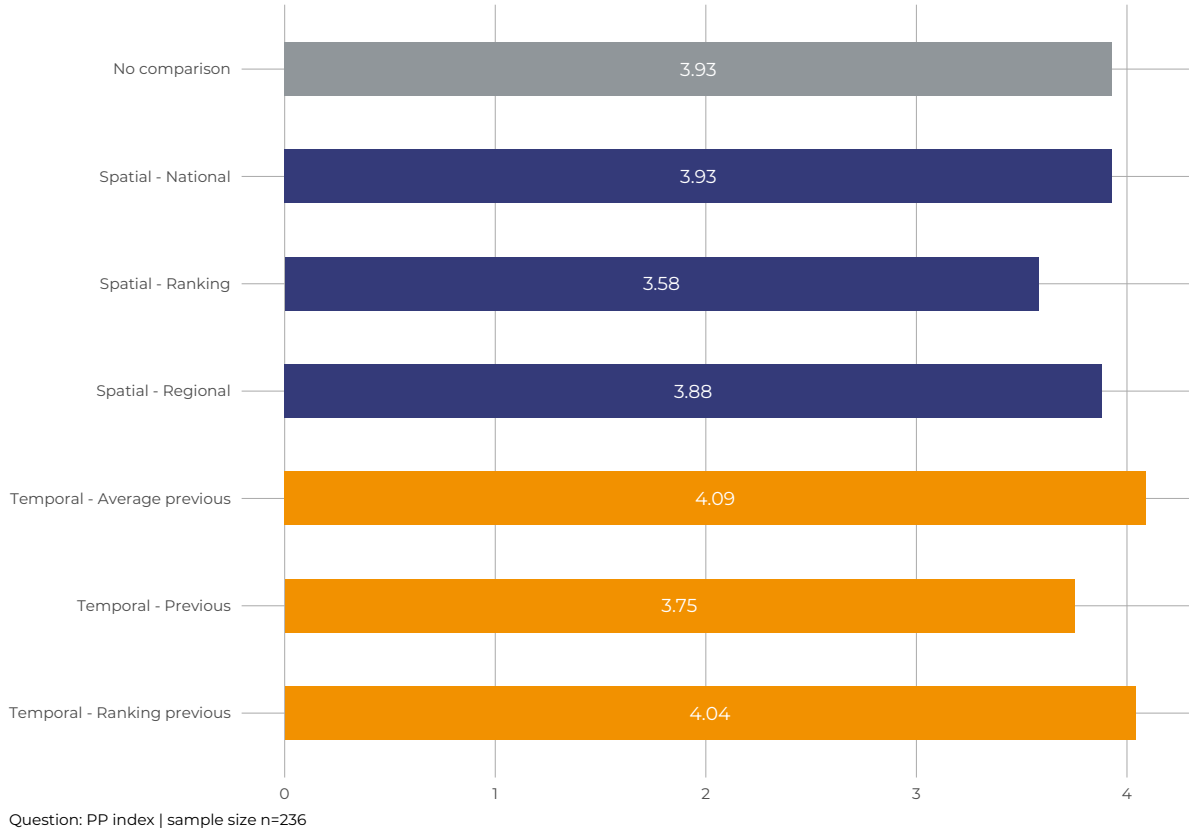
Figure 3: Evaluations - Closed-Ended Questions by *Treatment Status*



Question: The information provided in the video is.....? | sample size n=218

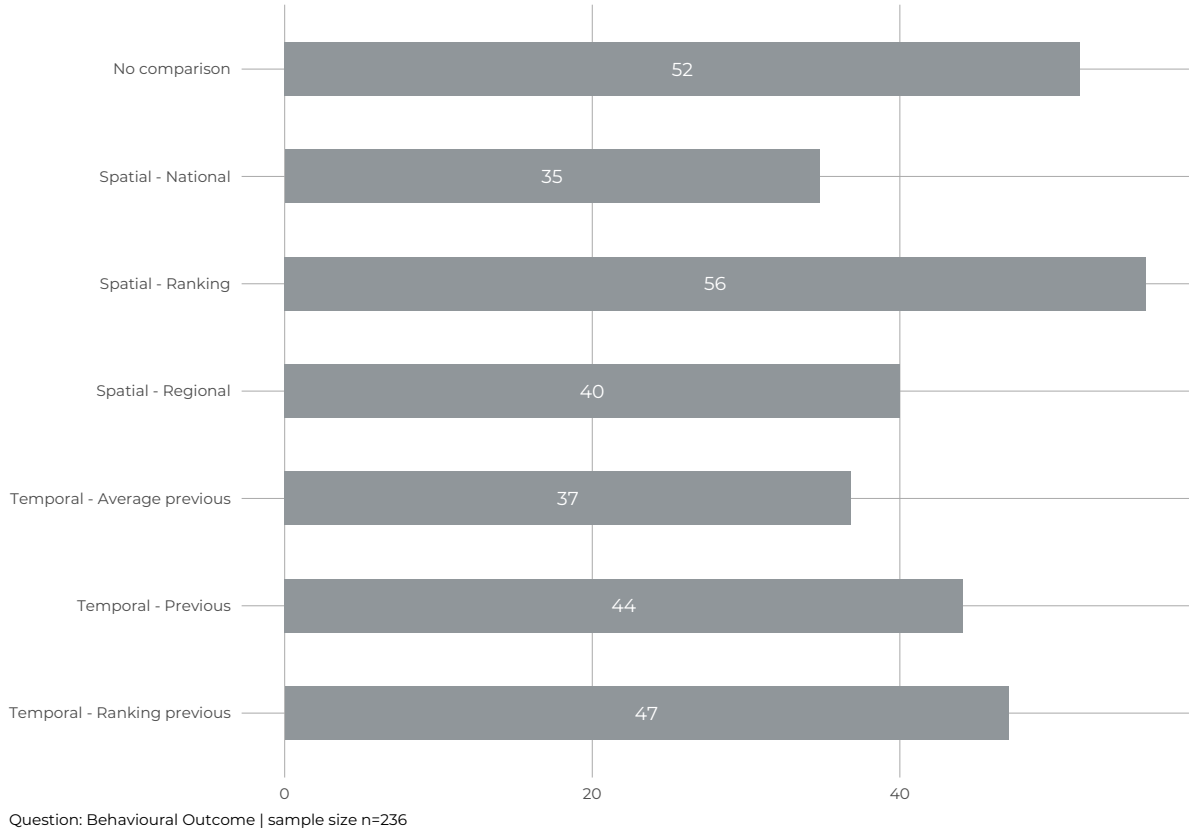
As in previous study, we constructed an index by averaging responses to all evaluation items. Based on this composite measure, the *Temporal - Average Previous* arises as the most compelling information strategy. Figure 4 presents the distribution of this index by treatment.

Figure 4: Evaluations - Closed-Ended Questions Index by *Treatment Status*



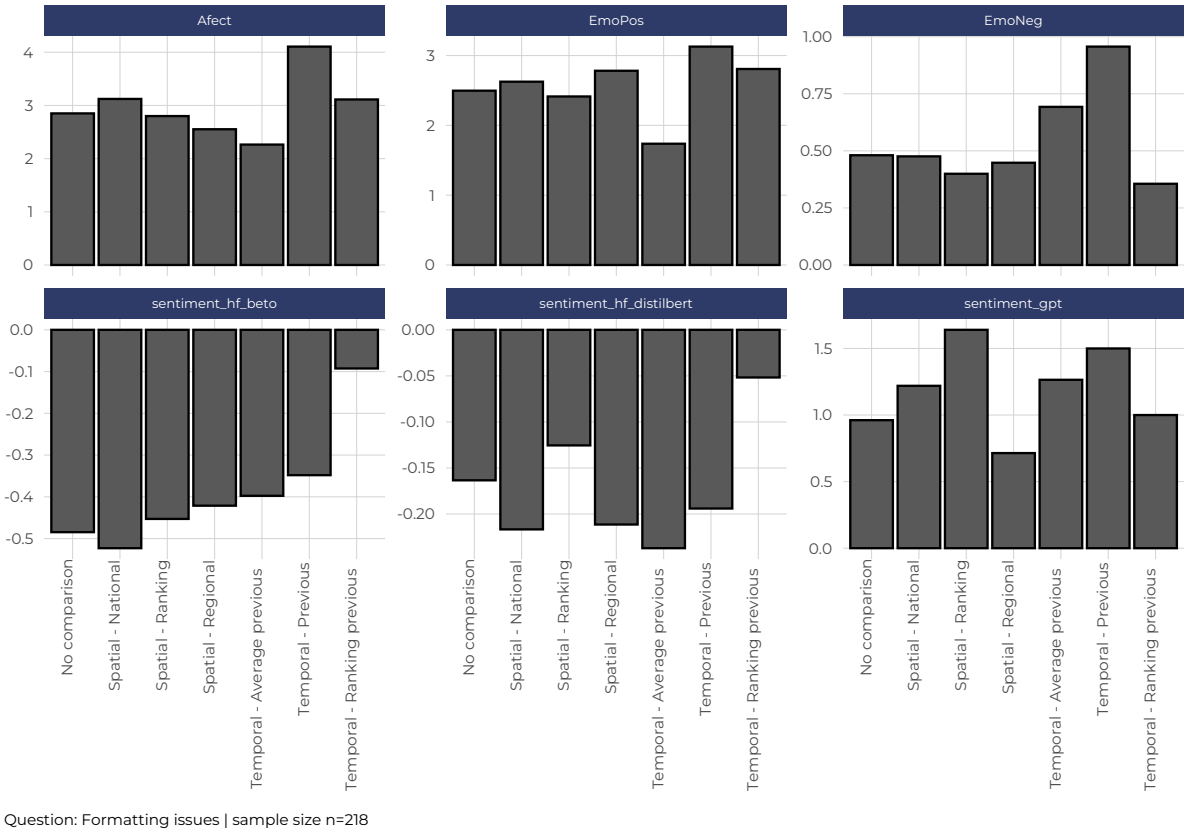
When we examine the results for the behavioral outcome, the *Spatial - Ranking* yields the higher CTR, with approximately 56% of respondents choosing to seek additional information about malfeasance in their local governments. This treatment yields similar rates to the reference condition, which has a 52% click-through rate (CTR).

Figure 5: Behavioural Outcome by *Treatment Status*



As pointed out before, we augmented our data collection by including an AI qualitative interview module in our survey. Based on this pilot data, there is a high degree of co-occurrence of features regarding the format and content provided in the different information treatments that respondents seem to care about. As a first approach, we calculated the sentiment scores of text generated from these interviews using different metrics. Negative values *sentiment_hf_beto* (-3 to -3) and *sentiment_hf_Distilbert* (-1 to 1) reflect a higher likelihood of a text containing negative emotions in the answers provided by the respondents. It stands out that the *Temporal Ranking* arm consistently yielded the lowest negative sentiment scores among all the treatments. Similarly, this treatment achieves the highest score for the *EmoPos* metric, which measures the prevalence of positive words in the text.

Figure 6: Sentiment Scores by *Treatment Status*



6 Discussion

This paper aimed to advance our understanding of which types of information citizens find most compelling and informative when learning about corruption in their local constituencies. Our research contributes to the growing body of work investigating the relationship between exposing elected officials' misconduct and subsequent political behavior, particularly voting. We placed special emphasis on the architecture of the informational signal, examining how the structure and content of corruption messages shape citizen engagement. Specifically, we identified which elements of message framing and content are most likely to capture the interest of the average citizen.

In order to achieve our research goal, we followed three different approaches: conducting multiple experiments and further examining the different features used in the malfeasance information literature. We began with a comprehensive number of treatment arms using a repeated-measures factorial design. Then, we transitioned to a small subset of treatment arms,

using conventional experiments and incorporating AI-powered qualitative interviews.

The results from these experiments identify two very different information strategies to be somewhat equally compelling. On one end, we find that *Standard Spatial*, which is the most straightforward and less sophisticated messaging strategy, performs well for the behavioral outcome, but poorly in the evaluation items. One simple conjecture that would explain why this arm performs well is that individuals may prefer uncomplicated and straightforward forms of information. In the *Standards* arms essentially, respondents received two statistics; therefore, the cognitive load respondents need to engage in learning about corruption should be reasonably low.

In the adaptive experiment, the evidence also shows that *Foregone Spatial* is an optimal information strategy. This treatment systematically yields higher CTRs and a high evaluation score. This treatment gradually obtains the highest posterior probability of being the best arm, with a considerably smaller variance. However, there is insufficient evidence to claim that this is the unique best arm. This finding also supports the extensive literature that argues that conveying information using a loss frame tends to be more effective. This treatment provided significantly richer content in terms of the complexity of the information presented in this arm. However, respondents tend to evaluate the clarity of this condition as very similar to the remaining treatments.

The results of the *Individual* arms are unclear. On one end, these arms perform very highly for the evaluation index but poorly for the CTR behavioral outcome. As raised before, these results can be explained due to lower levels of attentiveness found mainly for the *Individual Spatial* treatment. This issue is compounded by the fact that this arm received fewer observations[24], which would explain the inconsistent results we obtained for these treatments. Further batches of this treatment would help to determine with more certainty whether these arms are poor-performing ones.

With regards to the performance of the algorithm, *Exploration Sampling* algorithm indeed assigned non-zero treatment assignment probabilities to weak arms and less than 50% to the best-performing ones. However, the algorithm behaves quite erratically during the first batches. For example, it aggressively allocates over 40% of the respondents to the best arm, even though this arm received only 24 observations in the first batch. At the same time, under-performing arms received near-zero treatment assignment probabilities. Such treatment assignment shares

[24]There might be a group of inattentive respondents driving the results.

may not be appealing to researchers. An optimal strategy would be to favor exploration in the first batches and exploitation in the last ones.

An important limitation of this study is whether the selected outcome captures the persuasiveness and clarity of the different messaging strategies. An optimal approach would be to combine behavioral and evaluation measures of persuasiveness rather than relying on a single metric. Although using binary outcomes simplifies the estimation of the posterior probabilities of the best-performing arms, there may be a significant loss in capturing a more nuanced metric of the persuasiveness of the different information treatments.

Finally, given that the information treatments were hypothetical scenarios (based on actual data from audit reports), respondents may not engage enough, as the treatments did not contain information that may be relevant to them. A stronger and perhaps more robust design would be to implement this experiment in the field, utilizing information from their local constituency and a composite index. In this design, we cannot make claims about whether these information treatments shape individuals' political preferences or change beliefs. Still, it tells us that some information features are more compelling than others.

References

Agerberg, Mattias. 2020. "The Lesser Evil? Corruption Voting and the Importance of Clean Alternatives." *Comparative Political Studies* 53(2):253–287.

Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubín. 2018. Priors rule: When do Malfeasance Revelations Help or Hurt Incumbent Parties? NBER Working Papers 24888 National Bureau of Economic Research, Inc.

Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubin. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I.* Cambridge University Press chapter When Does Information Increase Electoral Accountability/ Lessons from a Field Experiment in Mexico.

Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubín. 2022. "Priors Rule: When Do Malfeasance Revelations Help Or Hurt Incumbent Parties?" *Journal of the European Economic Association* 20(4):1433–1477.
URL: *https://doi.org/10.1093/jeea/jvac015*

Augenblick, Ned and Matthew Rabin. 2021. "Belief Movement, Uncertainty Reduction, and Rational Updating*." *The Quarterly Journal of Economics* 136(2):933–985.
URL: *https://doi.org/10.1093/qje/qjaa043*

Avenburg, Alejandro. 2019. "Public Costs versus Private Gain: Assessing the Effect of Different Types of Information about Corruption Incidents on Electoral Accountability." *Journal of Politics in Latin America* 11(1):71–108.
URL: *https://doi.org/10.1177/1866802X19840457*

Bahety, Girija, Sebastian Bauhoff, Dev Patel and James Potter. 2021. "Texts don't nudge: An adaptive trial to prevent the spread of COVID-19 in India." *Journal of Development Economics* 153:102747.
URL: *https://www.sciencedirect.com/science/article/pii/S0304387821001140*

Banerjee, A., R. Hanna and S. Mullainathan. 2012. Corruption. In *Handbook of Organizational Economics*, ed. Mark P. Zanna. Princeton University Press.

Benartzi, Shlomo and Richard H. Thaler. 2007. "Heuristics and Biases in Retirement Savings Behavior." *The Journal of Economic Perspectives* 21(3):81–104.

Besley, Timothy. 2006. *Principled Agents.* Oxford University Press.

Besley, Timothy and A. Case. 1995. "Incumbent Behavior: Vote Seeking, Tax Setting and Yardstick Competition." *American Economic Review* 85:25–45.

Bhandari, Abhit, Horacio Larreguy and John Marshall. 2019. "Able and Mostly Willing: An Empirical Anatomy of Information's Effect on Voter-Driven Accountability in Senegal." Working paper.

Bidwell, Kelly, Katherine Casey and Rachel Glennerster. 2020. "Debates: Voting and Expenditure Responses to Political Communication." *Journal of Political Economy* 128(8):2880–2924.
URL: *https://doi.org/10.1086/706862*

Boas, Taylor C., F. Daniel Hidalgo and Marcus Andre Melo. 2019. "Norms versus Action: Why Voters Fail to Sanction Malfeasance in Brazil." *American Journal of Political Science* 63(2):385–400.

Bobonis, Gustavo J., Luis R. Cámara Fuertes and Rainer Schwabe. 2016. "Monitoring Corruptible Politicians." *American Economic Review* 106(8):2371–2405.

Botero, Sandra, Rodrigo Castro Cornejo, Laura Gamboa, Nara Pavao and David W. Nickerson. 2015. "Says Who? An Experiment on Allegations of Corruption and Credibility of Sources." *Political Research Quarterly* 68(3):493–504.

Breitenstein, Sofia. 2019. "Choosing the crook: A conjoint experiment on voting for corrupt politicians." *Research & Politics* 6(1):2053168019832230.

Buntaine, Mark T. and Brigham Daniels. 2020. "Combining bottom-up monitoring and top-down accountability: A field experiment on managing corruption in Uganda." *Research & Politics* 7(3):2053168020934350.
URL: *https://doi.org/10.1177/2053168020934350*

Caria, Stefano, Grant Gordon, Maximilian Kasy, Simon Quinn, Soha Shami, Teytelboym and Alex. 2020. "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan.".

Carpenter, Jeffrey, Emiliano Huet-Vaughn, Peter Hans Matthews, Andrea Robbett, Dustin Beckett and Julian Jamison. 2020. "Choice Architecture to Improve Financial Decision Making." *The Review of Economics and Statistics* 0(ja):1–52.

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(1):55–71.

Cronqvist, Henrik, Richard H. Thaler and Frank Yu. 2018. "When Nudges Are Forever: Inertia in the Swedish Premium Pension Plan." *AEA Papers and Proceedings* 108:153–58.

de Figueiredo, Miguel, Fernando Hidalgo and Yuri Kasahara. 2012. "When Do Voters Punish Corrupt Politicians? Experimental Evidence from Brazil." *SSRN Electronic Journal* .

Deshpande, Yash, Lester Mackey, Vasilis Syrgkanis and Matt Taddy. 2017. "Accurate Inference for Adaptive Linear Models." No notes.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan Hyde, Craig McIntosh and Gareth Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I.* Cambridge University Press.

Eggers, Andrew C., Nick Vivyan and Markus Wagner. 2018. "Corruption, Accountability, and Gender: Do Female Politicians Face Higher Standards in Public Life?" *The Journal of Politics* 80(1):321–326.

Engler, Sarah. 2020. "Fighting corruption or fighting the corrupt elite"? Politicizing corruption within and beyond the populist divide." *Democratization* 27(4):643–661.
URL: *https://doi.org/10.1080/13510347.2020.1713106*

Enrique, Jose Ramon, Horacio Larreguy, John Marshall and Alberto Simpser. 2019. "Online Political Information: Facebook Ad Saturation and Electoral Accountability in Mexico." Working paper.

Esposito, Bruno and Anja Sautmann. 2022. "Adaptive Experiments for Policy Choice : Phone Calls for Home Reading in Kenya.".

Ferejohn, John. 1986. "Incumbent Performance and Electoral Control." *Public Choice* 50(3):5–25.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123(2):703–45.

Ferraz, Claudio and Frederico Finan. 2011. "Electoral Accountability and Corruption in Local Governments: Evidence from Audit Reports." *American Economic Review* 101(4):1274–1311.

Figueroa, Rosa L., Qing Zeng-Treitler, Sasikiran Kandula and Long H. Ngo. 2012. "Predicting sample size required for classification performance." *BMC Medical Informatics and Decision Making* 12(1):8.

Fiorina, Morris. 2006. *Culture Wars: The Myth of a Polarized America, 2nd Edition.* Pearson Longman.

Franchino, Fabio and Francesco Zucchini. 2015. "Voting in a Multi-dimensional Space: A Conjoint Analysis Employing Valence and Ideology Attributes of Candidates." *Political Science Research and Methods* 3(2):221–241.

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager and Susan Athey. 2021. "Confidence intervals for policy evaluation in adaptive experiments." *Proceedings of the National Academy of Sciences* 118(15).

Hargreaves Heap, Shaun, Abhijit Ramalingam and David Rojo Arjona. 2017. "Social Information "Nudges": An Experiment with Multiple Group References." *Southern Economic Journal* 84(1):348–365.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/soej.12210*

Healy, Andrew and Neil Malhotra. 2013. "Retrospective Voting Reconsidered." *Annual Review of Political Science* 16(1):285–306.

Horvitz, D. G. and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47(260):663–685.

Humphreys, Macartan. and Weinstein Jeremy. 2012. Policing Politicians: Citizen Empowerment and Political Accountability in Uganda. Working paper, international growthcentre International Growth Centre.

Incerti, Trevor. 2019. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Survey Experiments." Working paper.

Jack, Bowden and Trippa Lorenzo. 2017. "Unbiased estimation for response adaptive clinical trials." *Stat Methods Med Res* 5(26):2376–2388.

Kahneman, Daniel. 1992. "Reference points, anchors, norms, and mixed feelings." *Organizational Behavior and Human Decision Processes* 51(2):296–312. Decision Processes in Negotiation.
URL: *https://www.sciencedirect.com/science/article/pii/074959789290015Y*

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2):263–291.

Kasy, Maximilian and Anja Sautmann. 2021. "Adaptive Treatment Assignment in Experiments for Policy Choice." *Econometrica* 89(1):113–132.

Klašnja, Marko, Noam Lupu and Joshua A. Tucker. 2020. "When Do Voters Sanction Corrupt Politicians?" *Journal of Experimental Political Science* pp. 1–11.

Klašnja, Marko, Noam Lupu and Joshua A. Tucker. 2021. "When Do Voters Sanction Corrupt Politicians?" *Journal of Experimental Political Science* 8(2):161–171.

Lagunes, Paul and Brigitte Seim. 2021. *The State of Experimental Research on Corruption Control*. Cambridge University Press pp. 526–543.

Larreguy, Horacio, John Marshall and Jr. Snyder, James M. 2020. "Publicising Malfeasance: When the Local Media Structure Facilitates Electoral Accountability in Mexico." *The Economic Journal* 130(631):2291–2327.

Liebman, Jeffrey B. and Erzo F. P. Luttmer. 2015. "Would People Behave Differently If They Better Understood Social Security? Evidence from a Field Experiment." *American Economic Journal: Economic Policy* 7(1):275–99.

Lierl, Malte and Marcus Holmlund. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press chapter Performance Information and Voting Behavior in Burkina Faso's Municipal Elections: Separating the Effects of Information Content and Information Delivery.

Mares, Isabela. 2003. *The Politics of Social Risk*. New York: Cambridge University Press.

Matz, SC, Kosinski M, Nave G and Stilwell DJ. 2017. "Psychological targeting as an effective approach to digital mass persuasion." *Proc Natl Acad Sci USA* 28(114(48)):12714–12719.

Nie, Xinkun, Xiaoying Tian, Jonathan Taylor and James Zou. 2018. Why Adaptively Collected Data Have Negative Bias and How to Correct for It. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, ed. Amos Storkey and Fernando Perez-Cruz. Vol. 84 of *Proceedings of Machine Learning Research* PMLR pp. 1261–1269.
URL: *https://proceedings.mlr.press/v84/nie18a.html*

Offer-Westort, Molly, Alexander Coppock and Donald P. Green. 2021. "Adaptive Experimental Design: Prospects and Applications in Political Science." *American Journal of Political Science* n/a(n/a).

O'Keefe, Daniel J. 2021. "Persuasive Message Pretesting Using Non-Behavioral Outcomes: Differences in Attitudinal and Intention Effects as Diagnostic of Differences in Behavioral Effects." *Journal of Communication* 71(4):623–645.
URL: *https://doi.org/10.1093/joc/jqab017*

Pallmann, Philip, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Lang'o Odondi, Matthew R. Sydes, Sofía S. Villar, James M. S. Wason, Christopher J. Weir, Graham M. Wheeler, Christina Yap and Thomas Jaki. 2018. "Adaptive designs in clinical trials: why use them, and how to run and report them." *BMC Medicine* 16(1):29.
URL: *https://doi.org/10.1186/s12916-018-1017-7*

Pande, Rohini. 2011. "Can Informed Voters Enforce Better Governance? Experiments in Low-Income Democracies." *Annual Review of Economics* 3(1):215–237.

Pande, Rohini, Abhijit Banerjee, Selvan Kumar and Felix Su. 2011. "Do Informed Voters Make Better Choices?".

Polk, Jonathan, Jan Rovny, Ryan Bakker, Erica Edwards, Liesbet Hooghe, Seth Jolly, Jelle Koedam, Filip Kostelka, Gary Marks, Gijs Schumacher, Marco Steenbergen, Milada Vachudova and Marko Zilovic. 2017. "Explaining the salience of anti-elitism and reducing political corruption for political parties in Europe with the 2014 Chapel Hill Expert Survey data." *Research & Politics* 4(1):2053168016686915.
URL: *https://doi.org/10.1177/2053168016686915*

Porten-Cheé, Pablo, Marlene Kunst, Ariadne Vromen and Michael Vaughan. 2021. "The effects of narratives and popularity cues on signing online petitions in two advanced democracies." *Information, Communication & Society* 0(0):1–21.
URL: *https://doi.org/10.1080/1369118X.2021.1991975*

Rafferty, Anna, Huiji Ying and Joseph Williams. 2019. "Statistical Consequences of using Multi-armed Bandits to Conduct Adaptive Educational Experiments." *Journal of Educational Data Mining* 11(1):47–79.
URL: *https://jedm.educationaldatamining.org/index.php/JEDM/article/view/357*

Russo, Daniel. 2020. "Simple Bayesian Algorithms for Best-Arm Identification." *Operations Research* 68(6):1625–1647.
URL: *https://doi.org/10.1287/opre.2019.1911*

Ryan, Timothy J. 2012. "What Makes Us Click? Demonstrating Incentives for Angry Discourse with Digital-Age Field Experiments." *The Journal of Politics* 74(4):1138–1152.
URL: *https://doi.org/10.1017/S0022381612000540*

Scott, Steven L. 2010. "A modern Bayesian look at the multi-armed bandit." *Applied Stochastic Models in Business and Industry* 26(6):639–658.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.874*

Shin, Jaehyeok, Aaditya Ramdas and Alessandro Rinaldo. 2019. "Are Sample Means in Multi-Armed Bandits Positively or Negatively Biased?".

Singh, Shane P. and Jason Roy. 2018. "Compulsory voting and voter information seeking." *Research & Politics* 5(1):2053168017751993.
URL: *https://doi.org/10.1177/2053168017751993*

Slivkins, Aleksandrs. 2020. "Introduction to Multi-Armed Bandits.".

Soll, Jack B., Ralph L. Keeney and Richard P. Larrick. 2013. "Consumer Misunderstanding of Credit Card Use, Payments, and Debt: Causes and Solutions." *Journal of Public Policy & Marketing* 32(1):66–81.

Thaler, Richard H. 2016. *Misbehaving: The making of behavioral economics*. Norton, London.

Thomas, Rosemary J., Judith Masthoff and Nir Oren. 2019. "Can I Influence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale." *Frontiers in Artificial Intelligence* 2.
URL: *https://www.frontiersin.org/articles/10.3389/frai.2019.00024*

Thompson, William R. 1933. "On the Likelihood that one Unknow Probability Exceeds Another in View of the Evidence of Two Samples." *Biometrika* 25(3-4):285–294.
URL: *https://doi.org/10.1093/biomet/25.3-4.285*

Vera, Sofia B. 2020. "Accepting or Resisting? Citizen Responses to Corruption Across Varying Levels of Competence and Corruption Prevalence." *Political Studies* 68(3):653–670.

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359(6380):1146–1151.
URL: *https://www.science.org/doi/abs/10.1126/science.aap9559*

Weitz-Shapiro, Rebecca and Matthew S. Winters. 2017. "Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil." *The Journal of Politics* 79(1):60–74.

Winters, Matthew S. and Rebecca Weitz-Shapiro. 2013. "Lacking Information or Condoning Corruption: When Do Voters Support Corrupt Politicians?" *Comparative Politics* 45(4):418–436.

Winters, Matthew S. and Rebecca Weitz-Shapiro. 2020. "Information credibility and responses to corruption: a replication and extension in Argentina." *Political Science Research and Methods* 8(1):169–177.

Wittenberg, Chloe, Ben M. Tappin, Adam J. Berinsky and David G. Rand. 2021. "The (minimal) persuasive advantage of political video over text." *Proceedings of the National Academy of Sciences* 118(47):e2114388118.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.2114388118*

Zhang, Kelly, Lucas Janson and Susan Murphy. 2020. Inference for Batched Bandits. In *Advances in Neural Information Processing Systems*, ed. H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin. Vol. 33 Curran Associates, Inc. pp. 9818–9829.
URL: *https://proceedings.neurips.cc/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf*

7 Appendix

Descriptives

This section contains some of the descriptive analyses referenced in the main sections of the manuscript.

Table 15: Proportion of Correct Answers by Treatment

Condition	CTR	Topic	Frame	Metric	All
Foregone Spatial	0.54	0.91	0.76	0.69	0.55
Foregone Temporal	0.53	0.91	0.92	0.75	0.71
Individual Spatial	0.41	0.91	0.54	0.84	0.44
Individual Temporal	0.39	0.96	0.90	0.90	0.82
Standard Spatial	0.54	0.88	0.88	0.83	0.77
Standard Temporal	0.52	0.85	0.87	0.84	0.78

Note: This table summarizes the proportion of people that answered correctly to the three attention check questions. The *Topic* column corresponds to the number of people who responded correctly to the attention check question asked about the video topic. The *Frame* column lists the proportion of people that answered correctly to the frame that asked respondents about the frame used in the video. The *Metric* columns report the proportion of respondents that responded correctly to the last attention check question that asked about the metric part of the video. Finally, the *All* column summarizes the portion of respondents that answered all three questions correctly.

Table 16: CTRs by Treatment, Frame, and Metric

Condition	CTR	SE
Foregone Spatial	0.54	0.03
Foregone Temporal	0.53	0.03
Individual Spatial	0.41	0.06
Individual Temporal	0.39	0.07
Standard Spatial	0.54	0.02
Standard Temporal	0.52	0.04
Frame		
Spatial	0.53	0.02
Temporal	0.51	0.02
Metric		
Foregone	0.53	0.02
Individual	0.40	0.05
Standard	0.53	0.02

Note: This table provides the Click-through rate (CTR) for each arm as well as the standard deviation of this metric. It also provides the same indicator by *Frame* and by *Metric*.

Table 17: Treatment Assignment

Treatment	Sample
Foregone Spatial	296
Foregone Temporal	207
Individual Spatial	68
Individual Temporal	49
Standard Spatial	408
Standard Temporal	172
Sum	1200

Note: This table reports the number of respondents assigned in each experimental condition. In total, 1,200 respondents took part in this study.

Table 18: Comparison between 100 and 200 First Batch Size

Treatment	N.Obs: 100	N.Obs: 200
Foregone Spatial	0.22	0.11
Foregone Temporal	0.00	0.16
Individual Spatial	0.22	0.01
Individual Temporal	0.04	0.02
Standard Spatial	0.26	0.45
Standard Temporal	0.26	0.25

Note: This table reports the results from a pilot using 100 and 200 observations in the first batch using *Exploration sampling*. We identified from this analysis that treatment assignment shares for the second batch are susceptible to the batch size. We can find quite sizeable differences for the *Foregone Temporal* and *Individual Spatial*.

Performance of the Exploration Algorithm The main feature of *Exploration sampling* is the exploitation-exploration trade-off by binding its exploitation motives, but also seeking to identifying the best arm with a high degree of certainty. This property means that assignment shares for poor-performing treatment arms are bounded away from zero and set a ceiling for best-performing ones. In figure 8, the left-hand panel shows the posterior probability of each arm being optimal. The panel on the right-hand side shows the treatment assignment shares obtained from *Exploration sampling*. As depicted in this figure, the algorithm quickly departs from the equal-size treatment proportion assigned on the first batch. It allocates roughly 45% units to the *Standard Spatial* treatment condition in the second batch [25]. The algorithm moderates its exploitation motive by assigning fewer observations to the *Standard Spatial* arm, even though the probability of this arm being the best is roughly 73%.

We also find that the algorithm continuously allocates units to second-best arms such as *Standard Temporal* and *Foregone Temporal*, as both treatments gradually converge to their posterior probability. In the two final batches, we observe that the algorithm allots a similar proportions in three out of the six arms, favouring exploration amongst the optimal arms. Finally, the *Foregone Spatial* arm is among the optimal messaging strategies throughout the experiment. This treatment condition yields the second-highest probability of being the best from batches two to ten, and in the last batch obtains the highest probability. Simultaneously, the *Standard Spatial* condition was in a steep downward decline from batch six onward.

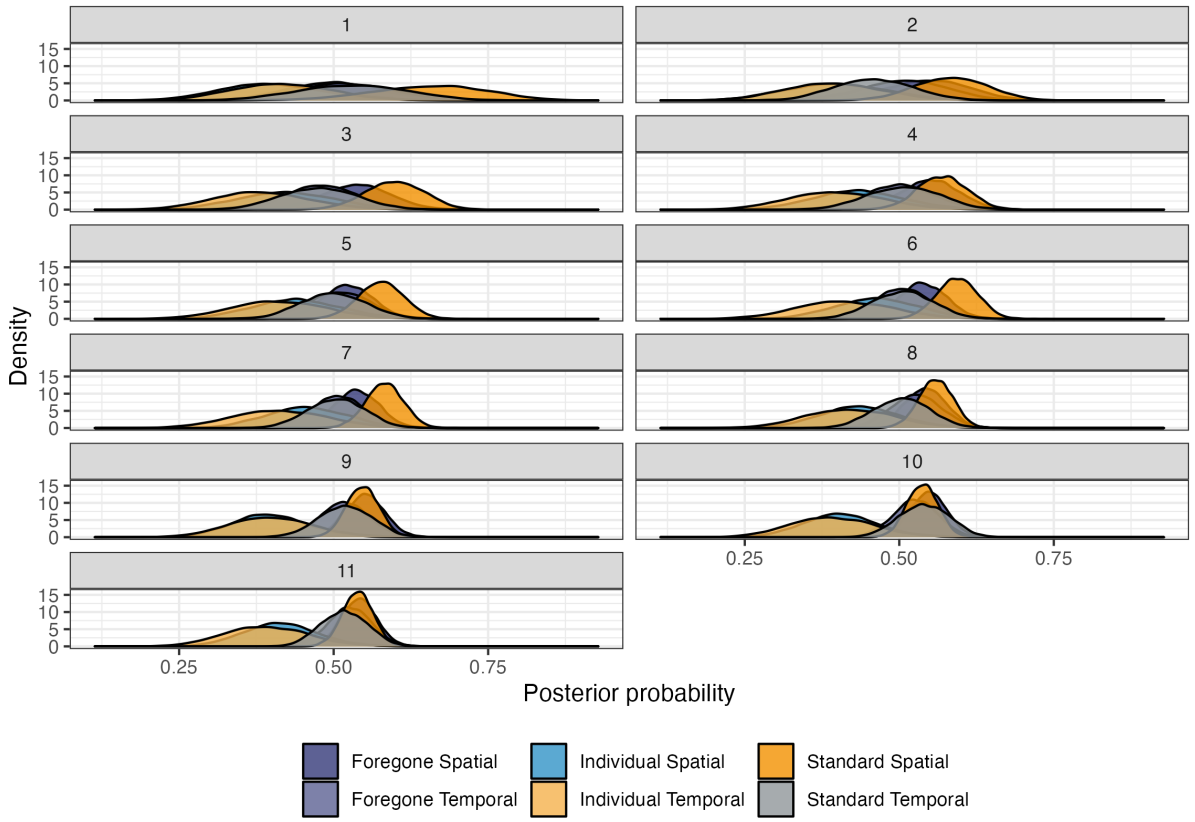
Table 19 in the Appendix summarizes the results of the posterior probability of each arm being the best, the standard deviation of the posterior probability distribution, and the sampling share for each arm broken-down by treatment and by batch. Figure 7 complements this analysis by depicting the posterior distribution of each arm by batch and by treatment.

[25] Using Thompson sampling, around 73% of respondents would be allocated to the *Standard Spatial* and roughly 100% using Epsilon Greedy.

Treatment	Batch	Posterior	Sd Posterior	Exploration
Foregone Spatial	0	0.17	-	0.17
Foregone Temporal	0	0.17	-	0.17
Individual Spatial	0	0.17	-	0.17
Individual Temporal	0	0.17	-	0.17
Standard Spatial	0	0.17	-	0.17
Standard Temporal	0	0.17	-	0.17
Foregone Spatial	1	0.05	0.07	0.11
Foregone Temporal	1	0.07	0.08	0.16
Individual Spatial	1	0.01	0.08	0.01
Individual Temporal	1	0.01	0.08	0.02
Standard Spatial	1	0.73	0.09	0.45
Standard Temporal	1	0.13	0.09	0.25
Foregone Spatial	2	0.28	0.07	0.33
Foregone Temporal	2	0.13	0.07	0.18
Individual Spatial	2	0.02	0.08	0.02
Individual Temporal	2	0.01	0.08	0.02
Standard Spatial	2	0.54	0.06	0.41
Standard Temporal	2	0.03	0.06	0.04
Foregone Spatial	3	0.18	0.05	0.34
Foregone Temporal	3	0.03	0.05	0.06
Individual Spatial	3	0.02	0.08	0.04
Individual Temporal	3	0.01	0.08	0.01
Standard Spatial	3	0.73	0.05	0.46
Standard Temporal	3	0.04	0.06	0.09
Foregone Spatial	4	0.29	0.05	0.31
Foregone Temporal	4	0.08	0.05	0.11
Individual Spatial	4	0.03	0.07	0.04
Individual Temporal	4	0.01	0.08	0.02
Standard Spatial	4	0.48	0.04	0.37
Standard Temporal	4	0.12	0.06	0.15
Foregone Spatial	5	0.11	0.04	0.19
Foregone Temporal	5	0.10	0.05	0.16
Individual Spatial	5	0.03	0.07	0.05
Individual Temporal	5	0.02	0.08	0.04
Standard Spatial	5	0.66	0.04	0.42
Standard Temporal	5	0.08	0.05	0.14
Foregone Spatial	6	0.11	0.04	0.24
Foregone Temporal	6	0.04	0.04	0.08
Individual Spatial	6	0.03	0.07	0.07
Individual Temporal	6	0.01	0.08	0.03
Standard Spatial	6	0.76	0.03	0.45
Standard Temporal	6	0.05	0.05	0.12
Foregone Spatial	7	0.14	0.04	0.25
Foregone Temporal	7	0.05	0.04	0.10
Individual Spatial	7	0.03	0.07	0.06
Individual Temporal	7	0.02	0.08	0.03
Standard Spatial	7	0.71	0.03	0.44
Standard Temporal	7	0.06	0.05	0.11
Foregone Spatial	8	0.25	0.03	0.27
Foregone Temporal	8	0.17	0.04	0.20
Individual Spatial	8	0.02	0.06	0.03
Individual Temporal	8	0.03	0.07	0.04
Standard Spatial	8	0.46	0.03	0.36
Standard Temporal	8	0.08	0.05	0.11
Foregone Spatial	9	0.39	0.03	0.34
Foregone Temporal	9	0.12	0.04	0.15
Individual Spatial	9	0.00	0.06	0.00
Individual Temporal	9	0.01	0.07	0.02
Standard Spatial	9	0.31	0.03	0.30
Standard Temporal	9	0.16	0.04	0.19
Foregone Spatial	10	0.33	0.03	0.30
Foregone Temporal	10	0.16	0.04	0.18
Individual Spatial	10	0.00	0.06	0.01
Individual Temporal	10	0.01	0.07	0.01
Standard Spatial	10	0.20	0.03	0.21
Standard Temporal	10	0.30	0.04	0.29
Foregone Spatial	11	0.33	0.03	0.30
Foregone Temporal	11	0.20	0.03	0.22
Individual Spatial	11	0.01	0.06	0.01
Individual Temporal	11	0.01	0.07	0.01
Standard Spatial	11	0.28	0.02	0.27
Standard Temporal	11	0.16	0.04	0.18

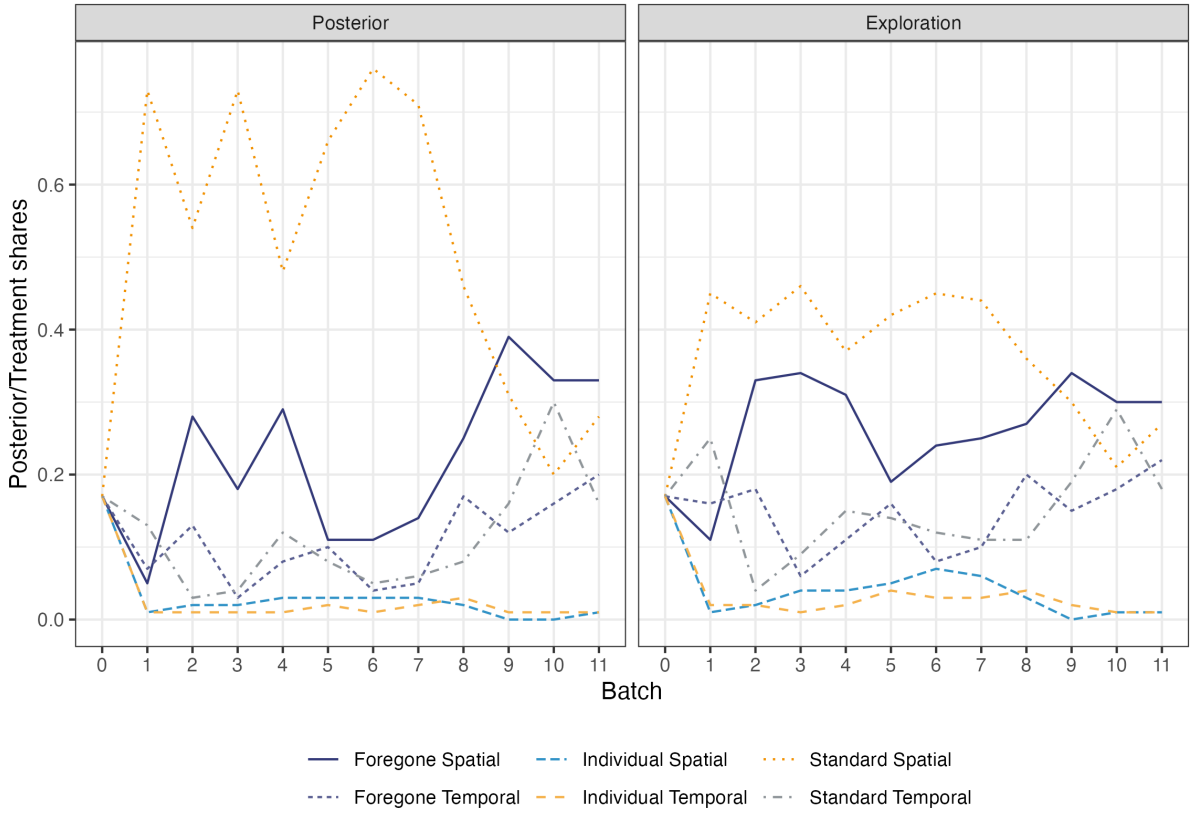
Table 19: Posterior probability, standard deviation posterior probability and exploration sampling share by treatment arm

Figure 7: Posterior Probability by Batch



Note: This figure depicts the posterior density of each arm being the best from batch 1 to batch 11. It is important to highlight that the posterior probability of being the best is not the mean of posterior probability density, but is the calculated based on many times each arm attains the highest posterior probability from each draw of the MonteCarlo simulations. We can observe that the mean of the distribution for the *Standard Spatial* and *Foregone Temporal* are centered around 0.55.

Figure 8: Posterior Probability and Treatment Shares from *Exploration Sampling*



Note: This figure depicts the probability of each arm being the best (left) and the treatment assignment shares obtained from the *Exploration sampling* algorithm (right). The figure report for both metrics for all eleventh batches. As we can see in the *Posterior* panel, the best-performing treatment arm is the *Foregone Spatial* treatment obtaining a posterior probability of around 33%.

Simulations

We conducted several simulations that informed this study's design and sampling algorithm selection. We simulated a static and an adaptive experiment with 4 and 6 treatment arms with 100 and 120 observations in each batch, respectively. In this version of the manuscript, we only reported the simulations with 6 conditions. Figure 9 reports the results of these simulations. These simulations are a modified version of Offer-Westort, Coppock and Green simulation scenarios. Some simulation parameters are based on the results obtained from a pilot conducted in March 2021, where we collected 1,276 evaluations of the four treatment arms of interest. Using this data, we computed the posterior probability of being the best arm for each arm. The results of this analysis are reported in Table 20.

Scenario 1 - *Clear winner*: In this scenario, there is a clear best-performing treatment arm. But, the remaining three arms have the same but substantially lower probability of being the best. Scenario 2 - *No clear winner*: All treatment arms have a similar success rate. Scenario three - *Second best*: here, there is a best-performing treatment arm, but there is also a competing second-best treatment condition. Table 21 reports the simulation parameters for these three scenarios.

Table 20: Posterior Probability Best-performing Arm - Results from the Pilot

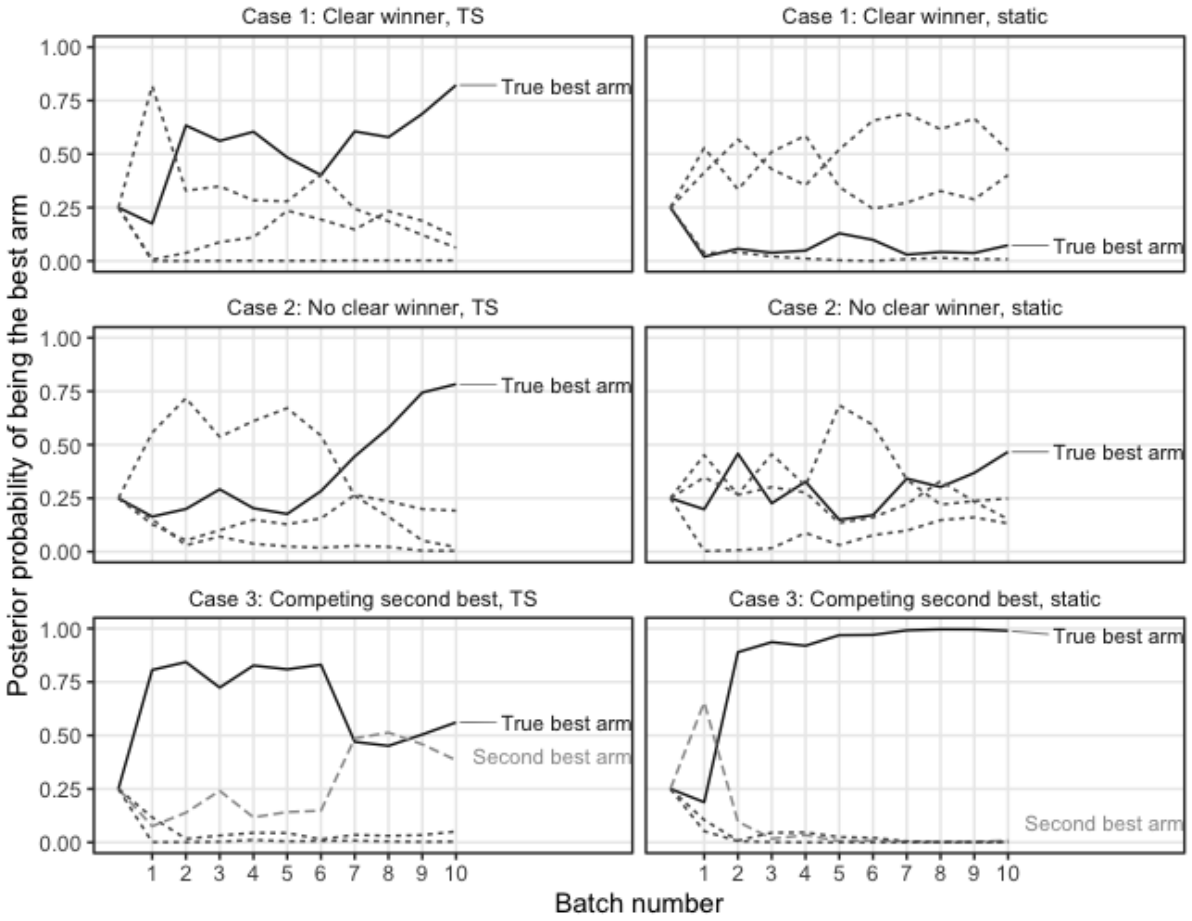
Treatment	Posterior probability
Individual Spatial	0.21
Individual Temporal	0.29
Foregone Spatial	0.20
Foregone Temporal	0.28

Table 21: Hyperparameter simulations

Parameters	Clear Winner	No clear winner	Second best
Prob. Best	0.30	0.28	0.30
Prob. Competing second	-	-	0.28
Prob. Other	0.23	0.24	0.21
Periods	10	10	10
Arms	4	4	4
Prob. Assign - Control $\hat{p}_{C,t}$	$q + R_t * (1 - q)$	$q + R_t * (1 - q)$	$q + R_t * (1 - q)$
Rescaling Prob - Others $\hat{p}_{k,t}$	$p_{k,t} * (1 - R_t) * (1 - q)$	$p_{k,t} * (1 - R_t) * (1 - q)$	$p_{k,t} * (1 - R_t) * (1 - q)$
Ceeling Prob. Best	0.9	0.9	0.9

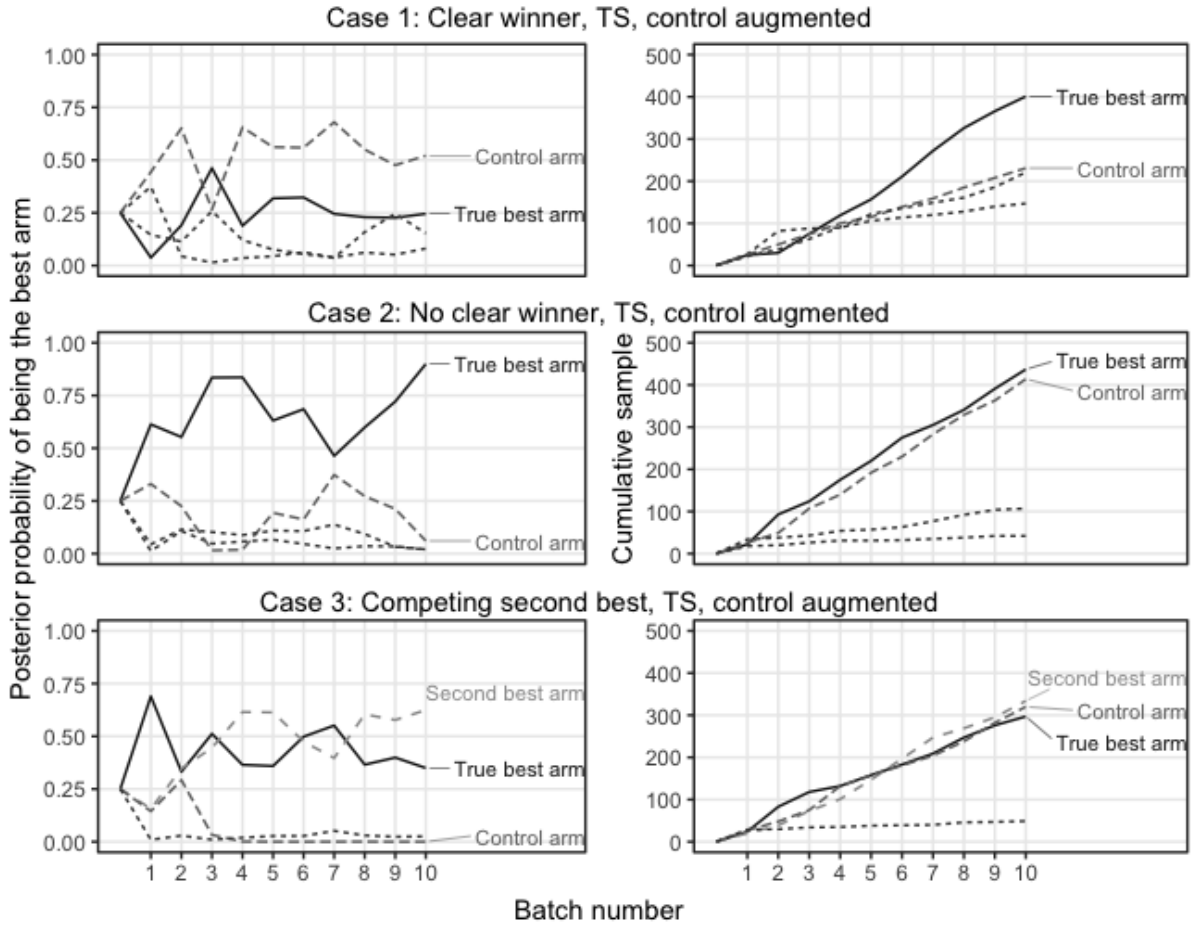
We compared the results of these scenarios using a Thompson sampling algorithm versus a static design where treatment assignment is equal to $1/K$, and K is the number of arms. As shown in Figure 9, the Thompson sampling algorithm successfully identifies the best-treatment arm in the *Clear winner* scenario. In the *No clear winner* scenario, we see that both designs identify the best arm, but the adaptive design yields a higher probability. Finally, in the *Second best* scenario, the static experiment outperforms the adaptive experiment. This result may be because the Thompson sampling algorithm struggles to reduce uncertainty about the quality of the "True" and "Second" arms in the final batches.

Figure 9: Simulated posteriors probabilities over Thompson Sampling and Static Designs



The second set of simulations compares the same three scenarios conducted in Offer-Westort, Coppock and Green but using their proposed Thompson Control-Augment algorithm. Their algorithm allocates a proportion of experimental units into the control condition, so the best arm and the control arm have similar cumulative samples. This modification would allow yielding accurate estimates of the treatment effects. Figure 10 shows that the Thompson Control-Augment sampling performs poorly in identifying the best-performing treatment arm in the *Clear winner* and in the *Competing second best* scenarios.

Figure 10: Simulated posteriors probabilities over Thompson Sampling and Control-Augmented



6 treatments - 1200 observations

- number of arms: 6
- number of batches (baseline): 10
- Total observations: 1200

Figure 11: Root mean Square Error by Number of Batches - 6 arms/1200 observations

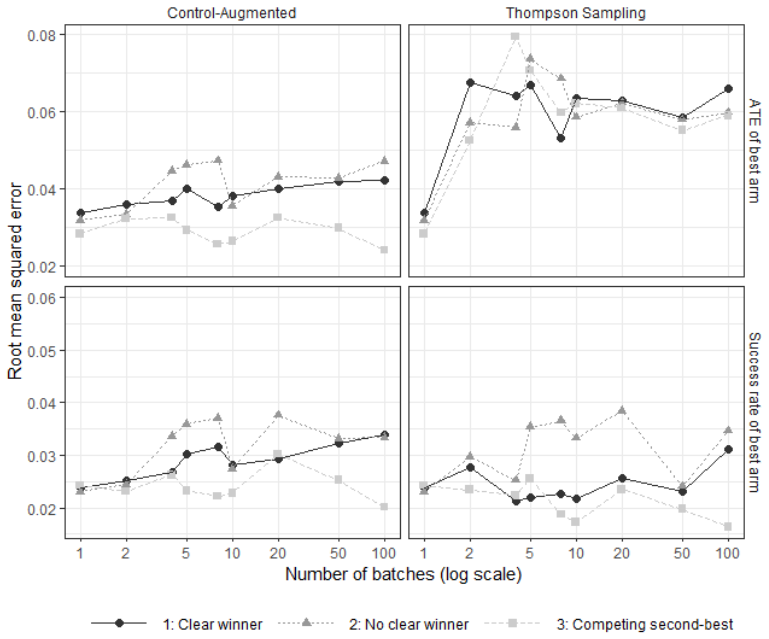


Figure 12: Arm Coverage by Number of Batches - 6 arms/1200 observations

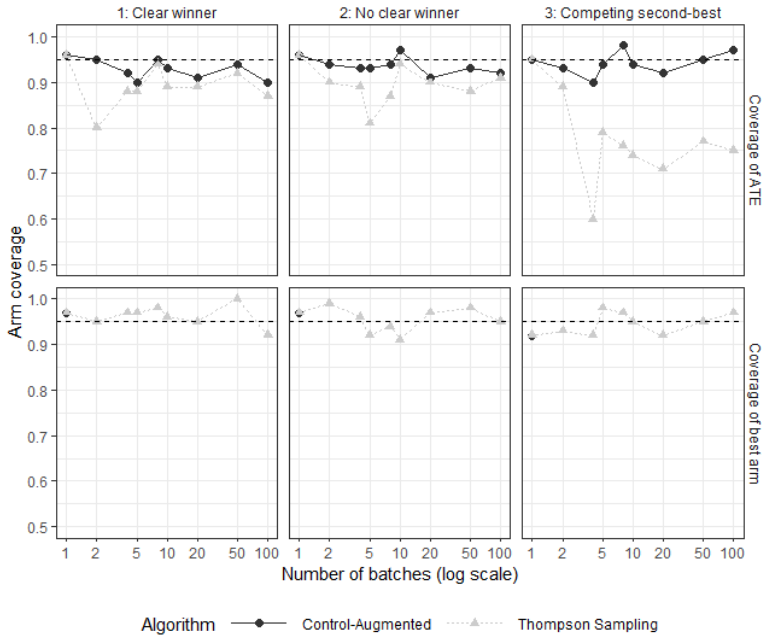


Figure 13: Root Mean Square Error by First Batch Size - 6 arms/1200 observations

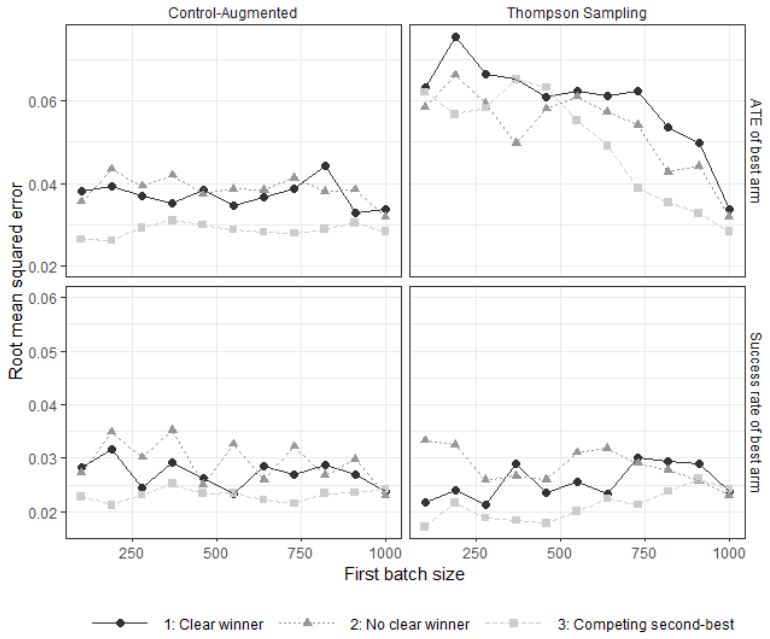
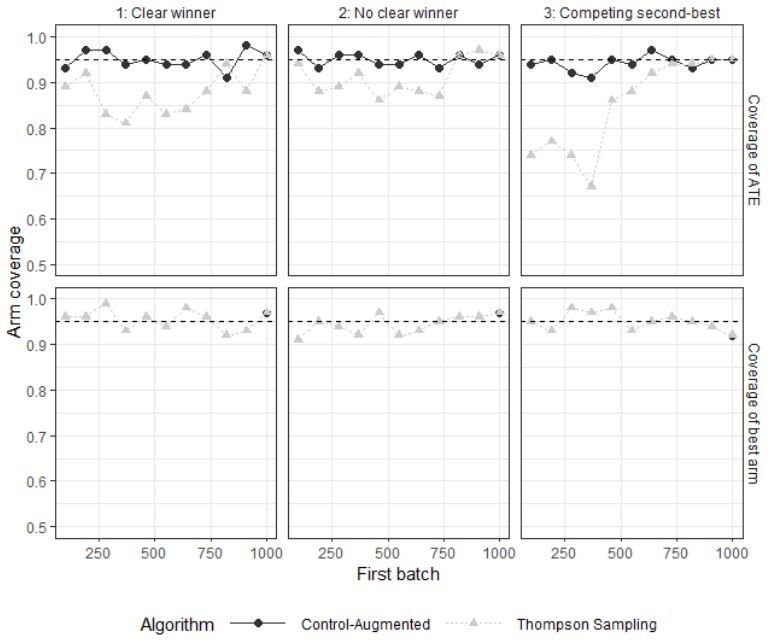


Figure 14: Arm Coverage by First Batch Size - 6 arms/1200 observations



Additional simulations We conducted a final round of simulations where the best performing treatment arm is the *Foregone Temporal* and the worst performing one is the *Standard Spatial*. In this case, we used a Thompson Sampling but set a probability floor of 10%. This means that all arms get at least 10% of the experimental units in each batch. 23 shows the results of an additional round of simulations using Kasy et al.'s exploration sampling. As we can see, although this algorithm slightly seeks more exploration across treatment arms, there is

still one arm with less than 100 observations. Finally, we report in Table 24 the mean of each using two weighting techniques: IPTW and Hájek-type stabilising weights.

Table 22: Results Thompson Sampling with a probability floor of 10%

Treatment	Successes	Trials	Mean
Foregone Temporal	60	288	0.208
Foregone Spatial	21	131	0.160
Individual Temporal	31	161	0.157
Individual Spatial	20	127	0.193
Standard Temporal	26	179	0.145
Standard Spatial	7	114	0.061

Table 23: Results Exploration Sampling Kasy et al's Exploration Sampling

Treatment	Successes	Trials	Mean
Foregone Temporal	29	182	0.159
Foregone Spatial	23	150	0.153
Individual Temporal	22	138	0.179
Individual Spatial	49	274	0.159
Standard Temporal	24	188	0.128
Standard Spatial	5	68	0.074

Table 24: Unweighted and IPW means Thompson Sampling

Treatments	Unweighted	IPW	Hajek
Foregone Temporal	0.208	0.204	0.208
Foregone Spatial	0.160	0.210	0.160
Individual Temporal	0.157	0.101	0.193
Individual Spatial	0.193	0.200	0.157
Standard Temporal	0.145	0.260	0.145
Standard Spatial	0.061	0.070	0.061

We can formalize how the Thompson sampling algorithm works as it reported in Offer-Westort, Coppock and Green and Russo expressed in equations 6 to 8. There are K treatments, and each arm k produces a successful outcome or reward with probability equal to θ^k and a failure with probability $(1-\theta^k)$. While we do not know the mean success rates for each arm k, this means that we cannot observe vector of θ ($\theta = (\theta^1, \theta^2 ... \theta^K)$), we can observe whether treatment k generates a success or a failure $x \in \{0, 1\}$. In the case of binary outcomes, we can set a prior that the probability of success for each treatment is equal to $FX|(\theta^k)$.

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta, a)p(\theta) \tag{6}$$

In period t, after units are assigned to treatments, outcomes are realised for each treatment arm k. We can also obtain the vector of responses for each treatment where $X_{n_t^k} = (X^k_{[1]} ... X^k_{n_t \cdot k})$, where n_t^k is the cumulative assignment to treatment k until period t included. Then, the

posterior probability in time t is given by the Bernoulli likelihood function and a beta prior distribution:

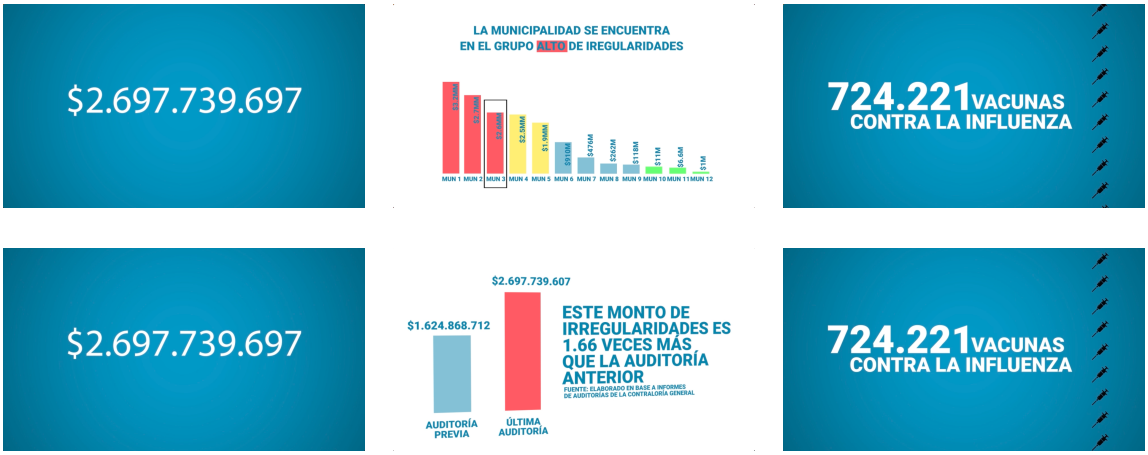
$$f_{\Theta|\theta_{n^k_{k,t}}}(\theta_k|x_{n^k_{k,t}}) \tag{7}$$

At the end of each period, we assign units to each arm based on their probability of being the best. For example, if the probability of treatment 1 being the best is equal to 0.2, in the following period, 20% of the experimental units will receive treatment 1. We can formalize the selection of the highest probability of success for each treatment as follows:

$$P\left[\Theta_k = \max_k\{\Theta_1, ...\Theta_k|(X_1^{n_{1,t}}, ..., X_K^{n_{k,t}})\right] \tag{8}$$

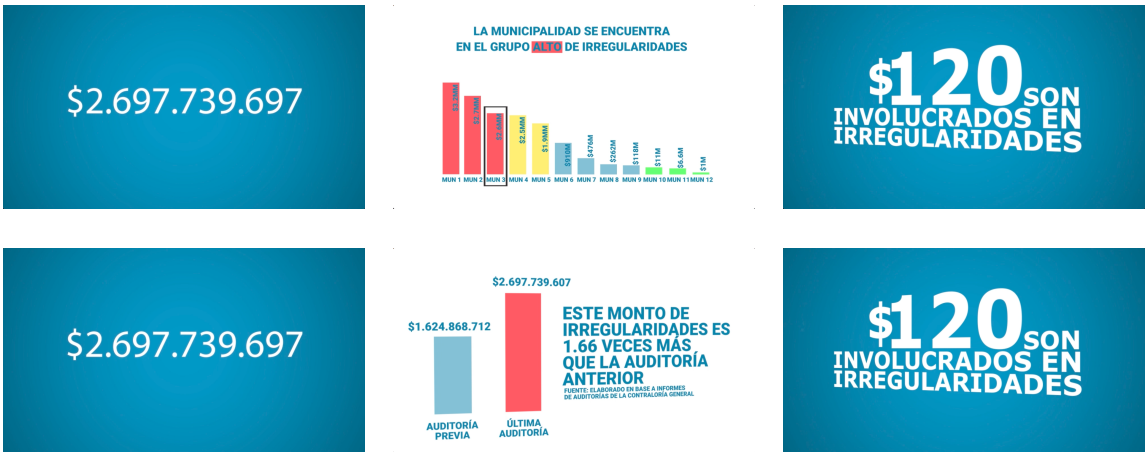
It is important to highlight that the posterior treatment assignment probabilities incorporates the uncertainty of the quality of each arm. This posterior probability is computed over the posterior distribution. Thus, if the posterior distribution has broad tails, the treatment assignment probability will be higher. Conversely, if the posterior distribution is clustered around its mean, the posterior treatment assignment probability will be smaller. We can formally state the steps of this sampling algorithm reported in Table 3.

Figure 15: Screen shoots - Foregone Treatment



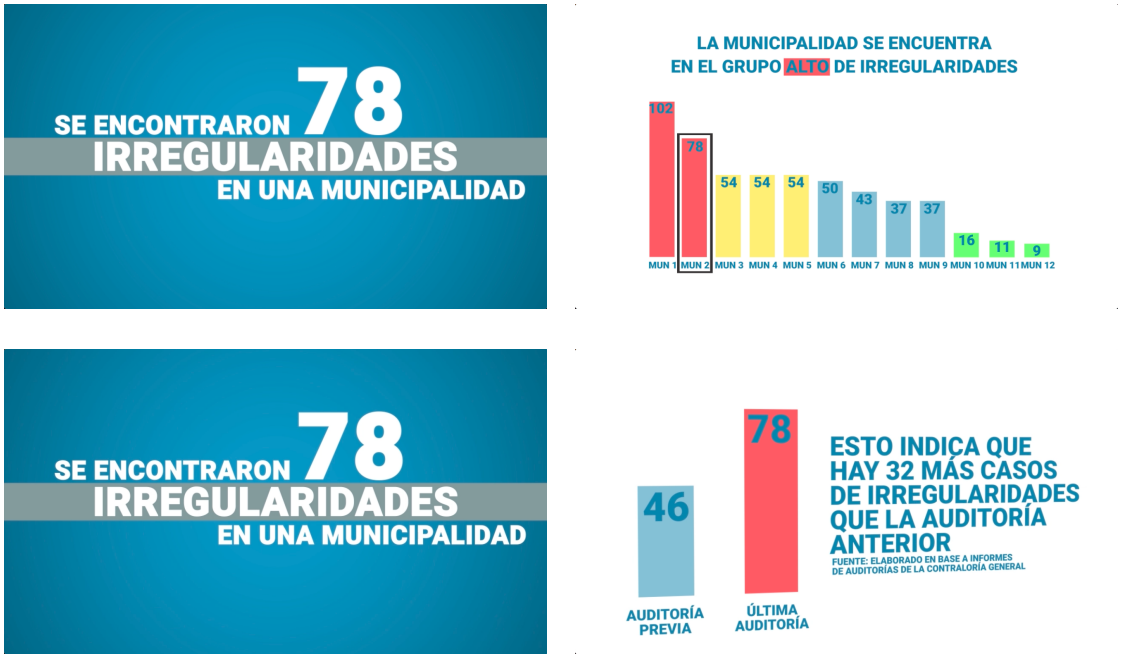
Note: This figure shows screenshots of the video information treatments that subjects received in this experiment. The top 3 images show screenshots of the *Foregone Spatial* treatment. The second row of screenshots corresponds to the *Foregone Temporal* condition. The first frame for both treatments reports the amount of malfeasance found in their municipality. The second screenshot on the top corresponds to the spatial bench-marking. The second frame on the bottom shows the temporal bench-marking. The third frame is the same for both treatments, and it reports the number of influenza vaccines that would have been bought with the money lost in corruption.

Figure 16: Screen shoots - Individual treatment



Note: This figure shows screenshots of the *Individual* spatial and temporal treatments. The top 3 images show screenshots of the *Individual Spatial* condition. The second row of screenshots corresponds to the *Standard Temporal* treatment. The first frame for both treatments reports the amount of malfeasance found in their municipality. The second screenshot on the top corresponds to the spatial bench-marking. The second frame on the bottom shows the temporal bench-marking. The third frame is the same for both treatments, and it reports the amount of malfeasance expressed as the amount out of every $1,000 the municipality spends.

Figure 17: Screen shoots - Standard treatment



Note: This figure shows screenshots of the video information treatments that subjects received in this experiment. The top 2 images show screenshots of the *Standard Spatial* treatment. The first frame for both treatments reports the number of irregularities in their municipality. The second screenshot on the top corresponds to the spatial bench-marking, which compares the number of irregularities across other local governments within the same region. The second frame on the bottom shows the temporal bench-marking, which compares the irregularities found in the last audit with those found in the previous audit.

Finally, in Figure 18 we reported the posterior probability of each arm being the best-performing treatment arm. WE conducted this analysis by making several modifications to the pilot. 1) We included only the four arms of interest 2) we split the data into equal-size batches of 100 observations. 3) we dichotomise the "convincing" outcome to simplify the posterior estimation of each arm. 4) we imposed an uninformative prior, where all arms have the same probability (1/24) of being the best arm. The results are in line with the results obtained in the previous models, where the Foregone Temporal arm shows an increasing posterior probability of being the best arm.

Figure 18: Posterior Probability of Best-Performing Treatment Arm

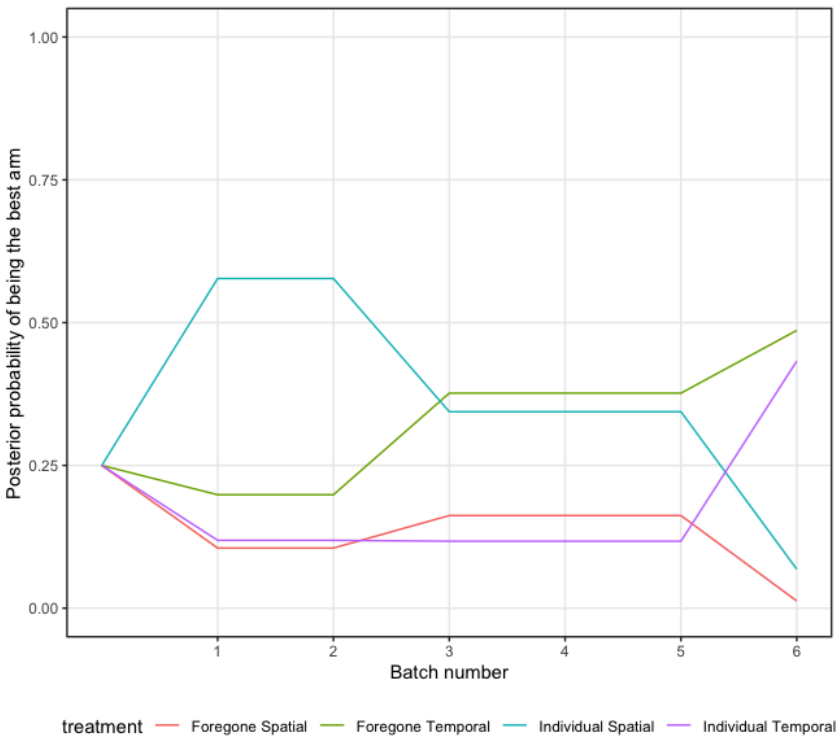


Table 25: Sample per Treatment Arm - Pilot

Treatment	Sample
Foregone Spatial High	170
Foregone Spatial Low	148
Foregone Temporal High	142
Foregone Temporal Low	165
Individual Spatial High	153
Individual Spatial Low	178
Individual Temporal High	166
Individual Temporal Low	154
Program Spatial High	175
Program Spatial Low	160
Program Temporal High	164
Program Temporal Low	168
Resources Spatial High	171
Resources Spatial Low	164
Resources Temporal High	164
Resources Temporal Low	170
Severity Spatial High	192
Severity Spatial Low	173
Severity Temporal High	169
Severity Temporal Low	149
Standard Spatial High	176
Standard Spatial Low - Baseline	179
Standard Temporal High	167
Standard Temporal Low	179

Multi-arm bandit algorithms

Scholars have proposed a wide range of algorithms that balances the exploration-exploitation trade-off in different ways.Westort et al put forward a *Control Augment Thompson sampling Algorithm.* The basic principle of this algorithm is that it modifies treatment assignment probabilities, so the cumulative samples of the control and the best-performing treatment arm are relatively similar. Table 26 outlines the steps of this algorithm.

Table 26: Algorithm Batch-wise Thompson Sampling - Control Augmented

Algorithm 2: Batch-wise Thompson sampling - Control Augmented
1: Iniatize priors such that $(\alpha_{k,1} = 1,\ \beta_{k,1})$ for $k = 1, ..., K$ 2: Calculate $p_{k,t} = P[\Theta_k = max\{\Theta_1, ...\Theta_K\}\|(\alpha_{1,t}, \beta_{1,t}), ..., (\alpha_{K,t}\beta_{K,t})]$ for $k = 1, ..., K$, excluding C 3: Retrieve the best arm at time t and calculate the difference between the cumulative sample assigned to the best arm (b) and the control arm $b = \text{argmax}_k\, p_{k,t}$ and $d = n_{b,t} - n_{C,t}$ 4: Calculate proportion (q) of the next batch needed for the control to match the cumulative sample of the best arm, setting a boundary $Z_t \in (0, 1)$. Find $q = min(max(d/n, 0), Z_t)$ 5: Calculate the probability of assignment to the control condition, for $k = 1, ..., K$: $\hat{p}_{C,t} = q + R_t * (1 - q)$ Where $R_t \in (0, 1)$ 6: Rescale posterior probabilities to the remaining sampling probability Compute $\hat{p}_{k,t} = p_{k,t} * (1 - R_t) * (1 - q)$ 7: Sample n observations using treatment probabilities in step 6. 8: Update posteriors, for $k = 1, ..., K$: $\alpha_{k,t+1} = \alpha_{k,t} +$ #success observed for arm k in period t $\beta_{k,t+1} = \beta_{k,t} +$ #success observed for arm k in period t

Similarly Russo (2020) proposes a *Top-two probability sampling* that adds a re-sampling step to the conventional Thompson sampling algorithm. This algorithm calculates the posterior probabilities of each treatment arm as the standard *Thompson sampling.* It then randomly assigns units to the best two arms with a probability β. This modification has several desirable properties, such as an exponential rate of posterior convergence[26].

[26]Slivkins (2020) provides an exhaustive analysis of the different sampling algorithms

Extension estimations

This section elaborates on two estimation and hypothesis testing strategies Batched OLS and AW-AIPW estimator.

BOLS estimator The authors formalize their BOLS estimator, as the difference in means of each arm for each batch $\Delta_t = \beta_{t,1} - \beta_{t,0}$ in time t, which is computed as follows:

$$\hat{\Delta}_t^{BOLS} = \frac{\sum_n^i (1 - A_{t,i}) R_{t,i}}{\sum_n^i 1 - A_{t,i}} - \frac{\sum_n^i A_{t,i} R_{t,i}}{\sum_n^i A_{t,i}} \tag{9}$$

In equation 9, $A_{t,i} \in \{0, 1\}$ represents a binary action chosen in batch period t. $R_{t,i}$ corresponds to the reward associated with each action i selected. On their approach, we can conduct the conventional hypothesis testing $H_0 : \Delta = c$ vs. $H_1 = \Delta = c$ using a t-statistics that is weighted combination of asymptotically independent normals.

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{\frac{\sum_n^i (1 - A_{t,i}) \sum_n^i A_{t,i}}{n\sigma^2}} (\hat{\Delta}_t^{BOLS} - c) \tag{10}$$

The basic idea of their estimation strategy is to compute an OLS estimator on each batch separately and then construct a Z-statistic for each wave and show multivariate normality. Their technique, in essence, takes advantage that the difference in means in each batch is asymptotically normal, and then we can weight the estimate by each batch.

AW-AIPW estimator We can compute the difference in means of this estimator ($\hat{\Delta}^{AW-AIPW}$) as follows:

$$\hat{\Delta}^{AW-AIPW} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}} Y_{t,1}}{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}}} - \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}} Y_{t,0}}{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}}} \tag{11}$$

In equation 11, $\sqrt{\pi^{(n)}}$ and $\sqrt{1 - \pi^{(n)}}$ are the variance stabilising weights, where $\pi^{(n)}$ is the probability of selecting each arm, conditional on the history $H_{t-1}^{(n)}$ of previous actions and rewards until $t - 1$.

The variance for this estimator is equal to $\hat{V}_0 + \hat{V}_1 + 2\hat{C}_{0,1}$ where:

$$\hat{V}_1 := \frac{\sum_{t=1}^{T} \sum_{i=1}^{(n)} \pi^{(n)} (Y_{t,1} - \hat{\beta}_1^{AW-AIPW})^2}{\left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}} \right)^2} \text{ and } \hat{V}_0 := \frac{\sum_{t=1}^{T} \sum_{i=1}^{(n)} \pi^{(n)} (Y_{t,0} - \hat{\beta}_0^{AW-AIPW})^2}{\left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}} \right)^2} \tag{12}$$

$$\hat{C}_{0,1} := \frac{\sum_{t=1}^{T} \sum_{i=1}^{(n)} \sqrt{\pi^{(n)}(1 - \pi^{(n)})} (Y_{t,1} - \hat{\beta}_1^{AW-AIPW})(Y_{t,0} - \hat{\beta}_0^{AW-AIPW})}{\left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}} \right) \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}} \right)} \tag{13}$$

Finally, by meeting some auxiliary conditions the student statistic is asymptotically standard normal:

$$\frac{\hat{\Delta}^{AW-AIPW}}{\left(\hat{V}_1 + \hat{V}_0 \right)} \xrightarrow[T \longrightarrow \infty]{d} \mathcal{N}(0, 1) \tag{14}$$

Figure 19: Screenshot survey

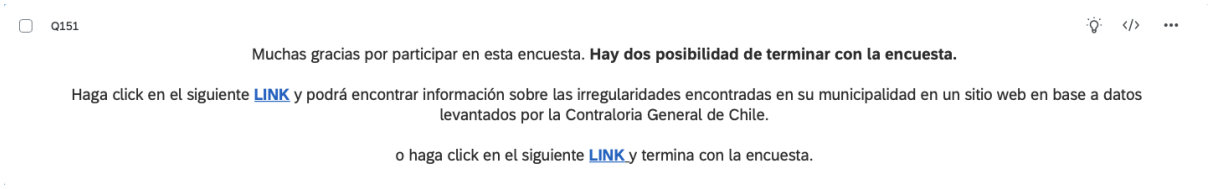
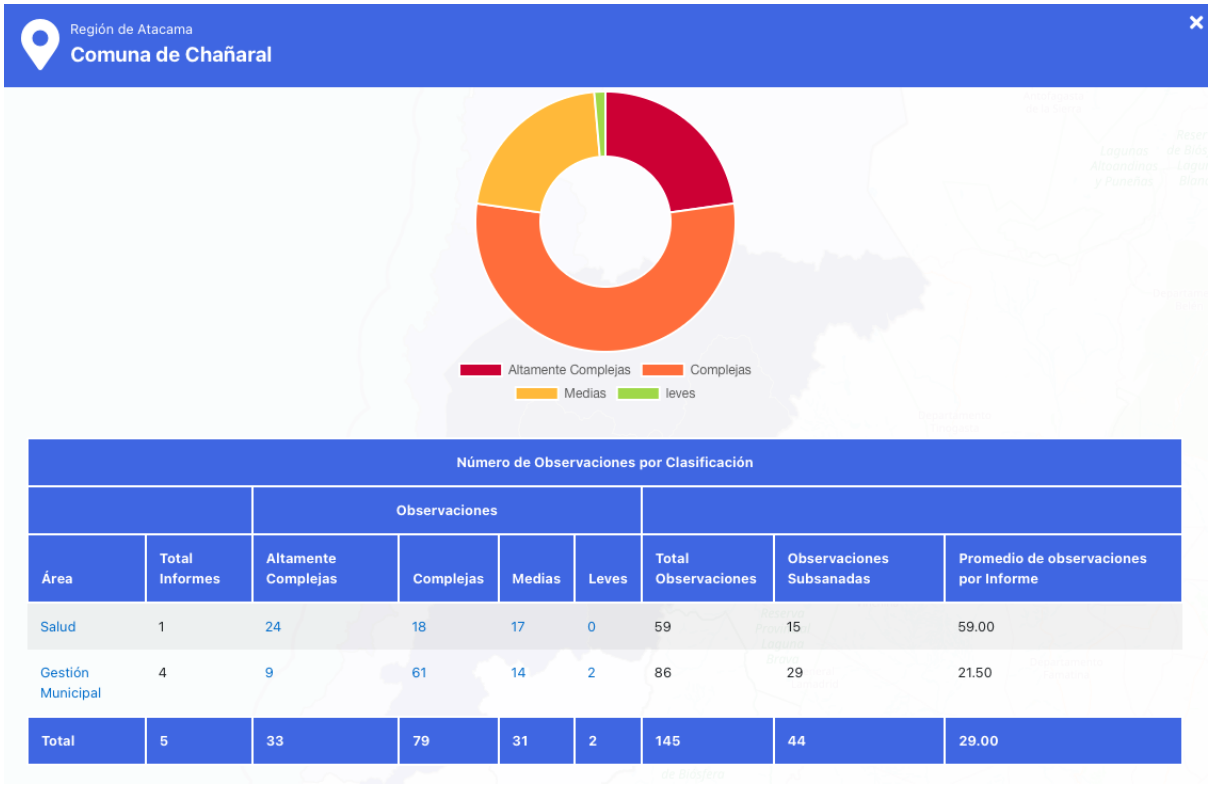


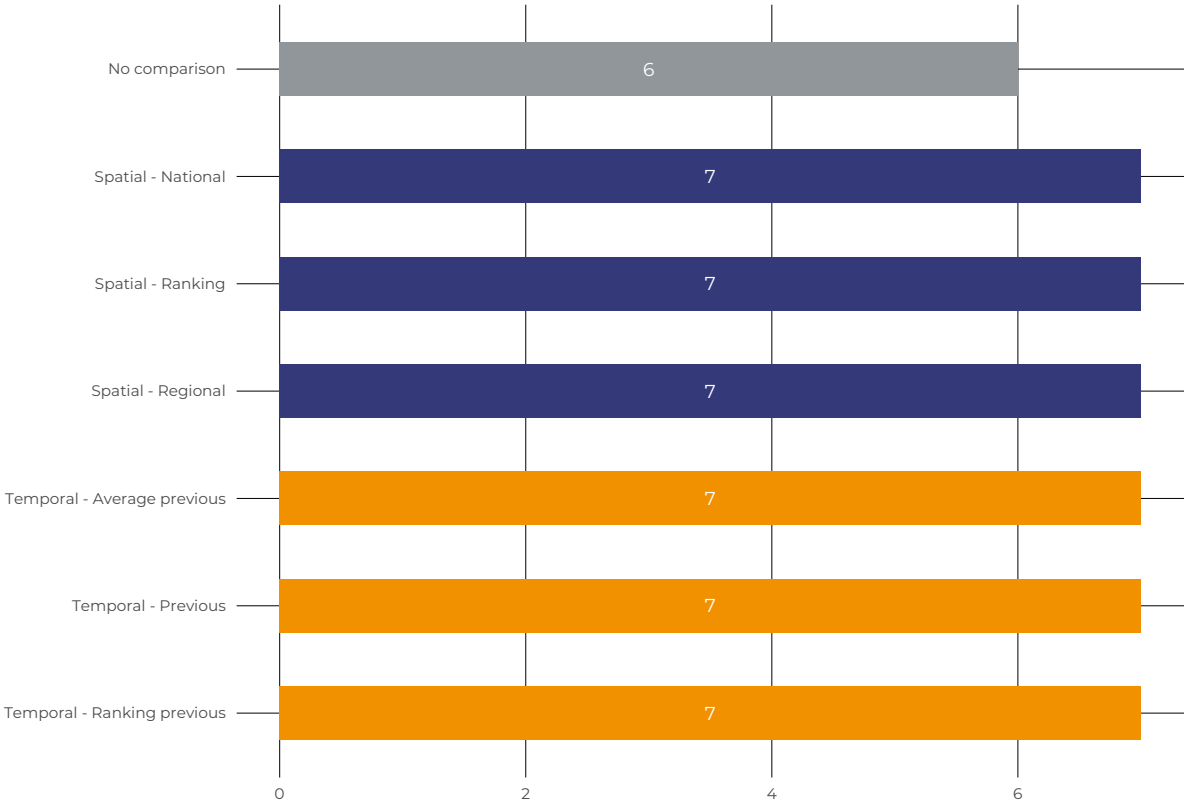
Figure 20: Screenshot survey

Figure 21: Screenshot Comptroller General Office's website



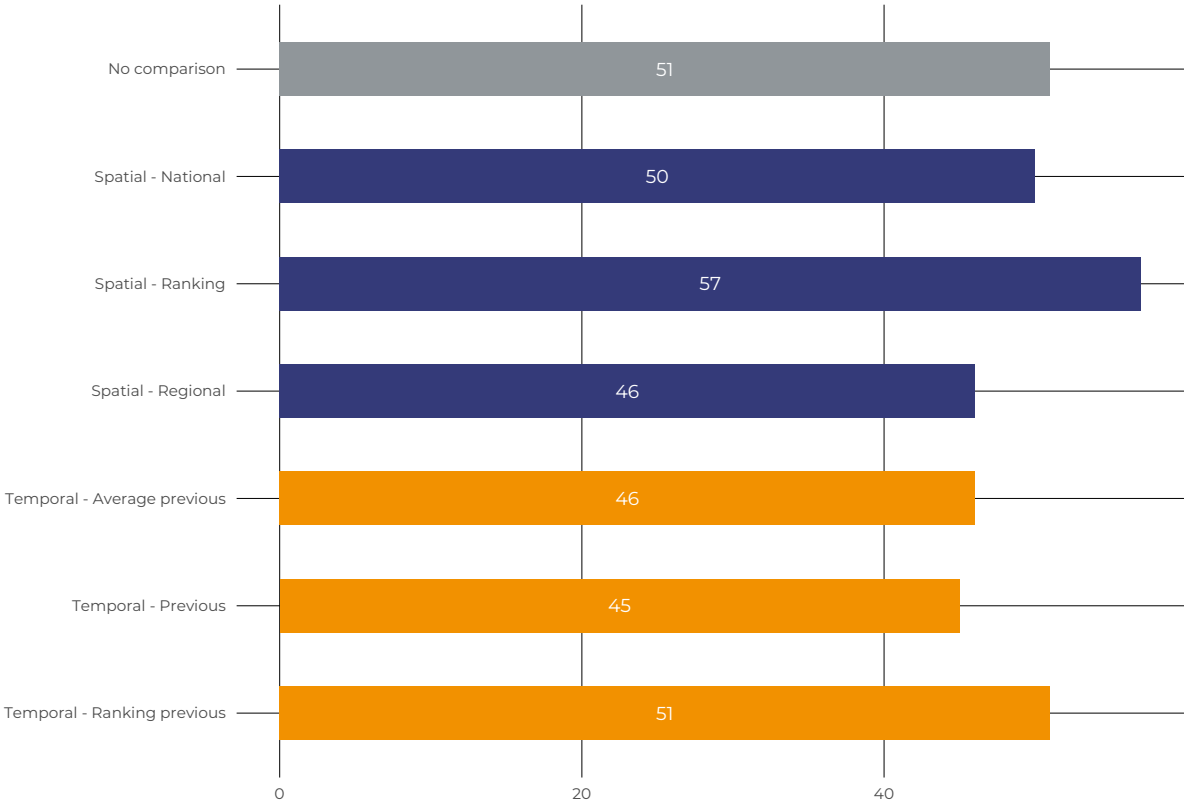
Número de Observaciones por Clasificación								
		Observaciones						
Área	Total Informes	Altamente Complejas	Complejas	Medias	Leves	Total Observaciones	Observaciones Subsanadas	Promedio de observaciones por Informe
Salud	1	24	18	17	0	59	15	59.00
Gestión Municipal	4	9	61	14	2	86	29	21.50
Total	5	33	79	31	2	145	44	29.00

Figure 22: Number of Answers by *Treatment Status*



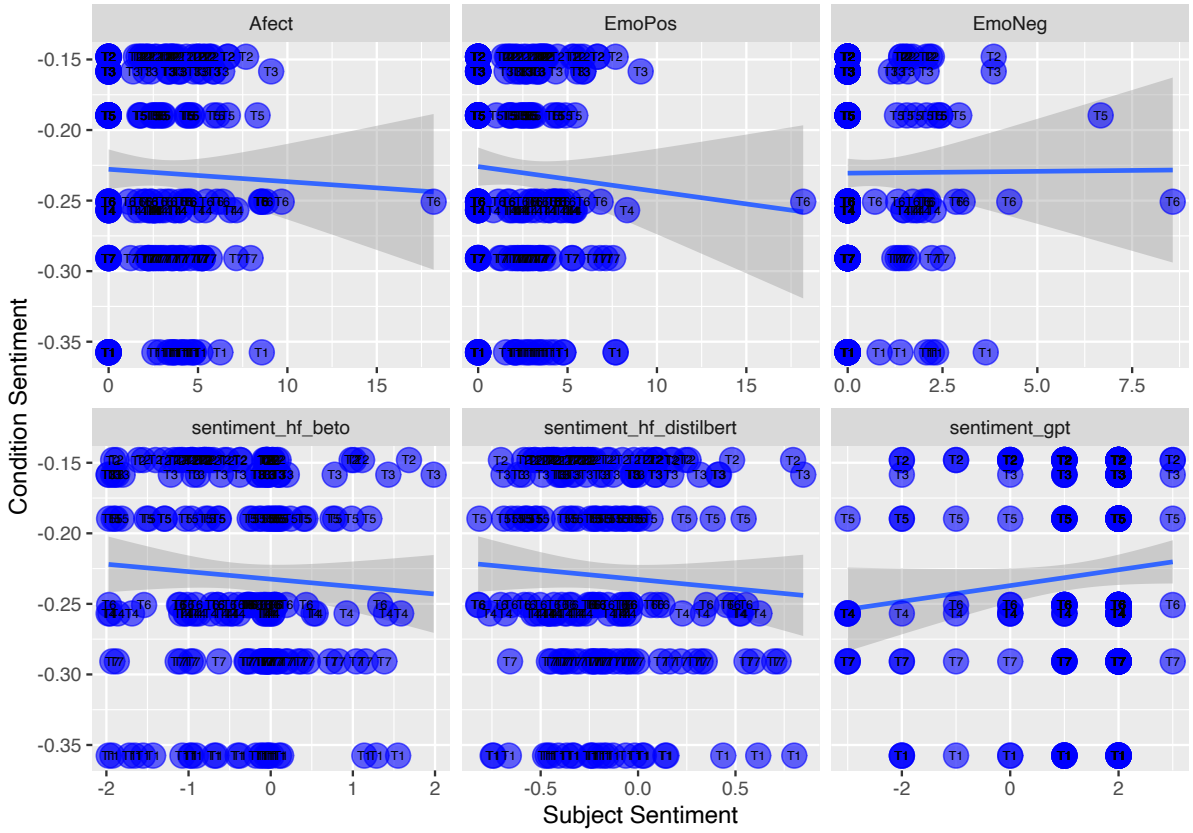
Question: Number of answers by treatment | sample size n=218

Figure 23: Lenght of answers by *Treatment Status*



Question: Number of words by treatment | sample size n=218

Figure 24: Message sentiment by *Treatment Status* each Individual



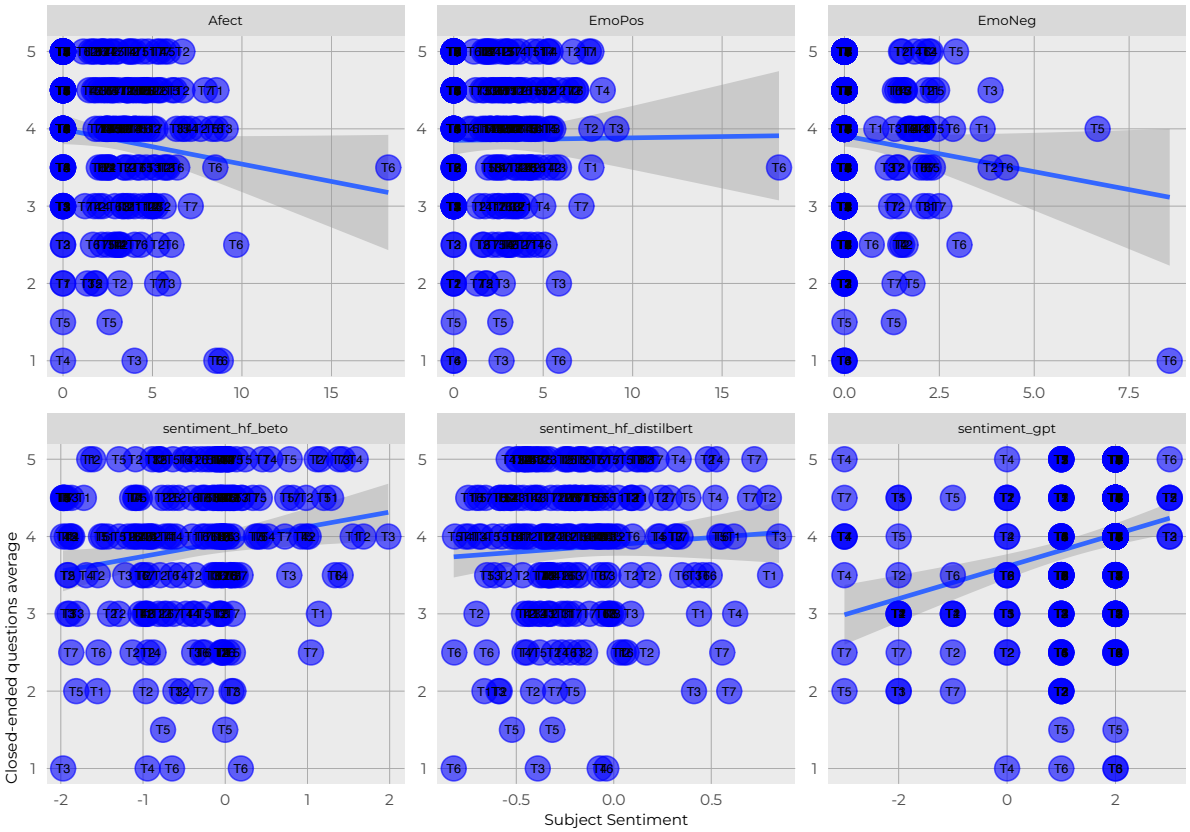
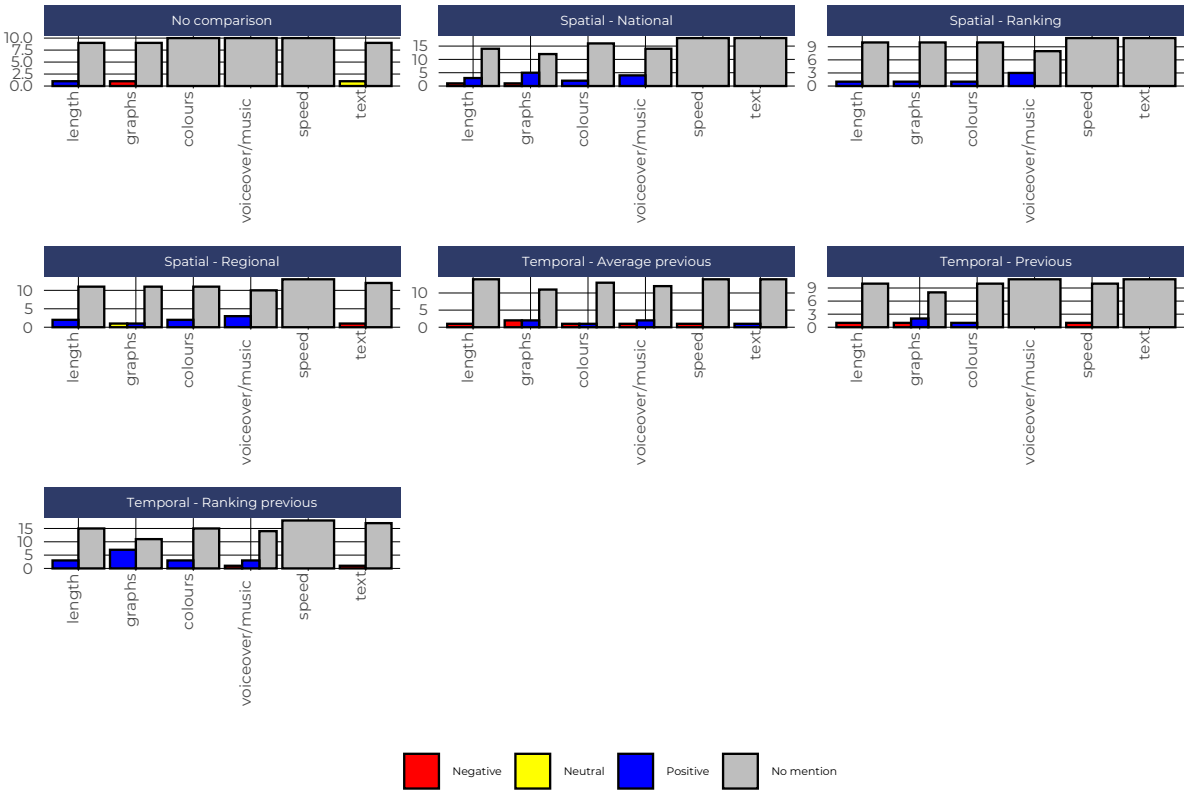
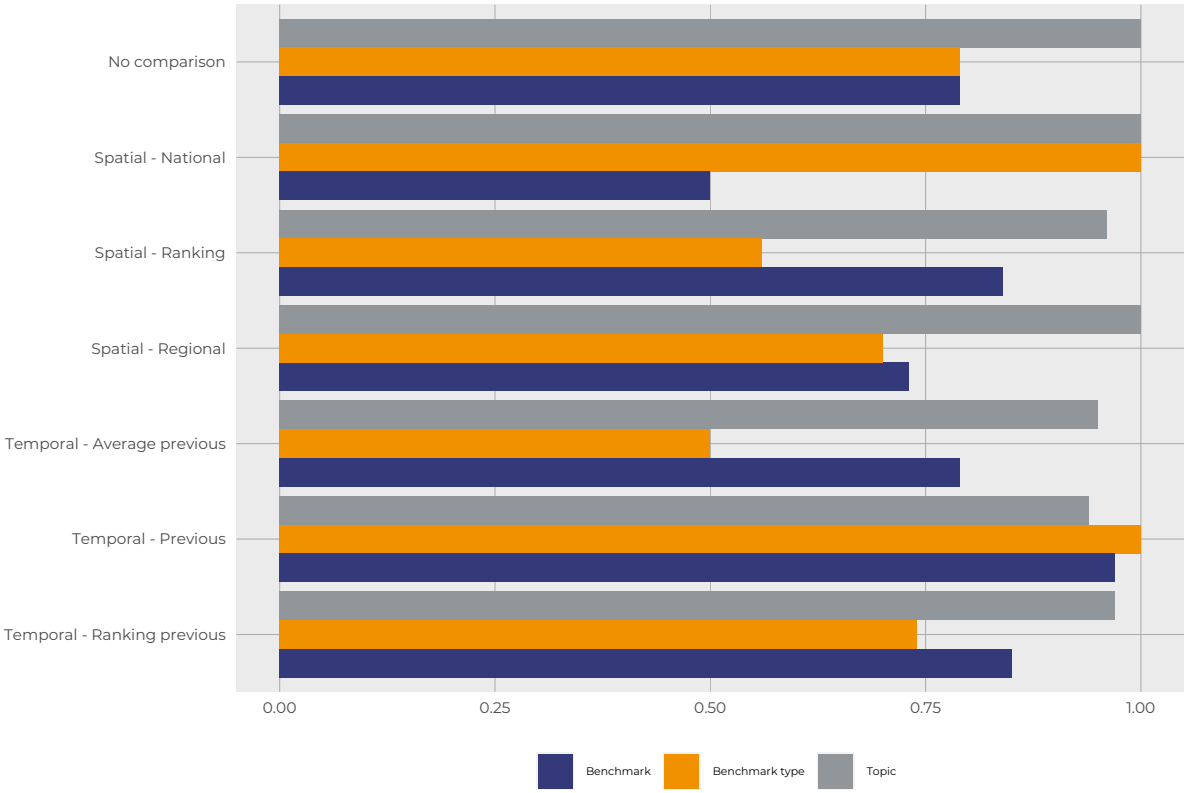

Afect
EmoPos
EmoNeg
sentiment_hf_beto
sentiment_hf_distilbert
sentiment_gpt
Closed-ended questions average
Subject Sentiment

Figure 25: Formatting issues by *Treatment Status*



Question: Formatting issues | sample size n=106

Figure 26: Attention checks by *Treatment Status*



Question: Attention check questions | sample size n=180