DPhil in Politics

Essays in Political Economy

Felipe Torres Raposo
Department of Politics and International Relations
University of Oxford
Pembroke College

Under the Supervision of:
Professor Raymond Duch
Dr Nelson Ruiz
Dr Konstantinos Matakos

June 27, 2023

Contents

1 Introduction **3**

1.1 Relationship between the three papers 4

1.2 Funding and Ethical Approvals 5

1.3 Disclaimer 5

2 Paper 1: Government Audits of Municipal Corruption and Belief Updating: Experimental Evidence **6**

2.1 Background of Audits and Context 9

2.2 Experimental Design 10

2.3 Estimating Belief Updating 14

2.4 Results of Malfeasance Information Experiment in Chile 20

2.5 Discussion and Conclusions 38

2.6 Appendix 42

3 Paper 2: Choice Architecture for Informing Corruption: Evidence from an Adaptive Experiment **70**

3.1 Motivation 70

3.2 Literature Review 73

3.3 Experimental Design 76

3.4 Treatment Effects Estimation and Inference 83

3.5 Results 84

3.6 Discussion 95

3.7 Appendix 98

4 Paper 3: Measuring Corruption Using RIRT **124**

4.1 Introduction 124

4.2 Randomised Responses 126

4.3 Survey Sampling, Instrument, and Design 130
4.4 Estimation and Inference 132
4.5 Results 135
4.6 Attention Checks and Estimation Adjustments 141
4.7 Comparison with direct question format from opt-in and representative samples . 146
4.8 Discussion 151
References 163

Chapter 1

Introduction

The following document outlines the research projects that I have conducted so far and would like to undertake in the upcoming academic year. The goal is to answer the following overreaching research questions:

What information is relevant to voters when it comes to getting information about malfeasance, and do voters update to 'informative' signals?

There has been a recent number of studies focusing on the relationship between information and accountability. Most recently, EGAP's Metaketa project [1]has shown that information tends to have little or no effect on party preferences or voting behaviour. Several explanations have been proposed to understand these findings. Furthermore, there is significant discrepancy between survey experiments and field experiments. This research project would seek to contribute to the field by providing empirical evidence on the different stages of the information chain.

This thesis is structured as a three-papers work. Paper 1 is a field experiment conducted in cooperation with the Comptroller General Office (CGO) in Chile and the local NGO Chile Transparente during the municipal elections held in May 2021. This paper looks at whether voters update in response to information about malfeasance at their local government. The study was conducted just before the municipal elections in 116 out of 345 communes in Chile.

Paper 2 focuses on the importance of informational architecture and how it affects the way citizens assess the reliability and trustworthiness of the information provided. A key component of Bayesian belief updating is that the signals should be informative and provide digestible quantitative information. This paper seeks to explore this premise by using an innovative methodological approach. I conducted an adaptive design experiment, which has rarely been implemented in the field of political science and compare its results with those of a conventional static experiment.

Paper 3 is a methodological paper that aims to assess the reliability of a Randomised Item Response Technique (RIRT) to measure the prevalence of corrupt behaviour at the local level.

[1]You can the results of the Metaketa here: https://egap.org/our-work-0/the-metaketa-initiative/round1-information-accountability/.

This paper is mainly motivated to explore unobtrusive and relatively new indirect survey techniques to measure the prevalence of corrupt behaviour. The purpose of conducting this technique is to yield both population and individual-level measures of experienced-based corruption. I also provide evidence of bias-corrected estimates due to inattentiveness. For the viva, I report population and individual-level estimates obtained from the RIRT.

1.1 Relationship between the three papers

There is extensive and recent research that has examined whether informing voters about corrupt behaviour of their local officials translates to into changes of electoral or non-electoral outcomes [2]. This research focuses on political and non-political measures such as party preferences, support for the incumbent, public goods contributions, and trust as the main outcomes of interest. I aim to contribute to the different stages of the chain that involves informing voters by focusing on belief updating and the informational architecture of the signal that citizens receive.

The first chapter focuses on the belief side, particularly on whether voters update their beliefs in response to the signal about malfeasance. Previous literature has shown mixed results on whether the information provided to voters is indeed informative or digestible, and how the signal interacts with prior beliefs. This paper looks at whether belief updating translates into some form of behavioural change on behalf of the voter. To this end, I focus on non-political outcomes such as costly contributions to public goods.

Paper 2 focuses solely on the signal. This paper aims to extensively scrutinise whether how the information about malfeasance is packed and presented to citizens change's respondents interest on this information. I conducted an adaptive experiment that shed some light on whether there is an optimal information strategy.

Finally, paper 3 looks at how corruption is measured. There is vast literature on social desirability [3], bias when it comes to measure the prevalence of sensitive issues such as race, religion, and corrupt behaviour. This paper aims to contribute to this evidence by providing findings of a relatively new extension of the Randomised Response technique that incorporates Item Response Theory. Furthermore, I expect to provide estimates of bias-corrected estimates due to inattentive respondents. Finally, I expect investigate on how robust this technique in yielding estimates both at the individual and group-level.

In summary, these three papers intertwine in their purpose of addressing some of the limitations identified in the existing literature. The first paper seeks to be an extensive exercise to find the frames and contents that are most informative and persuasive to the average citizen. The second paper takes a step back on the causal chain and examines beliefs as the main outcome. The logic here is that if voters are not updating, we should not expect changes in citizens' behaviour. Finally, a number of papers that have looked at informing citizens about malfeasance

[2]Some example of this research, you can find it her(Incerti, 2019; Larreguy, Marshall and Snyder, 2020; Grossman and Michelitch, 2018; Bhandari, Larreguy and Marshall, 2019)

[3]Some examples of this research can find it here (Blair, Chou and Imai, 2019; Wolter and Preisendörfer, 2013; Höglinger and Jann, 2018; Tourangeau and Yan, 2007; Ben, Jerke and Krumpal, 2012; Rosenfeld, Imai and Shapiro, 2016)

rely on survey-based measures of voting intentions, political preferences or experience-based corruption. However, survey-based measures could be imprecise and biased due to social desirability. Thus, the third paper seeks to address this limitation by providing a relatively new indirect questioning technique.

1.2 Funding and Ethical Approvals

This study is being funded by four main grants: 1) The Strategic Fund of the University of Oxford; 2) The National Research Centre in Chile (CONICYT); 3) the VIME Fund from the University of Santiago, Chile; and 4) the Nuffield Centre for Experimental Social Sciences. All these funds required ethical approval from different ethics committees. So far, I have received ethical approval from the Nuffield Centre for Experimental Social Sciences (Ethical approval number: OE_0048), the University of Oxford's Ethics Committee (Approval Reference: R68234/RE001), and the University of Santiago's Ethics committee (Approval number:037.2020). Given my previous affiliation with King's College London, I also obtained ethical clearance from King's College London (Approval number:LRS-19/20-14709)

1.3 Disclaimer

Professor Raymond Duch collaborated on Chapter 1. I am the guarantor of this paper and its principal researcher. I designed, implemented, and conducted the analysis of this study.

Chapter 2

Paper 1: Government Audits of Municipal Corruption and Belief Updating: Experimental Evidence

Extensive survey evidence suggests that, when asked, respondents, regardless of national context, express negative attitudes towards corruption in the public sector (Lagunes and Seim, 2021) and frequently question the honesty of their public officials (Chong et al., 2015). Hence the expectation that beliefs about malfeasance in government have important political consequences: Scandals have been shown to shape electoral accountability (Eggers, 2014); Grossman and Michelitch (2018) and Adida et al. (2022) implement field experiments with a job performance information frame; messages highlighting the quality of public service can affect accountability (Gottlieb, 2016). These beliefs about corruption in government may affect a broader range of behaviours. Beliefs about malfeasance by government officials might encourage corrupt behaviour. Ajzenman (2021), for example, estimates the impact of malfeasance information on civic values and dishonest behaviour in Mexico.

Citizens' beliefs about corruption are formed from diverse information sources. We assess the impact on these beliefs of a single source of information about corruption that is available in many countries; specifically, quantitative information about government malfeasance that is produced by audit agencies. There is a wide-spread global commitment to independent audits of government activities. An indication is the number of countries that have adopted these independent audits (OCDE, 2016) and the considerable aid funding that is directed at establishing independent audit agencies (Berliner and Wehner, 2022). There is an expectation that these independent audit reports generate credible information accessible to citizens that would in-turn promote democratic accountability (Weitz-Shapiro and Winters, 2017). A considerable body of literature explores the impact of audit reports on attitudes and political behavior although results vary: Some studies confirm the expected correlation between levels of reported malfeasance and electoral accountability (Larreguy, Marshall and Snyder, 2020; Ferraz and Finan, 2011); others have been less supportive of their impact on democratic accountability (Boas, Hidalgo and Melo, 2019; Adida et al., 2022; Chong et al., 2015); and some suggest they result

in the election of responsive but corrupt candidates (Bobonis, Cámara Fuertes and Schwabe, 2016). Context may be a conditioning factor that helps us understand these diverse findings (Breitenstein, 2019; Winters and Weitz-Shapiro, 2013)

Most of these studies assume that beliefs about corruption respond to the malfeasance information generated by audit reports. Our experimental design allows us to provide robust indications of how beliefs respond to the quantitative measures of malfeasance that are fairly typical of an audit agency report. An advantage of the design is that we can assess different models of how beliefs respond to informative and credible quantitative signals about the prevalence of corruption. We provide insights, for example, into whether updating reflects a Bayesian model, that focuses on the informative value of signals, as opposed to preference-based models that are less sensitive to the information content of signals (DellaVigna and Gentzkow, 2010). Another advantage of our design is that it allows us to assess how quantitative audit information might scale up as a policy tool for effectively informing citizens about corruption. A limitation is that we are not able to make claims as to the potential importance of this information, as opposed to other types of malfeasance information, in the overall assessments citizens have about corruption in government. The evidence we present here though suggests that the average citizen has little knowledge about this audit information and hence other factors currently play a much more important role in the formation of corruption beliefs.

Our experimental results indicate that the audit information treatments we designed can cause individuals to update their beliefs about malfeasance in their local government. Scaling this trial to the level of a national population would result in a meaningful improvement in citizens knowledge about malfeasance in local government. In the spirit of Finan and Mazzocco (2021) we argue that the cost of scaling up audit information should be included in efforts to assess optimal policy bundles for addressing corruption.

We present evidence from a field experiment in Chile specifically designed to test hypotheses that average citizens update their beliefs about municipal government corruption when presented with quantitative information generated by government audits. In partnership with the NGO Chile Transparente and the Comptroller General Office (CGO), treated subjects received a video, via WhatsApp, that summarised their municipality's actual audit results that were released by the CGO just prior to the experiment. The control group saw a placebo video. The effects of this information on the subjects' incentivised assessments of malfeasance in their municipal government is estimated one week and one month post-treatment. We estimated how much weight subjects assigned to objective audit information, relative to their prior beliefs about corruption.

Our subjects' priors regarding malfeasance in their municipal government weighed heavily in the formation of their malfeasance posteriors. These priors though are not particularly accurate. We compared participants' incentivised assessments of corruption in their municipal government with a composite malfeasance score calculated from historical annual audit data. There is at best a very weak correlation between the two.

Informing subjects about the audit results for their municipal government causes them to believe that there are higher levels of malfeasance in local government. Simply signalling to

subjects that there exists some non-zero level of malfeasance in their local government was sufficient to generate belief updating. This is consistent with very coarse preference-based models of belief updating.

Belief updating responds to quantitative details about local government malfeasance in a fashion consistent with richer models of belief updating. Belief updating is correlated with the amount of malfeasance reported in the audit. There is some evidence of Bayesian learning, i.e., subjects adjusting their priors to better accord with the reported prevalence of malfeasance in their municipal government. We report weak evidence that belief updating responds to a positive or negative audit report – there are expected differences but they are imprecisely estimated because of the relatively small number of cases in which audit results improved. Education is our proxy for cognitive abilities and we find that the better educated are more responsive to the quantitative details of the information treatment. Subjects who are partisans of the incumbent mayor have more positive priors regarding local government malfeasance and there is at best weak evidence that belief updating is conditioned on partisanship.

We designed the delivery of our experimental treatment and the belief elicitations so as to address frequent claims that information experiments elicit spurious reactions (Zwane et al., 2011). Pre-treatment measurement of priors was conducted four months prior to the intervention in order to minimise experimenter priming; the video intervention was delivered via WhatsApp rather than being embedded in a survey eliciting beliefs; posteriors were measured one week and then again one month after the intervention. Treatment effects persisted for a month after the intervention providing some assurance that the learning we do observe is not spurious.

In addition to the belief effect, we find that costly behavioural decisions also respond to our malfeasance information treatment. Our behavioural measure is a costly charitable contribution to the participant's municipal government that is solicited both pre- and post-treatment. This costly decision post-treatment responds to whether the municipality's malfeasance got worse (donations decline) or got better (donations increase).

These findings contribute to different strands of literature. They speak to the literature on how messaging about corruption affects political accountability; typically voting behaviour. Much of this literature focuses on estimating the causal effect of malfeasance information on important political outcomes and the findings are mixed. Important contributions such as those of Ferraz and Finan (2008), Chang, Golden and Hill (2010), Costas-Pérez, Solé-Ollé and Sorribas-Navarro (2012) and Arias et al. (2022) find that reporting information about mayoral malfeasance affects voting for incumbent candidates. On the other hand, recent field experiments find weak or null effects of malfeasance information treatments on incumbent vote (Adida et al., 2022; Boas, Hidalgo and Melo, 2019). Relatedly, Chong et al. (2015) find that negative information results in voter disengagement and lower voter turnout. Belief updating is one of the mechanisms that should contribute to the causal effect of malfeasance information on political behaviour. Hence we speak to this literature by helping understand how quantitative audit information causes corruption beliefs to update.

Our paper belongs to the political economy literature on persuasion that explores very broadly whether, and how, information can affect belief updating (DellaVigna and Gentzkow,

2010). Results from our study very much relate to the recent experimental evidence that quantitative details regarding resources windfalls can affect beliefs and participatory behaviours (Armand et al., 2020). Also of direct relevance to our study are efforts to understand whether political campaign messaging shapes voting behaviour (Coppock, Hill and Vavreck, 2020); how information about income distributions affect preferences for redistribution (Alesina, Stantcheva and Teso, 2018; Cruces, Perez-Truglia and Tetaz, 2013); and whether information frames on immigrants shapes redistribution preferences (Alesina, Miano and Stantcheva, 2018). A related contribution is to the extensive literature on how motivated reasoning, particularly when grounded in political partisanship, affects how individuals seek out and digest policy information (Peterson, Westwood and Iyengar, 2021; Mian, Sufi and Khoshkhou, 2021; Kunda, 1990).

We contribute to a recent literature that implements online experiments designed to measure the effects on beliefs of information about the macro-economy. Online experiments in which subjects are randomly assigned informative signals regarding the macro-economy have registered significant belief updating. Roth and Wohlfart (2020) find that information treatments consisting of professional forecasts strongly shifts macro-economic expectations towards the professional forecast and cross-sectional disagreement within the treatment arms declines. Similarly, Cavallo, Cruces and Perez-Truglia (2017) find that online inflation information treatments can move beliefs in the direction of these signals and also reduce the dispersion of expectations within treatment arms. Coibion et al. (2019) implement an online information experiment to assess the impact of inflation expectations on consumption. In a similar fashion, our online experiment explores belief updating by treating subjects with quantitative information about malfeasance levels in local government.

Our paper proceeds as follows. Initially, we present the experimental design followed by a discussion of how we estimate and model belief updating. A third section presents the results from the field experiment. We then conclude with a discussion of the implication of the results for ongoing studies of the public's beliefs about corruption.

2.1 Background of Audits and Context

The source of the information treatment in this experiment is the CGO (one of our implementing partners). Public opinion surveys commissioned by the CGO, research centres in Chile, and other independent governmental institutions confirm that the CGO has a reputation among citizens and civil servants as being non-partisan and reliable.[1] CGO audits have existed for over a decade; the CGO conducts annual audits of approximately 180 of the 345 municipalities in Chile. Audits usually take around 6 months. Each annual audit typically focuses on a different activity such as procurement, hiring processes or municipal finances. Audited municipalities are selected based on a scoring system that incorporates factors such as budget size, transfers to the private sector, and results from previous audits. This scoring system is not available to the public nor to municipalities. The factors and weights given to each factor change from one audit

[1]National Survey of Civil Servants and National Survey of Transparency, Council for Transparency, Chile. Survey waves from 2013 to 2020.

plan to another.

Subjects in our experiment are treated with municipal-level audit details that are actually generated by the CGO. The experimental results will indicate how citizens would respond *if* they received this information. This of course begs the question as to whether the average citizen would ever actually encounter these audit messages. In order to understand how these estimated treatment effects might scale up outside of the experimental context, we are interested in the likelihood that the average citizen would actually observe these quantitative metrics that are produced by audit agencies or reported on by NGOs such as Chile Transparente.

To answer this question we conducted a systematic analysis of media outlets that would likely inform the Chilean general public of the quantitative details in CGO audit reports. Social media is of course one of the venues in which we should be able to calibrate the attention citizens accord these municipal audit results. We analyse Twitter accounts from all 116 mayors and municipalities that were audited in 2020. First, we scrape the social media content from Twitter for the four month period preceding the release of the CGO audit reports and the four months following their release. For our modelling approach we used unsupervised Structural Topic Modelling. This modelling technique analyses a corpus of text (Tweets scrapped just prior and after the release of the audit reports) and identifies latent topics that best summarises the content of these online conversations. We identified that most of the discussion on Twitter focused on issues related to the COVID-19 pandemic, particularly vaccination and health. Figure 2.16 in the appendix reports the results from our structural topic modelling. We complemented this analysis by examining the term frequency and the inverse document frequency. This provides a measure of the relevance of words in a collection of documents. In our case, we set "documents" as the groups of municipalities based on their temporal frame (either negative or positive). We find that Tweets focused primarily on issues related to COVID-19. Results from this analysis can be found in the online appendix, Figure 2.15. Again, there is little indication of interest or knowledge of the Contraloria audit reports.[2]

We conducted the same analysis from a sample of 1058 respondents that shared with us their Twitter handle in the baseline survey. We find that in terms of the frequency of words, there is discussion related to national political figures and issues related to security, but no relevant mentions about corruption or audit reports. The results of this analysis can be found in the online appendix in Figure 2.15.

2.2 Experimental Design

Figure 2.1 provides a summary of the experiment design. In anticipation of the 2020 annual municipal audits we recruited a subject pool of approximately 57,708 subjects primarily via Facebook, Instagram and Twitter.[3] In January, 2019, the CGO selected 116 municipalities for the 2020 audit program. Our pool contained 46,723 residents in the 116 municipalities. We

[2]Note this is not consistent with the Google Trend analysis conducted by Ajzenman (2021) who finds spikes in searches on 'auditoria' corresponding with the release of Mexican audit reports.

[3]Full details on the recruitment campaign (including ads) are described in the Appendix.

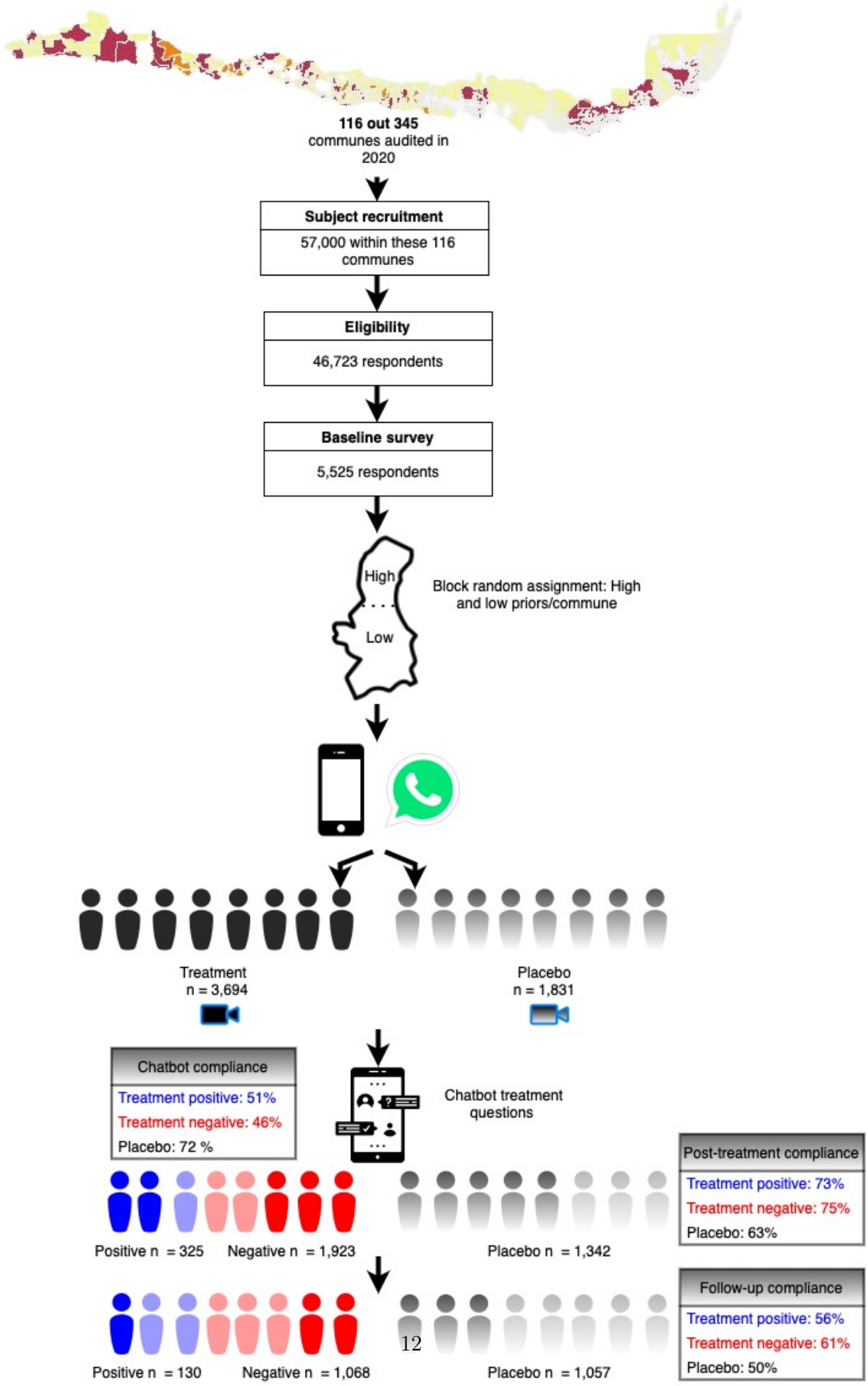
then conducted, in October-December, 2020, a pre-treatment survey with 5,525 participants, from our pool, who were resident in these 116 municipalities.[4] The Appendix benchmarks the demographics of the initial pool and final sample against the Chile census statistics. Our final sample has an age profile similar to the census; is more female; is significantly better educated; and is richer.

The CGO conducted their municipal audits during the months of January through December, 2020. As the CGO released audit results for each of the 116 municipalities, we produced a customised 45 second video for each municipality that summarised the reported irregularities.[5]. For each municipality we also produced a placebo video that provides neutral statistics on the municipality such as population and size of the commune. The table in Figure 2.1 presents the breakdown of treatment assignment.

[4]Subjects in the pool were given a window of more than 4 weeks to reply to the invitation to participate. Initially, 6,050 eligible respondents answered the baseline survey. A total of 5,525 respondents participated in the final study – we excluded 10 municipalities because the audit reports were not released on time by the CGO or were published very close to the election date. Hence, the effective response rate was about 25 percent.

[5]Videos for each of the 116 municipalities and details on the content are available on Dataverse

Figure 2.1 – Flow Chart of Experimental Design



Treatment assignment was conducted within each of the 116 municipality. Based on the pre-treatment survey we had a measure of each subject's malfeasance priors (based on both qualitative and quantitative measures of beliefs). Within each municipality, we blocked on high versus low malfeasance priors; subjects within each block were randomly assigned to either their municipal malfeasance treatment video or the placebo control video. This resulted in 3,694 assigned to treatment and 1,831 assigned to control.[6]

We observe our subjects making a number of decisions over the different phases of the experiment. Hence we actually have multiple indicators of whether the subjects were actually treated. These are summarised in Table 2.8. The treatment and control videos were sent to 5,525 participants via WhatsApp. Subjects received a short Chatbot survey after the video asking them attention check questions regarding the video content. A total of 3,190 subjects (57% of those treated) replied to the Chatbot survey. As we discuss in the Online Appendix, most of those who replied to the survey also correctly answered one of the attention check questions correctly. We define Chatbot Compliers as the approximate 56 percent of subjects who responded to the Chatbot survey and correctly answered the first and second attention check questions.

A post-treatment survey was administered to participants 3-6 days after they were treated with the videos. A total of 3,590 subjects responded to the survey: 2,249 out of 3,694 treated (61%) and 1,343 out of the 1,831 placebo subjects (73%). In the post-treatment sample, we asked subjects whether they recalled viewing the treatment videos: 846 (63 percent of those responding to the post-treatment survey) indicated they had viewed the Placebo video and 1,670 (74 percent of those responding to the post-treatment survey) indicated viewing the treatment videos. In Table 2.8, we define Post-Treatment Compliers as the approximately 45 percent of subjects, initially assigned to either placebo or treatment condition, who both answered the post-treatment survey and reported viewing the treatment videos. We will report intention-to-treat (ITT) results in the main text. The Online Appendix employs instrumental variable estimation to generate the complier average causal effects (CACEs) based on Post-Treatment Compliers (Sovey and Green, 2011).

Finally, a follow-up survey was administered one month after the video treatments with 1,198 out of 3,694 treated (32%) responding and 1,057 out of the 1,831 placebo subjects (58%) responding.

Figure 2.1 summarises the attrition registered over the four survey waves and also the rate of treatment compliance for the subjects in treatment, 3-6 day post-treatment and one-month treatment. In the Online Appendix, Table 2.9 reports the results of regressing subjects' treatment assignment (treatment versus placebo) on observed covariates, including both nested and full models. The initially assigned sample of subjects is balanced – only one of the covariate predictor variables, partisanship, is weakly significant. Table 2.9 also indicates that with rising attrition and declining compliance, the resulting samples remain balanced across the two treatment groups.

[6]The treatment assignment schedule for each municipality is available in the Appendix in Table 2.27

2.3 Estimating Belief Updating

In subsequent sections, we present reduced form evidence on how participants respond to the randomly assigned video information treatment compared to the placebo. We also present graphical comparisons of malfeasance beliefs for those in treatment and control. In this section we describe the specific elements of the malfeasance information treatment; we then propose different models of updating that inform our estimation of treatment effects.

The Corruption Signal Belief updating is sensitive to *what* and *how* individuals learn about government malfeasance. The experiment is designed to inform treated subjects of malfeasance audit outcomes with an effective messaging strategy – one that maximised the likelihood that subjects would learn about corruption in their local government. The information is delivered via WhatsApp in a short 45-second video because this is how many Chileans consume political information and also because we believe participants are more likely to engage with a video presentation of malfeasance information.[7]

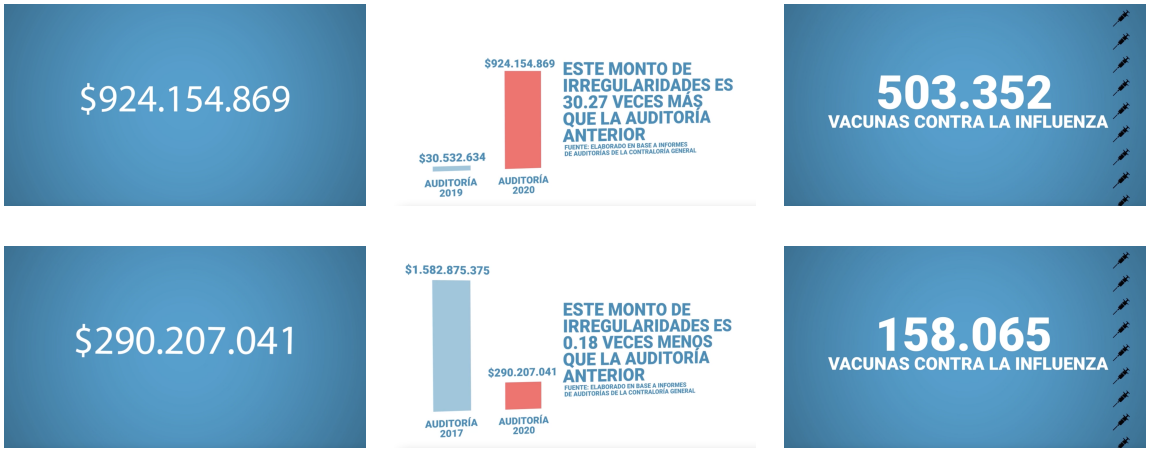
We adopt a novel strategy for determining what aspects of malfeasance should be included in this short 45 second video and how the information should be communicated. Existing research has adopted diverse strategies for designing corruption information treatments. For example the magnitude of malfeasance has been expressed in overall levels, in terms of percentages of the budget (Buntaine and Daniels, 2020) or as foregone expenditures (Arias et al., 2019; Larreguy, Marshall and Snyder, 2020). Bhandari, Larreguy and Marshall (2019), for example, find that "temporal" benchmarking (comparing current levels of malfeasance in a district performance with the district's previous levels) is more likely to affect updating than is the case with "spatial" benchmarking (comparing a district's malfeasance levels with those of other districts).

We explicitly built into our experimental design a strategy for identifying messaging that effectively signalled government malfeasance. Prior to producing the information treatment videos, we implemented a Chatbot experiment on WhatsApp that evaluated 24 different messaging strategies. Subjects were randomly assigned six different video messaging formats and asked to evaluate each of them. Results from this experiment suggested that a message identifying "foregone resources" and incorporating a "temporal" benchmark was the most effective framing of malfeasance information.

Figure 2.2 highlights the three core frames of the treatment video: First, respondents received information about the total amount of malfeasance found in the 2020 audit report. A second frame provided a temporal comparison of current reported malfeasance levels with those reported in the municipality's previous audit. This temporal comparison was depicted in a bar graph, allowing respondents to compare how levels of malfeasance have changed over the two audit reporting periods. The third frame of the video treatment expresses the malfeasance magnitude in terms of the number of foregone flu vaccines for the municipal population. The upper frame of Figure 2.2 illustrates an audit outcome that is negative, where we see an increase in the amount of corruption with respect to the municipality's previous audit. Conversely, in the lower frame is

[7]The baseline survey asked respondents about their WhatsApp usage – in particular their reliance on WhatsApp for political news and information.

Figure 2.2 – Screen shoots - Video treatment



Note: This figure shows screenshots of the video information treatments that subjects received in this experiment. The top 3 images show 3 screenshots for respondents that received a "negative" frame. The first frame reports the amount of malfeasance found in their municipality. The second screenshot shows a temporal comparison – in this case we see that corruption has increased over time in that municipality. The third frame on the top shows the "foregone losses", expressed in number of influenza vaccines. The three images on the bottom shows the same information, but for a municipality that decreased the amount of irregularities over time, which we call a "positive" frame.

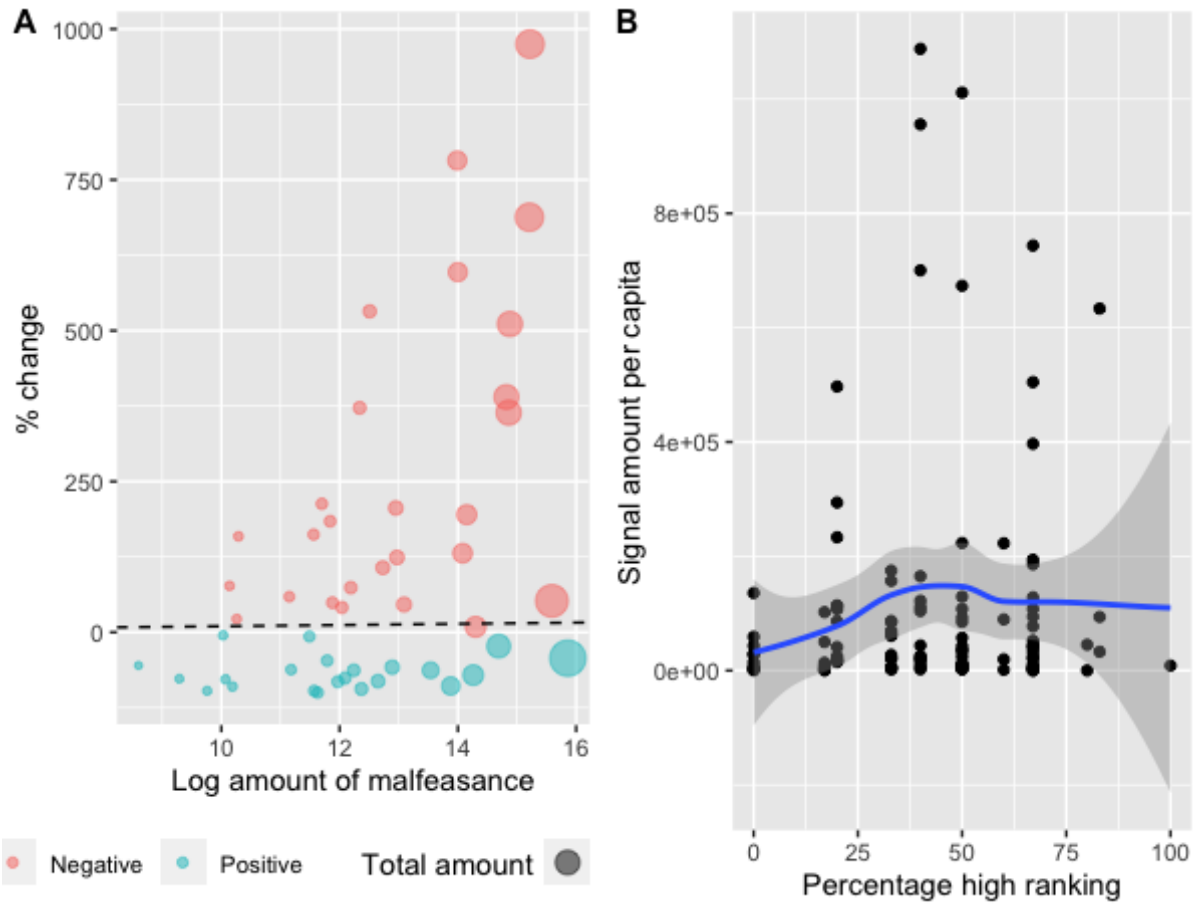
an example of a municipality with a positive audit result. In this case, the amount of corruption goes down with respect to the previous audit[8].

The information content of the videos is customised for each municipality so as to reflect the actual CGO audit results for the participant's local government. Figure 2.3A summarises two of the principal pieces of information conveyed to treated subjects. In the first frame of the video treatment, subjects are presented with the current 2020 values of audited malfeasance – this is summarised on the x-axis in Figure 2.3A as the logged value of the reported malfeasance expenditures. The second frame of the video treatment compares these 2020 malfeasance values to those reported in the municipality's previous audit - this is presented on the y-axis of Figure 2.3A as the percentage change in malfeasance magnitudes between the current and previous audit. Figure 2.3A plots percentage change against logged magnitudes. Negative values indicate a decrease in the levels of malfeasance and positive values an increase in irregular expenditures compared to the last audit. Malfeasance magnitudes vary significantly across municipalities: Changes in values from one audit to the next can be considerable and reported malfeasance is clearly more likely to increase than decrease.

Importantly, we make no broad claims about the corruption signal that our treated subjects observe. The treatment is a truthful summary of CGO's audit – as is the benchmarking information. We make no claims as to whether the audit report does, or should, characterise the overall extent to which a municipality is corrupt. In fact, a municipality's history of audit performance is not a particularly good predictor of how they perform in the 2020 audit report. Figure 2.3B

[8]Screenshots of the placebo video can be found in Appendix Figure 2.13

Figure 2.3 – CGO Audit Irregularities in Chilean Municipalities



Note: Figure A reports the relationship between the levels of malfeasance (log transformation) against the percentage change of the amount of malfeasance found in the audit report 2020 and the previous audit. Figure B reports the per capita amount of malfeasance found in each municipality (signal) that respondents received in the video treatments against the CGO's audit ranking scores.

plots the per capita magnitude of 2020 reported malfeasance expenditures against a metric of the municipality's overall audit performance between 2015 and 2020 – there appears to be no relationship between the two variables. It is also the case that individual priors regarding municipal corruption are uncorrelated with these audit metrics. To the extent that treated subjects recognise that there is a weak correlation between this signal and other metrics of malfeasance in local government we might expect them to discount its informative value (and possibly place much more weight on their priors). And this might in general characterise how citizens' beliefs about corruption respond to a typical audit report regarding local government malfeasance.

There are potential limitations to our randomisation protocol. The video treatments isolate three features of malfeasance that are expected to trigger belief updating in our outcome variables. A concern of course is that there are features of the video messaging that are both unmeasured and confounding (Fong and Grimmer, 2022). Secondly, the quantitative details reported in each of the video treatments are not randomly assigned. But as Figure 2.3B indicated though, there is little evidence of any systematic relationship at the municipal-level between historical audit performance and the 2020 audit reports. Individual priors regarding corruption

are uncorrelated with audit history and we find no evidence that a municipality's audit history conditions treatment effects.

Updating We observe subjects' beliefs about malfeasance levels in their municipal government pre-treatment, Corrupt_{ti}, and post-treatment, $\text{Corrupt}_{(t+1)i}$. For the treated subjects, our conjecture is that the audit information treatment will significantly affect posteriors – updated beliefs. Table 2.1 summarises the four models of belief updating we estimate with our experimental design. Each of the models characterises our hypotheses regarding signal extraction by the treated subjects. A vector of k covariates, X_{ki}, is included in each specification. In all specifications, we control for commune fixed effects represented by λ_c.

Table 2.1 – Models of Belief Updating

Model	Econometric specification
Baseline	$\text{Corrupt}_{(t+1)i} = \beta_0 + \beta_1 \text{Corrupt}_{ti} + \beta_3 T_i + \omega_k X_{k,i} + \lambda_c + \epsilon_i$
Malfeasance	$\text{Corrupt}_{(t+1)i} = \beta_0 + \beta_1 \text{Corrupt}_{ti} + \beta_2(\text{Log Malfeasance}_i) + \omega_k X_{k,i} + \lambda_c + \epsilon_i$
Negative	$\text{Corrupt}_{(t+1)i} = \beta_0 + \beta_1 \text{Corrupt}_{ti} + \beta_2(\text{Negative}_i) + \beta_3 T_i +$ $\beta_4(\text{Negative}_i \times T_i) + \omega_k X_{k,i} + \lambda_c + \epsilon_i$
Bayesian	$\text{Corrupt}_{(t+1)i} = \beta_0 + \beta_1 \text{Corrupt}_{ti} + \beta_2(\text{T}_i \times (\text{Corrupt}_{ti}^{True} - \text{Corrupt}_{ti})) +$ $\beta_3 T_i + \beta_4(\text{Corrupt}_{ti}^{True} - \text{Corrupt}_{ti}) + \omega_k X_{k,i} + \lambda_c + \epsilon_i$

At the most basic level, treated subjects learn that there exists malfeasance in local government. The treatment, T_i, is new information provided by the CGO about corruption levels in the subject's municipal government. This new information provides a signal of true corruption levels in the subject's municipal government, i.e., $\text{Corrupt}_{ti}^{True}$. Signal extraction proposed by the treatment effect in the *Baseline* model is very simple: Subjects who view any version of the information treatment from Figure 2.2 have resulting corruption posteriors that are higher than their priors. In this model, subjects respond to information that there exists malfeasance in their local government. The model specification assumes they do not extract much additional information from the signal.

Most efforts to model belief updating go a step further and assume subjects will respond to a quantitative metric. Recent experimental studies suggest, for example, that this is the case for information treatments that communicate consumer prices (Cavallo, Cruces and Perez-Truglia, 2017), public debt (Roth, Settele and Wohlfart, 2021), and immigration influxes (Grigorieff, Roth and Ubfal, 2020). In our case the metric is the level of malfeasance reported by the CGO in each municipality. Subjects in the second *Malfeasance* model in Table 2.1 are expected to respond to this quantitative information provided in the information treatment described in Figure 2.2. The *Log of Malfeasance* variable in the model is the log of the total malfeasance ex-

penditures identified by the CGO. In this case, subjects are expected to condition their updating on the magnitude of malfeasance reported to them in their video treatment (placebo subjects are "treated" with a magnitude of zero in this specification).

An additional refinement to the information treatment is that subjects learn whether malfeasance levels have risen or declined compared to their municipality's previous audit. Benchmarking the performance of elected officials is widely considered a critical element of responsibility attribution (Besley, 2006). Identifying the precise comparative frame that triggers belief updating related to malfeasance in government has been challenging. Information frames that compare the malfeasance of different elected officials across jurisdictions ("spatial" benchmarking) often have weak effects on updating (Arias et al., 2019; Campello and Zucco, 2016). There is though evidence that malfeasance belief updating responds to "temporal" benchmarking (Bhandari, Larreguy and Marshall, 2019). And as we noted earlier, our initial experiment exploring the persuasiveness of different malfeasance frames suggested temporal benchmarking is more likely to resonate with individuals than spatial benchmarks. In the *Negative* model in Table 2.1, we test this notion that belief updating responds to information comparing current municipal audit results to previous audits. A dummy variable, *Negative*, assumes a value of 1 if malfeasance levels have risen compared to the municipality's previous audit – zero otherwise. The expectation is that the information treatment effect should increase beliefs about corruption when *Negative* has a value of 1, i.e., the information treatment indicates malfeasance has risen.

Finally, we designed the interventions so as to discern whether subjects were using this new audit information provided in the video to update their malfeasance priors in a Bayesian fashion. Experimental evidence from Hill (2017) suggests that average citizens are not perfect Bayesian but that on balance Bayesian models are reasonable characterisations of how they process political information. A number of recent online experiments suggests that belief updating in response to information regarding macro-economic outcomes resembles Bayesian learning (Cavallo, Cruces and Perez-Truglia, 2017; Roth and Wohlfart, 2020). The term $(\text{Corrupt}_{ti}^{True} - \text{Corrupt}_{ti})$ measures the gap between subjects' priors regarding corruption and the true value of corruption in the municipal government. The extent to which belief updating is correlated with a gap between priors and actual corruption levels is captured by β_2 and β_4. This specification entertains the possibility that those in the control group, who do not receive an information signal, may update beliefs conditional on this gap between actual corruption and their priors. This is captured by β_4 and represents changes in beliefs in the control group that are conditioned on a gap between their priors and the audit information signal (which they do not observe). The β_2 coefficient indicates the extent to which this gap between priors and true corruption conditions the updating of treated subjects. The average treatment effect on subjects' corruption beliefs, not conditioned on this gap between priors and true corruption, is captured by β_3.[9]

A limitation to our specification of the Bayesian learning model is that pre-and post-treatment belief measures (Distribution and Resources) do not precisely reflect the malfeasance signals in the video treatment which consist of 1.) the value of malfeasance expenditures and 2.) the fore-

[9]This belief updating specification builds on Cavallo, Cruces and Perez-Truglia (2017). In our case there is random assignment to an information and placebo condition – hence the addition of the treatment dummy and treatment interaction.

gone resources associated with them. As we pointed out earlier, the information treatments (the signal, in effect) were based on the ChatBot experiments that identified the optimal messaging strategy. These were conducted after the original pre-treatment surveys (that occurred four months before the intervention) and did not precisely match the pre-treatment belief measures. Accordingly, we estimate a blunter version of Bayesian updating; one that likely underestimates the actually levels of Bayesian learning in the population.

Another version of this Bayesian learning model assumes that the impact of malfeasance information on updating is simply conditioned on whether individuals have relatively positive or negative malfeasance priors. Individuals with relatively negative priors regarding malfeasance in their local government respond to negative versus positive corruption signals differently than those with relatively positive priors. Evidence to this effect regarding citizen reactions to corruption audits in Mexico is reported by Arias et al. (2022). The effect of providing negative information about the malfeasance of a municipal government will be decreasing in individuals' prior beliefs that the government is malfeasant. Similarly, the effect of providing positive information would be decreasing in the individuals' prior beliefs that the government is not malfeasant. The reasoning here is that those with negative (positive) malfeasance priors may already anticipate audit results that report negative (positive) information about irregularities in their municipal government and hence there is minimal updating reflected in their posteriors.

Partisanship Information about local government corruption is decidedly political and will certainly trigger partisan considerations. There is an extensive literature documenting the extent to which partisan motivated reasoning colours how individuals process policy-related information (Peterson, Westwood and Iyengar, 2021; Iyengar et al., 2019; Mian, Sufi and Khoshkhou, 2021; Freeder, 2020; Kunda, 1990). This extends to corruption information. Most recent studies have documented the extent to which partisanship affects how individuals respond politically to information about government corruption (Solaz, Vries and de Geus, 2019; Anduiza, Gallego and Muñoz, 2013). Our focus is on one mechanism that might contribute to this partisan bias in electoral accountability for corruption. We argue that partisanship conditions belief updating when individuals receive new information about malfeasance by their local government. Alternatively, as Aarslew (2022) points out, partisans may not miss-perceive corruption information, other factors might simply outweigh this consideration in their vote utility function. Our conjecture has two implications for the results reported here. First, co-partisans of the governing municipal party will perceive the levels of municipal government malfeasance to be much lower than the levels perceived by opposition partisans. Second, our malfeasance treatment effect will differ for co-partisans versus opposition partisans.

Education The content of audit reports is decidedly quantitative – they quantify the extent to which public resources are expended in a corrupt manner. Processing this quantitative information can be cognitively demanding for the average citizen. There is evidence to suggest that the cognitively sophisticated segments of the population will be more discerning regarding the details of these audit reports (Weitz-Shapiro and Winters, 2017). Education is our proxy for cognitive facility with these quantitative details. For these quantitative elements of the mes-

saging, we conjecture that the malfeasance beliefs of the better educated subjects will be more responsive than those of the less highly educated (Hill, 2017).

Spurious or Genuine Learning Our experimental design aims to understand whether timely and accurate information about malfeasance results in meaningful changes in individuals' beliefs about local government corruption. The information treatments and belief outcome measurement are implemented as an online survey experiment. And while there are very persuasive reasons for scepticism regarding survey measures of malfeasance beliefs and their updating (Haaland, Roth and Wohlfart, 2020), there are features of our design that we believe minimise spurious belief updating. By measuring pre-treatment corruption beliefs four months prior to the information intervention we minimised biases that might be related to priming or anticipation (Zwane et al., 2011). The information treatment was a short 45-second video delivered via WhatsApp; a medium employed by most of the population for consuming political information. Belief updating was measured one-week and then one-month after the information treatment providing an indication of whether changes in beliefs persisted over a reasonable time period. Repeated, multiple, measures of the outcome minimised spuriousness and contributed to the precision of the estimations (Clifford, Sheagley and Piston, 2021).

Our outcome of interest is a quantitative assessment of municipal government malfeasance and the information treatment consists of quantitative measures of malfeasance. A concern with the design of these "statistical" information treatments is that these measures solicited from subjects after treatment will be subject to unconscious numeric anchoring (Roth and Wohlfart, 2020; Coibion et al., 2019). The design of the outcome measures addresses this concern: We include both qualitative and quantitative measures of beliefs regarding local government malfeasance. In addition, the incentivised quantitative outcome measures have scales that are not identical to the information treatments.

2.4 Results of Malfeasance Information Experiment in Chile

Variables and Measurement The outcome of interest is beliefs about malfeasance in municipal government. We implemented qualitative and quantitative measures of beliefs about government malfeasance along with a behavioural measure.[10] Table 2.2 presents the treatment and control group mean values and standard deviations for the outcome variables measured at each of the three survey waves. In the case of all three core outcome variables (Qualitative malfeasance, Resources, and Distribution) we observe that for treated subjects average values rise in the post-treatment wave and then decline somewhat in the third follow-up wave. And as expected the average values of subjects in the control treatment are similar across all three waves.

The qualitative measurement question adapted from Enrique et al. (2019) was: 'How much

[10] A number of recent studies have highlighted the importance of robust measurement strategies in order to avoid experimental measurement error (Duch, Laroze, Reinprecht and Robinson, 2020; Engel and Kirchkamp, 2018; Gillen, Snowberg and Yariv, 2019; Haaland, Roth and Wohlfart, 2020; Broockman, Kalla and Sekhon, 2017).

of the municipal government expenditures are corrupt?' (10 point scale from none to all). It's average value for treated subjects rises from 4.49 in pre-treatment to 5.34 post-treatment and then declines somewhat to 5.15. We also asked respondents to indicate how certain they were about their response to qualitative question.

Table 2.2 – Descriptive mean values for malfeasance belief outcome measures and covariates

	Survey Wave					
	Pre-Treat		Post-Treat		Follow-up	
	Control	Treatment	Control	Treatment	Control	Treatment
Qualitative Measures						
Subjective malfeasance scale 1-10	4.47	4.49	4.60	5.34	4.39	5.15
	(2.44)	(2.44)	(2.35)	(2.34)	(2.30)	(2.29)
Certainty subjective malfeasance scale 1-10	5.44	5.39	5.28	5.74	5.46	5.52
	(2.72)	(2.70)	(2.72)	(2.59)	(2.72)	(2.59)
Quantitative Measures - incentivised						
Resources	3.83	3.85	3.78	4.57	3.52	4.22
	(2.36)	(2.39)	(2.36)	(2.60)	(2.22)	(2.48)
Distribution	1.90	1.91	1.93	2.29	1.89	2.19
	(0.78)	(0.79)	(0.79)	(0.87)	(0.78)	(0.84)
Temporal	-	-	4.75	5.92	4.46	5.45
	-	-	(1.98)	(2.34)	(1.92)	(2.20)
Behavioral Measures						
500 CLP Municipal Donation	32%	33%	17%	15%	11%	10%
Covariates						
Female	64.79%	64.69%	64.78%	64.61%	64.81%	66.61%
Education (Yrs)	15.09	15.05	15.10	15.04	15.06	15.03
Age	36.42	35.61	36.41	35.37	36.45	35.17
Partisanship (1 to 10)	4.55	4.43	4.58	4.38	4.58	4.31
Political alignment	28%	27%	28%	26%	28%	26%
Income (USD)	1003	904	853	921	793	759
Sample	1832	3696	1343	2249	1057	1198

Note: Mean values are reported with standard deviations in parentheses. The *Resource* measure is the quantity out of $1,000 subjects believed is lost due to corruption by their municipality. The ordinal categories vary from 0 to 10. The *Distribution* variable is subjects' reported ranking of their municipality in terms of levels of malfeasance: Low (1); Moderate (2); High (3); and Very High (4). The *Temporal* variable has values from 1 to 9, 1 being equal to "decreased more than 60%" to 9 being equal to "Increased more than 60%". We included the means of covariates used in our estimations. We reported values for each of the waves broken down by treatment status.

As Table 2.2 indicates, we included two incentivised measures of corruption beliefs. Subjects are asked to rank their municipality in terms of audited malfeasance. They are presented with the actual distribution of Chilean municipalities in four categories (see the Online Appendix for the survey question). For each subject we have an ordinal score ranking: 1 (low); 2 (moderate); 3 (high); and 4 (very high). Participants earn 300 CLP if they correctly categorise their municipality. For treated subjects, in pre-treatment the average Distribution score is 1.91, just below moderate malfeasance, rising to 2.29 in the initial post-treatment survey and then declining to

2.19 in the follow-up.

A second incentivised quantitative measure asks participants to indicate how much out of every 1000 CLP municipality expenditures is associated with corrupt activities. Participants earn an additional 300 CLP if they get the correct answer. This Resource measure has been converted to ordinal scores ranging from 1 ("Less than 100 CLP") to 2 ("Between 100 and 200 CLP) until 10 ("More than 900 out of 1000 CLP"). For treated subjects, in pre-treatment the average Resource score is 3.85; in post-treatment the average increases to 4.57, but it declines in the follow-up to 4.22.

An additional incentivised question was included in the post-treatment and follow-up waves. We asked participants to estimate the percentage change in corrupt expenditures between the municipality's current audit and its previous audit. This question had 9 categories, from 1 (decreased more than 60%), 5 (Stayed the same), 9 (increased more than 60%). For treated subjects, the average score on the post-treatment survey was 5.92 and then decreased slightly on the follow-up survey to 5.45.

We include a measure of whether subjects would contribute to the provision of a local public good (Bonica, 2018; Grigorieff, Roth and Ubfal, 2020; Roth, Settele and Wohlfart, 2021). Subjects are asked whether they would donate 500 CLP (about 15 percent of their earnings). In the initial wave, about one-third of treated respondents donated 500 CLP – this drops to 15 percent in the post-treatment survey for treated subjects and then drops to 10 percent.

In terms of key covariates, we do not see substantial changes across waves. The exception is average individual monthly income that drops from US$853 and US$921 in post-treatment to $793 and $759, respectively, in the follow-up survey. There are relatively minor differences between treatment and control subjects. Our sample is disproportionately female. Most of our sample had completed primary and secondary education and had completed some form of tertiary education. The mean partisanship across all wave, and for both treated and control subjects, is Left-leaning with a score of 4.4 measured on a 1 to 10 Left-Right scale. We define a dichotomous partisan alignment variable that assumes a value of 1 if the subject indicated they will vote for the incumbent mayoral candidate in the upcoming municipal election and a 0 otherwise (*Aligned*). Approximately one-quarter of the respondents indicated they would vote for the incumbent mayor.

The Chilean public's priors regarding malfeasance in their local government are uncorrelated with CGO audit reports. Based on our pre-treatment survey of the 116 audited municipalities in Chile, we calculate an average for each of the three malfeasance outcome measures (these are based on a median sample size per municipality of 47 respondents).[11] For the past 6 years of CGO audits we generated the average malfeasance score for each of these 116 municipalities (the median number of municipalities that are audited by the CGO in any year is 226 out of a total of 356). Figure 2.12 in the Online Appendix plots each of the three average malfeasance scores against the actual average CGO audit results for each municipality. Our three outcome variables measuring malfeasance priors are decidedly uncorrelated with the historical levels of municipal malfeasance reported by the CGO. This is consistent with Arias et al. (2022) who

[11] 4.48 for the Qualitative measure, 384 for the Resource outcome, and 53.90% for the Distribution outcome.

Table 2.3 – Regression results of Malfeasance Beliefs - All Outcomes

	Qualitative			Resources			Distribution		
	Baseline	Negative	Malfeasance	Baseline	Negative	Malfeasance	Baseline	Negative	Malfeasance
Intercept	2.867***	5.861***	2.888***	3.354***	5.862***	3.368***	1.301***	2.043***	1.302***
	(0.392)	(0.551)	(0.393)	(0.679)	(0.739)	(0.685)	(0.142)	(0.174)	(0.142)
Prior	0.406***	0.406***	0.406***	0.412***	0.412***	0.412***	0.320***	0.321***	0.321***
	(0.019)	(0.019)	(0.019)	(0.018)	(0.019)	(0.019)	(0.021)	(0.021)	(0.021)
Treat	0.733***	0.530*		0.728***	0.204		0.342***	0.085	
	(0.080)	(0.233)		(0.091)	(0.186)		(0.029)	(0.069)	
Negative		-3.009***			-2.545***			-0.763***	
		(0.314)			(0.281)			(0.105)	
Treat x Negative		0.238			0.613**			0.300***	
		(0.249)			(0.207)			(0.075)	
Malfeasance IV			0.051***			0.051***			0.024***
			(0.006)			(0.006)			(0.002)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	3439	3439	3439	3439	3439	3439	3439	3439
R2	0.319	0.319	0.319	0.268	0.270	0.269	0.280	0.284	0.283
R2 Adj.	0.290	0.290	0.290	0.237	0.238	0.238	0.250	0.253	0.252

Note: Regression estimates are reported for the *Baseline, Negative* and *Malfeasance* models listed in Table 2.1 for all three outcomes. *Malfeasance IV* is the instrumented variable for the log malfeasance reported in the video. The *Negative* variable is a dummy variable that takes value equal to 0 for respondents in municipalities where corruption decreased over time, and 1 if respondents where informed that corruption increased since the last audit. We clustered standard errors at the commune level. For all specifications, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

find no correlation between malfeasance priors and municipal-level malfeasance reported by the Mexico ASF.

Information Matters Treated subjects observe a 45-second video indicating there are malfeasance expenditures occurring in their municipal government. The information treatment also includes quantitative details about actual levels of malfeasance expenditures in their local government. Treatment effects are estimated using linear regressions that include pre-registered pre-treatment covariates in order to increase precision. Random assignment is conducted within municipalities and hence the standard errors are clustered at the municipality level.

Regression estimates of the treatment effects are estimated in Tables 2.3 for the Qualitative, Resources and Distribution outcome variables for the *Baseline*, *Negative*, and *Malfeasance* models that were presented earlier in Table 2.1.[12] Across all the model specifications and for all outcome measures, posteriors are strongly correlated with the subjects' priors regarding malfeasance. Our *Baseline* model only includes a control for Priors and a dichotomous Treat variable. Table 2.3 presents estimates for this simple model of updating that excludes the quantitative details of the information treatment. Across all three outcome measures, Qualitative, Resources and Distribution, the treatment coefficient is large and statistically significant.

As Figure 2.3 indicated, treated subjects received two quantitative details about malfeasance.

[12]As we have pointed out, there is some attrition that occurs over the survey waves. This of course could result in "missingness" that is not at random and potentially correlated with the treatment arms. The Online Appendix has a section that addresses the implications of attrition bias for the results we report in Table 2.3. We implement a number of different strategies that have been suggested for addressing attrition bias. In the cases of each of these different estimation strategies, we estimate treatment effects that are very similar to those reported in Table 2.3

In the second frame of the video these subjects received a temporal comparison of their 2020 municipal audit results with the municipality's previous results. Of the total 116 municipalities in the study, 94 registered worse malfeasance audits in 2020 than in their previous audit; 22 of the municipalities had improved outcomes. This is the basis for categorising subjects into a Negative versus Positive temporal frame.

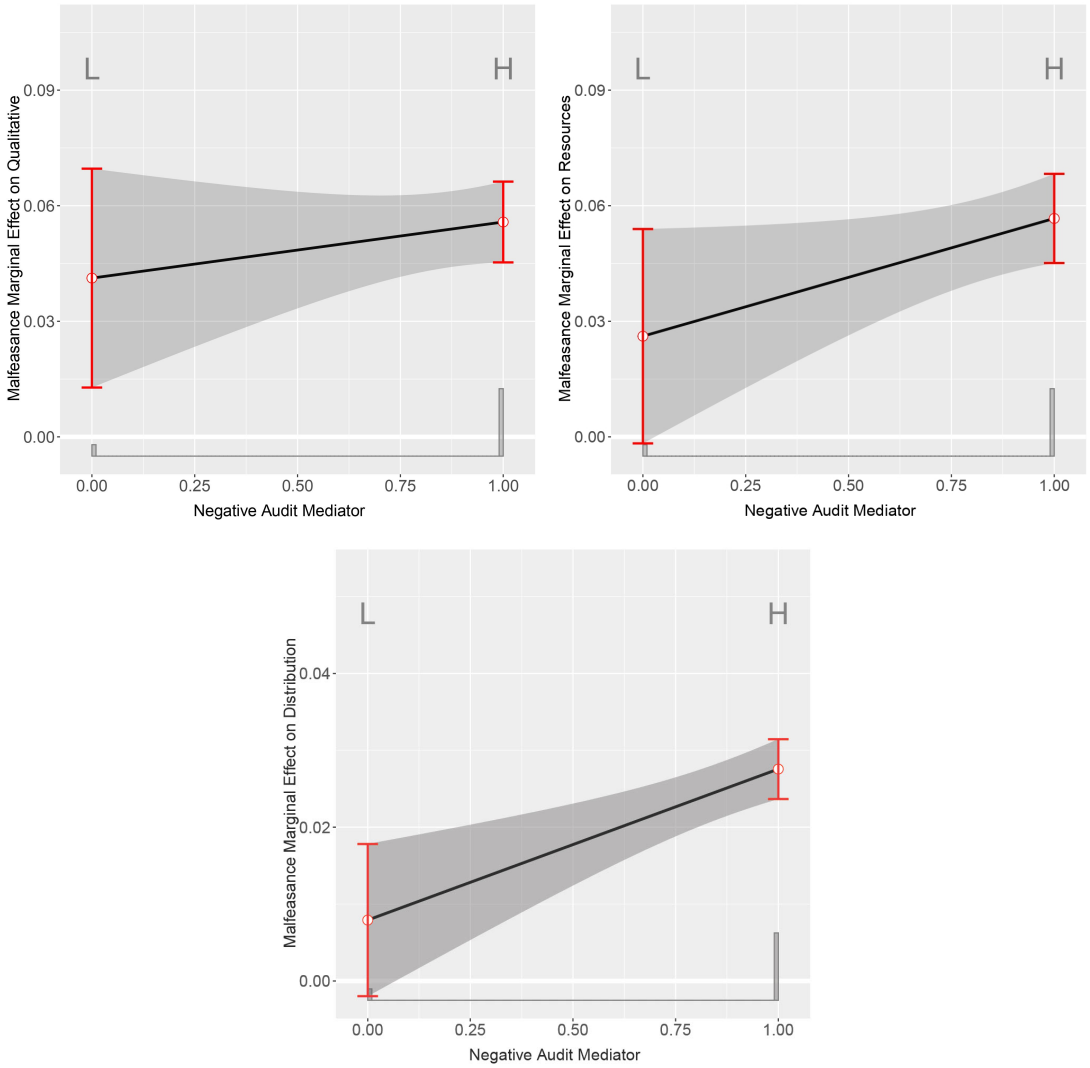
In the *Negative* models in Table 2.3, we interacted the *Negative* dummy variable (indicating whether or not treated subjects observed a deteriorating corruption record) with the *Treat* variable. The coefficient on the interaction term is significant in the case of the two incentivised malfeasance outcome measures – Distribution and Resources. Treated subjects, in municipalities with a deteriorating malfeasance record, update more negatively than those in municipalities with improving audit records. Although, again, subjects in both the negative and positive frame have, on average, posteriors that are more negative than their priors.[13]

The last frame of the information treatment in Figure 2.2 presented the total cost of reported malfeasance from which we created a malfeasance level variable (we log transform the foregone losses reported in this frame). Table 2.3 reports the estimates from our *Malfeasance* model. Subjects in the control condition get no information about malfeasance while those in treatment are informed about actual reported levels of malfeasance. As we pointed out earlier, the "dosage" of malfeasance reported for treated subjects is not randomly assigned. It certainly is possible that there are confounding unobservables associated with being treated with high versus low malfeasance reports. It is important clear up that all municipalities have positive levels of malfeasance. This means that even thought there some municipalities with positive audit reports (i.e. corruption went down) they still have positive level of malfeasance. Here we implement the instrumental variables (IV) estimation strategy proposed by Angrist and Imbens (1995), instrumenting for the amount of malfeasance observed by each subject using random assignment to treatment or control.[14] In the case of all three outcome measures, our instrumented *Malfeasance* variable is consistently significant. This is strong evidence for our claim that subjects are responding to the variation in malfeasance magnitudes in their local audit reports.

[13]The results for those who see the "positive" temporal comparison frame should be treated with caution. As it happens, only 22 of the municipalities had audits that improved over their previous CGO audit. This leaves us with slightly less than 500 subjects in the Positive temporal category. As a result, we are under-powered to estimate treatment effects for this group of subjects.

[14]The dosage cumulative density functions for treatment and placebo is presented in the Online Appendix Online Figure 2.9. It indicates there are no significant overall differences in the malfeasance "dosages" presented in the audit reports of subjects in treatment and control. This essentially confirms that there was successful random assignment within each commune.

Figure 2.4 – Linear marginal effects - All outcomes



Note: This Figure reports conditional marginal effects using the conventional linear interaction model and binning estimator. The moderator in this case is the dichotomous *Negative* variable. The explanatory variables "D" are the predicted values obtained from regressing the endogenous malfeasance variable on the exogenous treatment variable. Each plot reports the binning estimates and their standard errors and 95% confidence intervals. The plot also shows the percentage of observations within each bin.

Negative Interactions. We estimate a model that includes the interaction of the *Negative* dichotomous variable with the *Malfeasance* variable. Full results are reported in Table 2.24 in the Online Appendix. Following Hainmueller, Mummolo and Xu (2019) we summarise the magnitude and statistical significance of these interaction effects. Figure 2.4 reports the marginal effects of the malfeasance treatment for dichotomous values of the Negative moderator variable (black line) along with the binning estimator (red dots). The horizontal bars on the x-axis indicate the relative density of data on which the marginal effects are estimated. In the case of all three outcome variables, the marginal effect of reported Malfeasance is higher for subjects

who observed an audit report that was worse than their municipality's previous audit. As the data density plots in Figure 2.4 indicate, the positive Malfeasance effect is precisely estimated for the large number of subjects we observe in the *Negative* condition. As we noted earlier, we observe relatively fewer subjects in the *Positive* condition and as a result the marginal estimated effects of Malfeasance for these subjects are much less precise – encompassing zero in the case of the Resources and Distribution outcome variables. Subjects who were treated with an audit report that was worse than their locality's previous audit have a significant positive correlation between malfeasance dosage and belief updating. This is less likely to be the case for those who observed an improving audit report. This is additional evidence that 1) subjects were responding to the details of the audit report treatment – both temporal and malfeasance levels; and 2) there is some asymmetry to their updating.

Bayesian Learning Another conjecture is that belief updating reflected Bayesian learning. This implies that subjects update posteriors so as to reduce the distance between their prior beliefs about corruption levels and the actual corruption levels provided in the signal. In our case, treated subjects receive an indirect, or "noisy", signal that we can only assume is highly correlated with malfeasance beliefs. For each of our two incentivised quantitative measures (Resources and Distribution) we subtract the subject's belief from the actual information based on the 2020 CGO Audit Report. The treatment video provides subjects with a "noisy" signal about these two beliefs: quantitative information on the cost of malfeasance both in absolute terms and as a percent of the municipal budget; the absolute change in malfeasance expenditures since the previous audit; and foregone resources in terms of vaccine treatments. Bayesian learning is defined here is whether treated subjects (who receive quantitative audit information) update their assessment of foregone resources (Resources) and the corruption ranking of their municipality (Distribution) so as to reduce the extent to which their priors deviated from actual Resources and Distribution audit metrics for their municipality. This is a coarse indication of Bayesian learning because treated subjects only receive an indirect signal about the accuracy of their prior beliefs.

Our Bayesian conjecture is that post-treatment participants become more accurate about malfeasance levels in their municipality. Table 2.4 presents the pre- and post-treatment accuracy results for the Resources and Distribution measures for subjects in treatment and control. Bayesian learning suggests that, post-treatment, the gap between actual corruption metrics and beliefs should be smaller – closer to zero. As we would expect, the accuracy of beliefs, for both outcome measures remain quite stable pre- and post-treatment for the control subjects. Results for the treated subjects are contradictory. Results for the Distribution measure are consistent with our Bayesian learning conjecture: Subjects in treatment have more accurate beliefs – from 0.36 pre-treatment to essentially zero in post-treatment. On the other hand, results for the Resources outcome variable suggests that subjects have less accurate beliefs about corruption in post-treatment – their accuracy goes from -1.61 to -2.22.

Multivariate estimations in Panel B of Table 2.5 confirm this pattern. The learning model in Table 2.5 includes our measure of the difference between actual corruption levels in the municipality as reported in the 2020 audit report and the subject's responses to the two incentivised

Table 2.4 – Difference between Signal and Beliefs - Raw estimates

Panel A: Raw comparisons				
	Resources		Distribution	
Condition	Pre	Post	Pre	Post
Control	-1.63	-1.56	0.40	0.39
Treatment	-1.61	-2.23	0.37	0.02

Note: This table reports the differences of respondents beliefs minus the signal. We report both pre and post-differences. The closer to zero this statistics, the more accurate are respondents' belies.

Table 2.5 – Difference between Signal and Beliefs - Learning Model

	Resources	Distribution
Prior	0.422***	0.373***
	(0.021)	(0.021)
Treat	0.738***	0.315***
	(0.095)	(0.031)
Corrup diff	0.006	0.014
	(0.018)	(0.016)
Treat x Corrup diff	0.007	0.061*
	(0.029)	(0.025)
Covariates	Yes	Yes
FE	Yes	Yes
Num.Obs.	3439	3439
R2	0.268	0.282
R2 Adj.	0.237	0.251

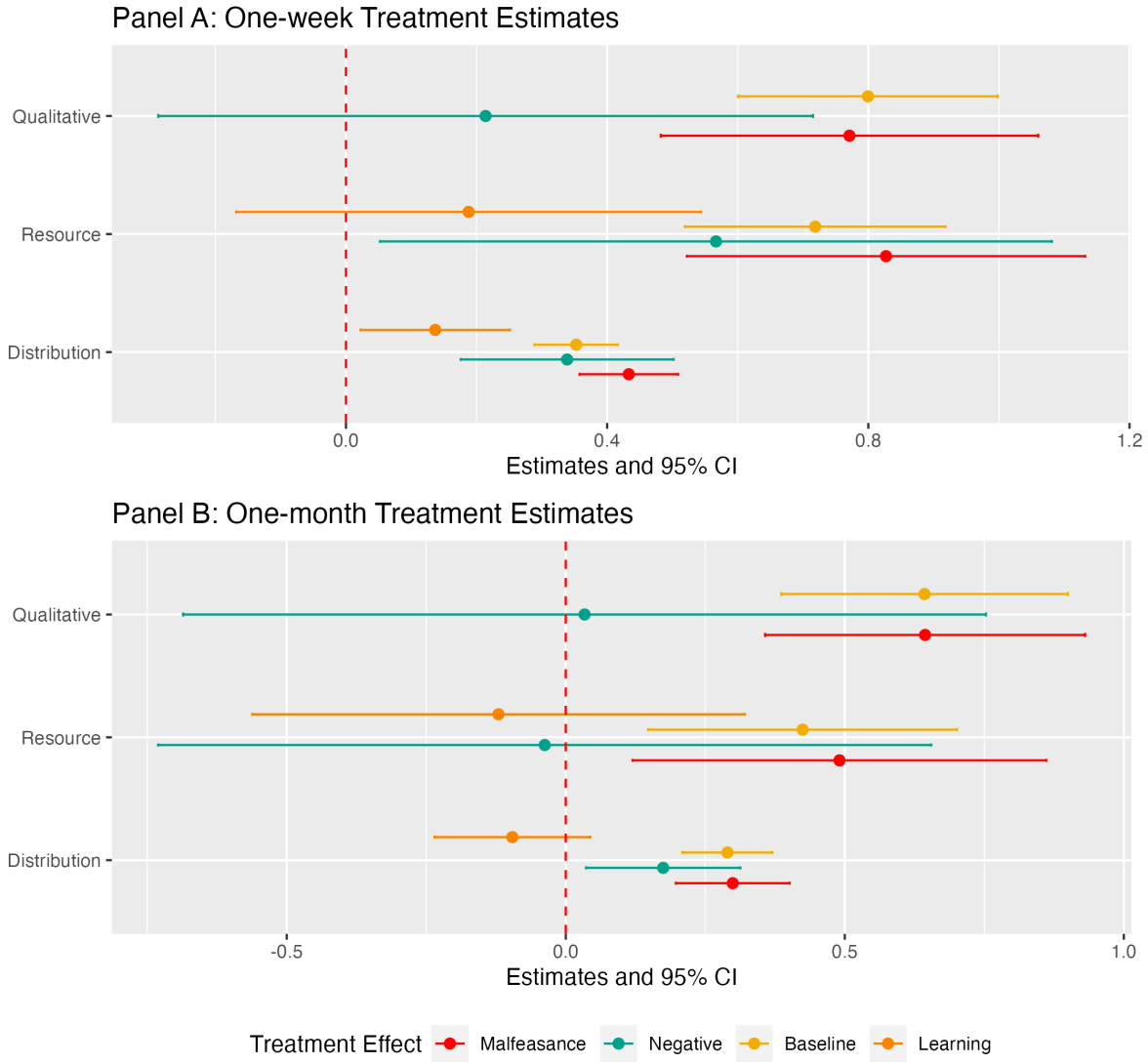
Panel A reports the average difference of respondents' priors and the corruption signal reported in the video treatment. We breakdown this metrics by treatment status: Treatment and Control and by survey wave: Pre-treatment and Post-treatment. The more accurate respondents' beliefs are, the closer this metric should be to zero. The farther away (either positive or negative) the less accurate are respondents' beliefs. Panel B, reports the regression coefficients of the *Learning* model reported in Table 2.1. The coefficient of interest in this specification is *Treat x Corrup Diff*, where *Corrup Diff* is categorical variable that captures the difference the respondents' priors beliefs about corruption and the signal reported in the video. For all specifications, we control for commune fixed effects. $+p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < 0.001$.

quantitative questions. Of interest here are the coefficients on this measure of the gap in beliefs interacted with the treatment variable: *Treat x Corrup Diff*. This measures the extent to which the posteriors of treated subjects is affected by this gap between their priors and actual malfeasance outcomes. For the Distribution model, the estimated coefficient for this interaction term, *Treat x Corrup Diff*, is statistically significant. Treated subjects do incorporate the malfeasance information in updating their quantitative priors about malfeasance. The interaction term is not significant in the Resources model (consistent with what we observe in Panel A). Results for the incentivised Distribution measure provide evidence of Bayesian learning.

A simpler version of Bayesian learning suggests that this updating should be conditioned on subjects' prior beliefs; specifically, on whether individuals held relatively negative or positive priors about government malfeasance. For each of the three malfeasance outcome variables we grouped subjects into those who score higher versus lower than the median value for their municipal sample (a high/low score indicates subjects believe malfeasance is high/low). Table 2.25 in the Online Appendix presents the high/low results for our three model specifications: *Baseline*, *Negative*, and *Malfeasance*. In each case we report results for the total sample; for those with high priors; and for respondents with low priors. For the *Baseline* model that only includes the *Treat* variable plus covariates, and for all three outcome variables, the coefficient on *Treat* is virtually identical for the Pooled, High Prior and Low Prior samples. A similar pattern is observed for the model that includes the *Negative* treatment variable and the *Malfeasance Levels* treatment measure. We observe some differences although in none of the cases are the estimated coefficients significantly different from each other. There is no support in our data for this simpler model in which levels of malfeasance priors condition the treatment effect.

Summary Treatment Effects Figure 2.5 summarizes the effects of the malfeasance message treatment effects estimated in the four models of belief updating: *Baseline* treatment; *Negative* frame; *Malfeasance* dosage; and *Learning*. Panel A in Figure 2.5 presents the estimated treatment effects measured one week after the intervention. Our *Baseline* model in Table 2.1 of belief updating predicts that assessments of prevailing malfeasance levels in local government would increase when subjects are simply informed that there are irregularities in government spending. The treatment in our post-treatment *Baseline* model increases posterior Qualitative beliefs, on average, by about 0.9 (which is about a third of a standard deviation for the change in belief measures that ranges between -10 and 10 and has a standard deviation of 2.45 for the control group). For the Resources model, the basic one-week post-treatment effect increases posterior malfeasance beliefs by about 0.71 (which is about a quarter of a standard deviation for the change in belief measures that ranges between -10 and 10 and has a standard deviation of 2.57 for the control group). The change in Distribution beliefs measure ranges between -3 and +3 and has a standard deviation of 0.4. The basic one-week post-treatment effect raises posterior malfeasance beliefs by about two-fifth of a standard deviation of 0.841. Simply informing citizens, regardless of their partisanship, that malfeasance exists in their local government will trigger significant, and negative, belief updating.

Figure 2.5 – Regression results of Malfeasance Beliefs - One-week and One-month Results



Note: Average treatment effects for all three outcome variables. *Baseline* is the treatment effect for those who received a video reporting audit outcomes. *Negative* indicates average treatment effect for subjects who observed a positive audit report. *Malfeasance* is the average treatment effect for subjects who received a video reporting above average levels of malfeasance. *Learning* is the effect of the *Treat x Corrup Diff* interaction term in the Bayesian updating model. The *Corrupt Diff* variable represents the difference between the respondent's prior and the signal that received in the video. Confidence intervals were constructed using a 95% confidence level with clustered standard errors. For all models, we control for commune fixed effects.

Panel A in Figure 2.5 also presents the one-week post-treatment effect for the interaction of the *Negative* dummy with the *Treat* variable. Consistent with the results reported in Table 2.3, they are positive for all three outcome variables and precisely estimated for the Distribution outcome variable. Beliefs regarding corruption updated positively if malfeasance performance improved compared to the last audit and update negatively otherwise. In Figure 2.5 *Malfeasance* is dichotomised into High and Low Malfeasance. All three outcome variables in Panel A have positive and significant Malfeasance dosage effects. Subjects treated with high levels of reported malfeasance have higher posterior malfeasance beliefs than those subjects viewing reports of low malfeasance. Finally, Panel A in Figure 2.5 presents the estimated effect of *Treat x Corrup Diff* from the *Learning* model. The *Corrup Diff* variable that represents the difference between the respondent's prior and the signal. As we can see in Panel A, there is some evidence of learning for the Distribution outcome variable.

Our information treatment consists of a single video delivered via WhatsApp shortly after the CGO released the audit results for the subjects' municipality. Any observed belief updating could be spurious – although, as we pointed out earlier, our design attempts to minimise the likelihood of spurious treatment effects. Very short-lived belief updating would suggest spurious treatment effects. To assess this, malfeasance beliefs are measured one week and then one month after the intervention. We estimated the identical models reported in Table 2.1 with the exception that the outcome variables are subjects' posterior beliefs one-month after the intervention.[15] There is strong evidence that the information treatments persist over this one month period. Panel B in Figure 2.5 summarises the one-month follow up treatment effects from our different model specifications and for each of the three outcome measures. For the *Baseline* model, the coefficient magnitudes on the treatment variable persist over the month period. The one-month and one-week *Negative* model effects are similar. As was the case in the one-week results, the *Negative* treatment is significant in the Distribution model but not in the other two models. The effects of the instrumented *Malfeasance* variable measured in the one-month follow-up are similar to the one-week estimates in Panel A. And in the case of the *Learning* model, the significant one-week coefficient for the Distribution model is indistinguishable from zero in the one-month follow-up. The overall persistence of our information treatment effects suggests that we are observing meaningful updating of beliefs regarding local government malfeasance.

Partisanship. We entertained two hypotheses regarding partisanship and malfeasance beliefs. One expectation was that partisans of the incumbent municipal mayor would believe that malfeasance levels are much lower than the levels expressed by non-partisans. The upper Panel A of Figure 2.6 presents the mean scores for the three outcome variables, pre-treatment, of aligned and not aligned subjects. In the case of all three outcome variables non partisans have much higher malfeasance beliefs than partisans: on the Qualitative measure, the difference is 1.76 (CI of 1.62,1.89; p-value of 0.000); for Distribution it is 0.49 (CI of 0.45, 0.53; p-value of 0.000); for Resources it is 0.96 (CI of 0.83, 1.14; p-value of 0.000).

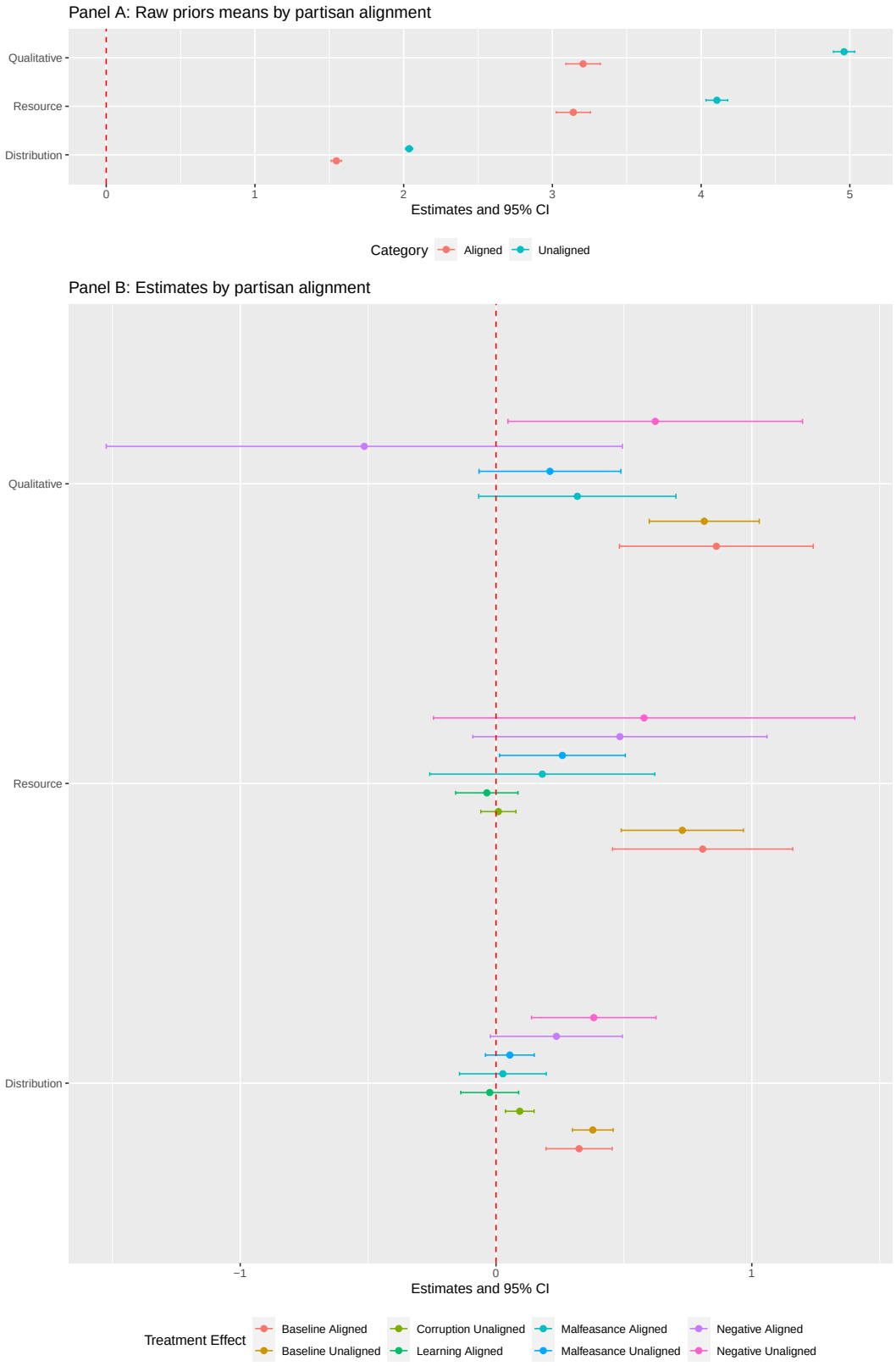
We also conjectured that subjects who are co-partisan would be less likely to update than non co-partisan of the incumbent municipal mayor. In the Online Appendix we separate the

[15]The full set of multivariate results for the follow-up outcome measures is provided in Table 2.23 of the Online Appendix.

sample into aligned and not aligned subjects and then estimate the models in Table 2.1 for each of these two groups. The results are reported in Table 2.26. Figure 2.6 presents the results for both aligned not aligned subjects in the lower Panel B. The *Baseline* treatment effect for all three outcome variables (Qualitative, Distribution and Resources) is large and significant for both partisans and non-partisans of the incumbent local mayor. There is, at best, weak evidence that aligned and not aligned differ in how they respond to quantitative details in the information treatments. The only notable difference with respect to the *Negative* model is for the Qualitative outcome variable: The aligned subjects register a negative although insignificant treatment effect while the not aligned subjects register a positive and statistically significant effect. There are small differences in how aligned and not aligned subjects respond to information about levels of *Malfeasance* provided in the audit reports. There is some evidence of a partisan effect in the *Learning* models. For the Distribution outcome variable in Figure 2.6 the estimated coefficient for the *Treat x Corrup Diff* coefficient from the *Learning* model has a significant positive coefficient for the not aligned subjects but the coefficient is negative and not significant for the aligned subjects.[16]

[16]In Table 2.22 we extended this analysis by looking at partisan alignment for respondents with strong and weak partisanship and find similar differences in treatment effects for the aligned and not aligned subjects.

Figure 2.6 – Partisan Priors and Regression Results of Malfeasance Beliefs by Partisan Alignment



Note: Panel A reports the average pre-treatment measures for Baseline, Resources and Distribution. Panel B: Average treatment effects for all three outcome variables. *Baseline* is the treatment effect for those who received a video reporting audit outcomes. *Negative* indicates average treatment effect for subjects who observed a negative audit report. *Malfeasance* is the average treatment effect for subjects who received a video reporting above average levels of malfeasance. *Learning* is the effect of the *Treat x Corrup Diff* interaction term in the Bayesian updating model. Confidence intervals were constructed using 95% confidence level with clustered standard errors. For all models, we control for commune fixed effects.

Partisans have significantly more positive priors regarding levels of malfeasance in their municipal government. But partisan preferences have a relatively limited impact on how subjects update their malfeasance beliefs. In most cases, the overall treatment effect is similar for both partisans and non-partisans of the incumbent municipal government. We observe small differences in updating related to the quantitative details of the information treatment.

Behaviour We tested whether informing average citizens about malfeasance in local government has behavioural consequences. The information treatments inform subjects about rent seeking associated with the provision of their local public goods. Our conjecture is that the belief updating caused by this messaging is meaningful and hence the information treatment effect should also be observed when subjects are asked to make a costly contribution to the municipality. Included in the survey is an incentivised behavioural measure of the subjects' confidence in local government delivery of public goods: In both pre- and post-treatment we give subjects the opportunity to make a 500 CLP (around 18% of their subject payments) contribution to their local municipality. We treat this is a costly behavioural signal of subjects' willingness to contribute to the cost of providing local public goods.

The outcome variable of interest is whether, or not, subjects donated 500 CLP[17] (around 15% of their subject payment) to their municipality. We estimate three models of the audit information treatment on donations: 1) a simple malfeasance treatment effect – whether subjects are in treatment or placebo; 2) temporal benchmarking - whether subjects respond to a negative frame (corruption goes up) or positive (corruption goes down); and 3) a malfeasance magnitude effect – whether treated subjects respond to the magnitude of the malfeasance. Table 2.6 reports the results of a logistic regression of donation on the information treatment variables along with covariate controls. We also control for their donation amount in the pre-treatment survey. The baseline model combines treated subjects who observe both a negative and positive temporal frame. In this case, the treatment coefficient is not significant – simply receiving information about malfeasance does not affect willingness to donate. The Positive and Negative models in Table 2.6 compare treatment effects on donations for subjects in the positive and negative information treatment frames. The coefficients for the binary treatment dummy variable indicate that those receiving the positive framed audit reports were more likely to donate and subjects in the negative framed treatment group were less likely to donate. In both cases the effect sizes are large and precisely estimated.[18] What stands out here is that this costly decision to donate is responsive to the temporal benchmarking described in the video treatments. Donations are significantly higher for treated subjects when they observe a positive audit for their municipality. Conversely, donations are significantly lower for subjects informed that their municipality audit is negative (corruption goes up).

The Malfeasance model in Table 2.6 estimates whether this donation responds to the amount of malfeasance reported in the video treatment. The coefficient for the Malfeasance term is statistically insignificant. The willingness to contribute is not correlated with the levels of

[17] 500 CLP is around 50 cents.

[18] While we provided the donation option in the follow-up survey we did not expect this donation effect to result in "serial" donations. And in fact, there was no information affect on donations in the follow-up data.

Table 2.6 – Likelihood of Making a Public Donation

	Baseline	Positive	Negative	Malfeasance	Qualitative	Resources	Distribution
Treat	-0.099	0.950*	-0.222+				
	(0.113)	(0.379)	(0.121)				
Previous	2.126***	3.254***	1.998***	2.126***	2.150***	2.133***	2.147***
	(0.114)	(0.391)	(0.121)	(0.114)	(0.115)	(0.114)	(0.115)
IV Malfeasance				-0.006			
				(0.008)			
Updating					-0.064**	-0.035	-0.145*
					(0.023)	(0.021)	(0.063)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	498	2941	3439	3439	3439	3439
AIC	2596.6	353.3	2259.8	2596.7	2589.3	2594.6	2592.1
BIC	3468.9	551.2	2978.1	3469.0	3461.6	3466.9	3464.4

Note: This table reports logistic regression results for a binary dependent variable: whether subjects donated 500 CLP (Around 0.50 USD dollars) to their municipal government. Controls include covariates, whether subjects donated in pre-treatment survey (*Previous*), how much subjects update in response to the treatment (*Updating*). The first three models regress donations on treatment controlling for donation in pre-treatment. The second and third model separate subjects into those viewing a Negative versus Positive audit performance. The fourth model reports results of regressing donation decision on the amount of malfeasance reported in the information treatment. The last three columns report results of models that regressed the donation decision on belief updating (for each of three outcome variables). For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

malfeasance identified in local government.

Our expectation is that this behavioural response is mediated by belief updating. This of course is difficult to ascertain here because the belief updating is not randomly assigned. With this important caveat in mind, we can nevertheless report results suggesting there is an indirect effect between the information treatment and donation behaviour that can be associated with belief updating. We measure the amount of belief updating as simply the difference between posteriors and priors on each of our three outcome variables (Qualitative, Resources and Distribution). Table 2.6 reports the results of regressing the donation decision on these three different measures of belief updating. In the case of two of the outcome measures – our qualitative indicator and our incentivised distribution measure – post-treatment changes in beliefs about corruption are correlated with donations. Subjects who have posterior beliefs about levels of malfeasance that are higher than their priors tend to donate less. Ideally we would like to parse out the direct causal contribution of information acquisition to donation behaviour from the indirect causal effect via the belief updating mediating variable.

In the case of a costly donation to a local government-provided public good, participants condition their donation on whether they observe a positive or negative audit report and there is some evidence that this is mediated by updated beliefs regarding malfeasance in their municipal government.

Education Our *Baseline* model of the intent to treat subjects with the malfeasance information video results in a very robust and significant estimated treatment effect. This model implies a very limited engagement with the quantitative content of the message – beliefs update when subjects are simply reminded that there is malfeasance in local government. There is also support for our richer model in which belief updating responds to the quantitative content of the malfeasance information messages. This model presumes that subjects engage with the quantitative content of the video treatment. Hence the likelihood that subjects will engage in the more demanding quantitative features of this information treatment will be conditioned on their cognitive ability. We use education as a proxy for cognitive ability. We have two expectations: We expect the treatment effect for the baseline model, with a dichotomous treatment variable, to be similar across education segments in the subject population. On the other hand, we conjecture that the higher educated are more responsive to the quantitative details of the malfeasance information treatment.

We exploit machine learning techniques in order to evaluate 1) whether education is a source of heterogeneous treatment effects; and 2) education-related heterogeneity varies by treatment. Our non-parametric modelling strategy uses Bayesian Additive Regression Trees (BART) (Duch, Laroze, Robinson and Beramendi, 2020; Green and Kern, 2012; Hill, 2011). This strategy estimates multiple regression trees, where each tree within the "forest" aims to explain only a small (and separate) proportion of the variance from the remaining trees. Fitted values of the outcome variable are estimated by summing the predictions across the trees. Since each tree chooses which variables to split on at nodes (and at what value), the BART model flexibly adapts to approximate the underlying functional form of the data, rather than relying on some assumed, prescribed parameterisation.

We recover ATE estimates by first using the BART models to generate predicted outcomes for the observed data, and then for a set of counterfactual observations. To train the BART model, we provide the observed outcome of interest (this is one of the Qualitative, Resources or Distribution measures) and a training data matrix consisting of the information treatment variable and the covariates of interest. We then simulate a second set of outcomes based on a separate test data matrix that contains "synthetic" observations that are identical to the training data except that the treatment assignment is inverted ("treated" becomes "control" and vice versa). The test data are only used post-estimation to predict counterfactual outcomes given the results of the BART model using the observed, training data. Estimating outcomes for *both* the observed and synthetic observations ensures that for each observation there is a corresponding counterfactual "control" case, where the only difference between the two is the value for treatment assignment. The individual-level effect of being treated is simply the difference between these two predictions.[19] Conveniently, with individual-level predictions, we can assess the extent to which education, taking into consideration other covariates, is a principal source of heterogeneity in the malfeasance video treatment effect.[20]

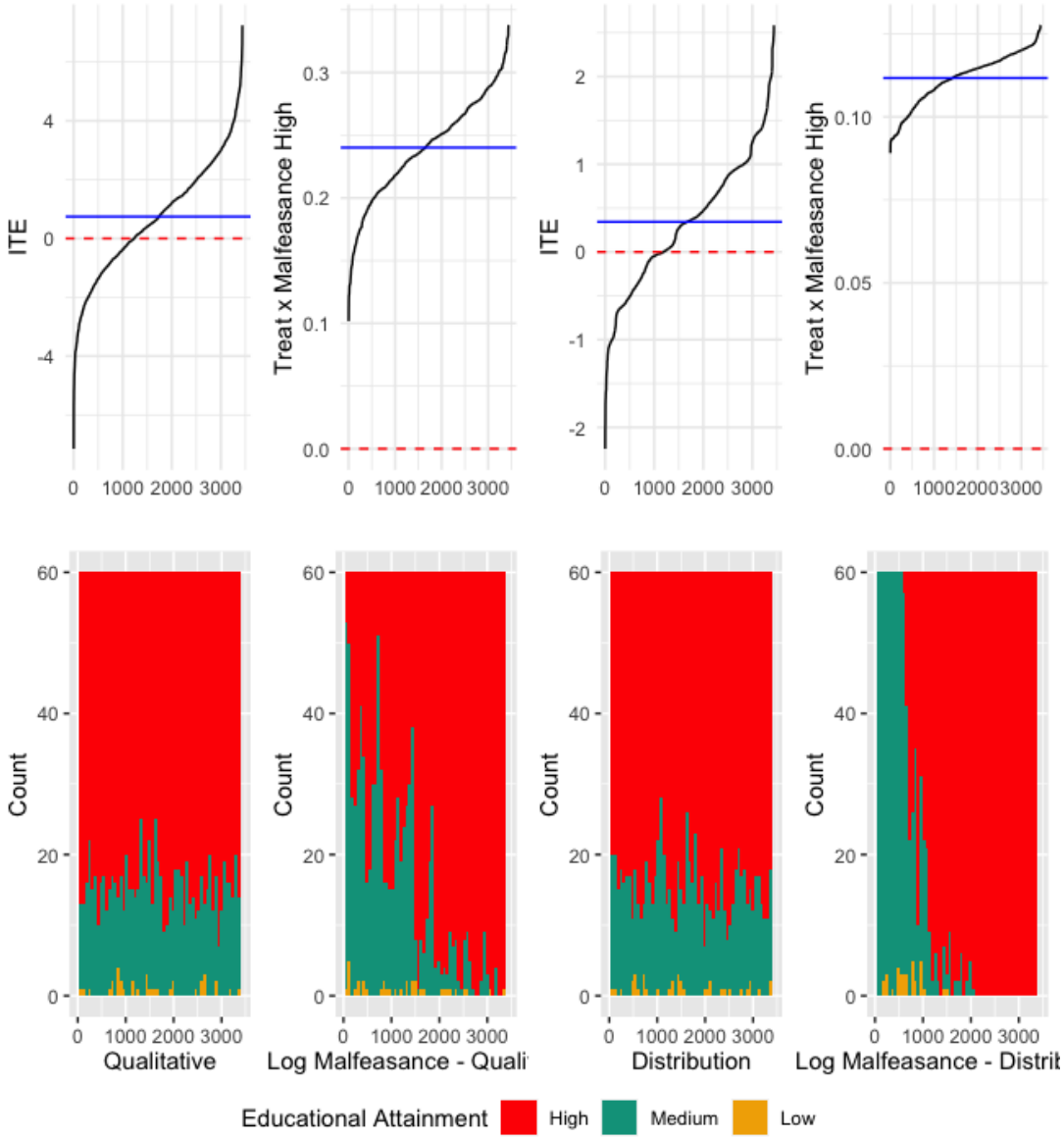
[19] Our BART model of heterogeneous effects is a simple specification generated using the BART R package with inputs described as above. All other options within BART are left at their default value. The R code and all of the data employed here is available on the replication site: https://doi.org/10.7910/DVN/SGTS1K

[20] Although there are clear data constraints here – we do not claim to have measured and estimated the effect of all possible covariates. Instead our claim is with respect to the various "pre-treatment" variables collected as

In the case of the Qualitative and Distribution outcomes, the results for BART analyses of the Baseline and Malfeasance treatment effect models confirm our conjectures. Figure 2.7 plots the distribution of the estimated individual-level treatment effects by magnitude along with the histogram of the corresponding education covariate values beneath. For these two outcome variables, the Baseline model has a significant malfeasance treatment effect. The plot in the top part of panel (a) of Figure 2.7 clearly demonstrates that the size of this effect is heterogeneous in the sample population: the overall estimated ATE for the Qualitative measure (that varies between 1 and 10) is approximately 0.75 (identical to the estimated coefficient in Table 2.3) and it ranges between -7.2 to 7.2. The plot in the lower part of panel (a) of Figure 2.7 indicates the density of the three education co-variate values over the range of estimated CATEs. The distribution of CATEs generated by BART suggests that education is not correlated with the magnitude of the treatment effects. The effect of being exposed to an information treatment reporting audited malfeasance is very similar across education groups in the sampled population. Panel (b) of Figure 2.7 reports similar BART analyses of the Baseline malfeasance treatment effect for the Distribution outcome variable. Again, we find that exposure to our audit information treatment generates belief updating effects that have similar density distributions for our three education groups. These results confirm our expectation that simple belief updating in response to any information about malfeasance would not be conditioned on educational levels.

part of the surveys.

Figure 2.7 – Heterogeneous Conditional Average Treatment Effects: Qualitative and Distribution Corruption Beliefs



Note: BART estimated heterogeneous effects by education; predicted CATEs are in the top panel of each column; and corresponding histograms of education distributions are in the bottom panel.

Our BART analysis does detect evidence, consistent with our conjecture, that the effect of malfeasance magnitudes would be conditional on education levels.[21] Results of the BART analyses of the malfeasance magnitude effect, for both the Qualitative and Distribution outcome variables, are reported in Panels (c) and (d) of Figure 2.7. In both cases, the medium-educated subjects have CATEs that are skewed in the direction of the lower part of the continuum while the higher educated are skewed in the higher part of the CATE continuum. The better educated clearly are much more responsive to the quantitative information regarding malfeasance

[21] To enable more straightforward estimation of the difference in predicted outcomes, we dichotomise the log of malfeasance variable into high (1) and low (0) respectively – we dichotomisee at the median value of 8.0.

magnitudes.

When we model the treatment in a very simple fashion with a dichotomous malfeasance variable (whether or not the subject observed a video discussing levels of malfeasance in their local government) the effect is highly significant and similar across education categories. This was our expectation given that the treatment effect does not imply any significant cognitive engagement with the quantitative content of the video. Belief updating is also correlated with malfeasance magnitudes that are described in the information treatments. This correlation we believe will be conditional on education because, we conjecture, the quantitative details of the information treatment require more cognitive engagement on the part of subjects. And we find that the malfeasance magnitude treatment effect is conditioned on education levels. We observe no evidence of education-related heterogeneous treatment effects in the case of the Resources outcome variable.

2.5 Discussion and Conclusions

We implement a field experiment in Chile that incorporates the 116 CGO municipal audit results from 2020 into video information treatments that are randomly assigned to 5,525 online subjects. Our goal was simply to understand how information about corrupt governance causes average citizens to update their beliefs. Understanding this simple relationship we believe is a critical precursor to modelling the relationship between malfeasance and responsibility attribution. This Chilean field experiment provides nine important insights into the updating of beliefs regarding malfeasance: citizens' priors regarding malfeasance in local government are quite uninformed; viewing a short malfeasance information video causes citizens to update beliefs; belief updating responds much more strongly to the general corruption topic rather than to the quantitative information in the messaging; updating is correlated with the magnitude of reported corruption; there is limited evidence of updating that resembles Bayesian learning; the better educated are more likely to respond to the quantitative details in the information treatments; partisans of local government incumbent parties have significantly lower beliefs about malfeasance prevalence; non-partisans and partisans of the local mayor respond for the most part similarly to the information treatments; and malfeasance information messaging has behavioural consequences.

We should be sceptical as to whether citizens are informed about corruption levels. The average Chilean citizen has quite uninformed priors regarding corruption in their local municipal government. This insight is based on pre-treatment measures of corruption belief benched against objective measures of the prevalence of malfeasance (from historical CGO audits). This is consistent with other efforts to calibrate beliefs regarding corruption and the performance of elected officials (Dunning, Grossman, Humphreys, Hyde, McIntosh, Nellis, Adida, Arias, Bicalho, Boas, Buntaine, Chauchard, Chowdhury, Gottlieb, Hidalgo, Holmlund, Jablonski, Kramon, Larreguy, Lierl, Marshall, McClendon, Melo, Nielson, Pickering, Platas, Querubín, Raffler and Sircar, 2019). And it is consistent with evidence regarding the accuracy of the public's priors regarding macro-economic outcomes (Cavallo, Cruces and Perez-Truglia, 2017; Lenz, 2013).

Citizens' beliefs about corruption respond to messaging about government malfeasance. The strongest belief updating treatment effect we record is simply whether or not the subject received a treatment versus a control video. Receiving a video message summarising the CGO's audit of irregular activities in their municipality causes Chileans to update their beliefs about corruption in their municipality. Regardless of whether their municipality's performance is more or less corrupt than its previous audit, treated subjects believe, post-treatment, that levels of corruption are higher in their municipality. This estimated treatment effect is very robust: We measure the outcome variable, beliefs about malfeasance, with three different variables: a qualitative variable (non-incentivised) and two incentivised quantitative variables. In all three cases the treatment effect is large and precisely estimated. Moreover, the treatment effect persists: Our initial post-treatment survey was conducted 3-6 days after the video intervention. In a follow-up survey, conducted one-month after the intervention, we observed the same large and significant treatment effects.

Subjects respond to the quantitative details of the malfeasance messaging. We provide them with two quantitative pieces of information in the videos that could improve the accuracy of their posteriors. Subjects are told of the magnitude of malfeasance, in monetary terms, in their municipality. We report a robust correlation between the "dosage" of malfeasance in the treatment and the magnitude of belief updating.

The information treatments also inform subjects of how much better or worse the audit results are compared to the municipality's previous audit report. For both incentivised quantitative outcome measures there is some evidence of higher updating of corruption beliefs if the municipality's performance deteriorated. But there was in fact a relatively small percentage of the subjects who observed malfeasance reports that improved over their previous audit. As a result the treatment effects we observed for these 'positively; treated subjects are imprecisely estimated and the comparisons between Negative and Positive treatment groups are directionally correct but not statistically significant.

We explore whether this belief updating incorporates in some fashion the relationship between priors and the informative signal provided in the treatment videos. There is some evidence in the case of the Distribution quantitative measure of classic Bayesian learning. The gap between priors and an informative signal about corruption correlates with updating but the estimates are imprecise. We examine whether the belief updating of subjects with negative versus positive priors regarding malfeasance respond differently to the information treatments. We find little conclusive evidence that this is the case.

Segments of the population exhibit variation in how they update their malfeasance beliefs. As has now been extensively document, partisanship colours the manner in which individuals consume information. Beliefs about malfeasance in government are affected by political partisanship. Co-partisans, not surprisingly, have less negative beliefs about levels of malfeasance. But both segments of the sampled population respond similarly to being treated by a malfeasance video. This is certainly the case for our robust *Baseline* treatment. We find limited evidence that partisans and non-partisans respond differently to quantitative details of information treatment.

The malfeasance information communicated in our video treatments was designed to be

accessible to the average Chilean citizen. Nevertheless, there are quantitative details that the average individual might overlook in 45 seconds or find difficult to understand. If we model updating as simply a response to having been treated with a video about malfeasance, we observe a similar large robust effect for all education categories. On the other hand, with respect to the cognitively challenging quantitative details, we find evidence that the educated are more responsive to the amount of reported malfeasance. There is evidence to this effect from two of our outcome variables; the Qualitative and Distribution measures.

The goal is to determine whether a 45 second video summarising audit results for a local municipality can cause *meaningful* changes in attitudes and behaviour. Our design and measurement strategy suggests the observed belief updating is meaningful: pre-treatment measurement of outcomes occurs four months before the intervention; the video treatment is relatively unobtrusive; and a one-month follow-up measurement of beliefs confirms the earlier observed treatment effects. Our measure of behavioural change is an opportunity for subjects to donate some of the earnings to their local municipal government. The donation is a costly direct contribution to a public good and we treat the donation as an endorsement of the municipality's provision of public goods. This behavioural measurement strategy is very much in the spirit of other studies that have used donations to candidates as a predictor of policy preferences (Bonica, 2018; Pérez, 2015) or donations in survey experiments to NGOs as an expression of preferences for lower government spending (Roth, Settele and Wohlfart, 2021). Costly contributions confirm our belief updating results: Subjects reduce their municipal donations when they are informed that their municipal audit is worse than its previous audit. Moreover, we have some evidence that this behavioural reaction to the information treatment is mediated by updated beliefs.

These results tell a fairly complete story of how the information in audit reports can have a meaningful impact on the public's beliefs about malfeasance in local government. As a point of departure, we note that the typical practice of simply publishing annual municipal audit reports has little noticeable impact on the public's beliefs about local corruption. In spite of having very regularly published audit reports, the Chilean public's beliefs about local governmental malfeasance are not well-informed and our social media metrics suggest that public engagement with this information is minimal. Simply establishing audit programmes, increasing their frequency, and publishing their results (Bobonis, Cámara Fuertes and Schwabe, 2016), in itself, will not likely have much impact on public beliefs about corruption.[22]

Our experimental results though suggest that there are messaging strategies that deliver significant belief updating in response to the malfeasance information reported in municipal government audits. Scaling this messaging strategy to a national population, such as Chile, is feasible. Delivering, via social media platforms like WhatsApp, short 45 second videos that provide a synopsis of audit outcomes will cause the general public to update their malfeasance beliefs. Regardless of their specific details, i.e., relatively positive or negative, this will generate a significant increase in malfeasance beliefs nation-wide because the public updates negatively to any messaging about corruption in local government.

[22]This is not to say that elected officials are not affected by the existence of audit programmes and the frequency with which they are published. There is strong evidence to suggest this is the case.

Can we successfully communicate to the average citizens the more nuanced, but important, quantitative details that are typically included in these audit reports? The reported treatment effects suggest yes, this is feasible: subjects engaged in temporal benchmarking; malfeasance beliefs increased in proportion to the amount of malfeasance reported in the video treatment; and we have some evidence of Bayesian learning. While statistically significant the effect sizes of these quantitative details are small relative to the overall treatment effect of simply viewing the video about malfeasance. The goal of future research should be identifying strategies for improving the communication of these more nuanced quantitative details to the general public. Our efforts to explore heterogeneity in treatment effects provide some guidance in this respect. Partisanship does not significantly undermine the effective communication of malfeasance information – partisans update similarly to non-partisans. Education does matter – there is scope to enhance belief updating by making this quantitative information accessible to individuals with lower levels of education. As a precursor to scaling up an information campaign, we should explore these and other communication strategies that will enhancing the information treatment effects.

Governments currently invest considerable resources in generating audit reports. Finan and Mazzocco (2021) make a convincing case that these audit efforts, as currently designed, are not necessarily the most cost effective policy initiatives for reducing corruption. Their modelling efforts, for example, suggest that audit probabilities of mayors should be weighted to heavily favour auditing the last term of an incumbent. We believe, and most modelling efforts seem to concur, that a public that is better informed about these audit results will enhance the cost effectiveness of audit programmes. Our trial results identify messaging strategies for audit report findings that will significantly increase the public's knowledge about malfeasance in local government. And we have some evidence that this information affects politically relevant behaviour. The marginal cost to existing audit programmes of scaling up messaging strategies similar to those in our trial is relatively low; the increased information levels are likely to enhance the audit impact on corrupt behaviour at the local level; and hence scaling up would be welfare enhancing.

2.6 Appendix

Subject Recruitment

In the initial phase of the experiment occurred from September to January 2019. We recruited 49,883 Chilean residents 18 years or over as potential participants. We recruited subjects using Facebook, Instagram, Twitter ads. We complemented our online recruitment with traditional forms of subject recruitment such as fliers, posters and newspaper ads. For online ads, study participants filled out a short form providing information about the municipality they reside, their gender and contact details. For subjects that were recruited via traditional forms of subject recruitment, they were invited to complete a detailed form with basic demographic information. An example of the Facebook ad employed in recruitment is presented in Figure 2.8. In January, 2019 the CGO selected the 116 municipalities that would be audited in 2020. Our recruitment efforts focused on these communes.

Figure 2.8 – Example Facebook Ad used to recruit study participants



Note: This figure shows a version of one of the Facebook ads that we used to recruit respondents. Respondents were invited to join the Nuffield Centre for Experimental Social Sciences subject pool. Study participants were asked their names, gender, and email account. Facebook ads were targeted to recruit study participants within the 116 municipalities that were audited in 2020. We recruited study participants at least one year before the audits were released, and subject recruitment took around 4 months (From January to March 2019).

We recruited our sample from a population of Chileans who have a social media presence, particularly on Facebook. The recruitment took place over two phases. In the first phase, we recruited 49,883 subjects into a pool of potential subjects. A total of 22,755 of these recruits resided in the 116 municipalities audited by the CGO and hence were eligible for the experiment. We sent invitations, to participate in the pre-treatment survey, to these 22,755 eligible subjects. We had a response rate of 25 percent that left us with 5,525 subjects participating in the pre-treatment survey. Table 2.7 compares the demographic features of the subject pool (including eligible respondents) and the final sample with statistics from the Chilean national census. Note that the pool and the sample are very similar in terms of age, education, gender, and income. There are differences between our sample, the subject pool and Chile's demographic population. Our sample is disproportionately younger and it has higher education attainment. In terms of gender, more women took part of the study, representing 65% of our sample. Our sample reports a relatively smaller income compared to the subject pool.

Table 2.7 – Profile of Subject Pools and Final Sample

	Census	CASEN	Subject pool	Eligible	Sample
N*	17,574,003	2,164,439	49,883	31,971	5,525
Age	40	37	40	39	36
Education					
Secondary	40%	37%	15%	17%	67%
Tertiary	24%	22%	83%	82%	31%
Female	51%	52%	51%	50%	65%
Income	-	607	1291	1324	996

Note: This table basic demographic features of the subject pool and sample used in the study and compares it to the key demographic gathered from the 2017 Chilean census. We can observe that our sample is disproportionately female and has a higher income than the average Chilean citizen. It is also slightly younger than the average population and it has attained lower levels of education.

Compliance and Complier Average Causal Effects (CACE)

Table 2.8 summarises treatment compliance. A Placebo video was sent via WhatsApp to 1,831 subjects and a treatment video to 3,694 subjects. We have no information on whether subjects actually clicked on the video link they received on their personal device. Immediately after the video is played on subjects' personal device they are given a short ChatBot survey. We can characterise compliance here as subjects who responded to the Chatbot survey that appeared right after the video treatment. This would assume that those who actually watched the video also responded to the short post-video Chatbot survey. Response rates to the ChatBot survey were the following: For Placebo 59% responded to the Chatbot survey and for Treatment it was 56%.

The Chatbot survey asked two attention check questions:

- First question: "What was the topic of the video?" 98% answered correctly in Placebo and 95% answered correctly in Treatment.
- Second question: "What was the information reported in the video?'" 97% answered correctly in Placebo and 99% answered correctly in Treatment.

We see that for both attention check questions, both treated and placebo subjects answering correctly. As we indicated in the main text, we define ChatBot compliance as respondents that answered the second attention check questions correctly: 59% for placebo subjects and 56% for subjects in the audit treatment.

A post-treatment survey was administered to participants 3-6 days after they were treated with the videos. A total of 3,592 subjects responded to the survey: 2,249 out of 3,694 treated (61%) and 1,343 out of the 1,831 placebo subjects (73%). In the post-treatment sample, we asked subjects the following question about their recall of receiving the treatment videos: "Do you remember receiving a video from the NGO Chile Transparente?" 846 (63 percent of those responding to the post-treatment survey) indicated they had received the Placebo video and 1,670 (74 percent of those responding to the post-treatment survey) indicated receiving the treatment videos. Subjects who reported viewing the video were also asked the following attention check questions:

- Question:"Do you remember what was the content of this video?" 80 percent answered correctly in Placebo and 84 percent answered correctly in Treatment.

We will define Post-Treatment Compliers as the approximately 38 percent of subjects initially assigned to either placebo or treatment condition who both answered the post-treatment survey and reported viewing the treatment videos. In the main text we reported intention-to-treat (ITT) results in the main text. We will use these two definitions of compliance (Chatbot Compliers and Post-Treatment Compliers) to estimate the Complier Average Causal Effects (CACE).

Table 2.8 – Treatment Compliance

Treatment Status	Placebo	%	Treatment	%	Total	% T. sample
Videos sent	1831		3694	-	5525	100
A:Chatbot survey						
Clicked video	No information					
Respondent Chatbot survey	1077	100	2054	100	3131	57
First compliance question - correct	1055	98	1961	95	3016	
First compliance question - incorrect	22	2	112	5	134	
Second compliance question - correct	1042	97	2039	99	3130	
Second compliance question - incorrect	34	3	15	1	49	
B:Post-treatment survey						
Total eligible	1831	100	3694	100	5525	
Responded	1343	73	2249	61	3592	65
No response	489	27	1447	39	1936	
C:Post-treatment - compliance						
Total eligible	1343	100	2249	100	3592	
Reported receiving the video	846	63	1670	74	2516	46
Attention check question - correct	680	51	1401	84	2081	
Attention check question - incorrect	166	12	269	16	435	
D:Chatbot and Post-Treatment survey						
Total eligible	1343	100	2249	100	3592	
Responded both surveys	987	73	2055	91	3042	55
Attention check Chatbot correct	967	72	1962	87	2929	
Attention check Chatbot incorrect	20	1	93	4	111	
F:Follow up - compliance						
Total eligible	1057	100	1198	100	2255	
Reported receiving the video	534	100	729	100	1263	23
Attention check question - correct	371	69	548	75	919	
Attention check question - incorrect	163	31	181	25	344	
G:Chatbot and Follow up survey						
Total eligible	1057	100	1198	100	2255	
Reported receiving the video	534	44	729	60	1263	22
Attention check Chatbot correct	411	39	559	45	970	
Attention check Chatbot incorrect	3	0.3	25	0.02	28	

Note: This table reports levels of compliance and attention checks for all three waves. Panel A provides a breakdown of correct answers in the attention checks included in the Chatbot survey. These questions came up immediately after respondents watched the video. Panel B reports how many of the initial baseline subjects answered the post-treatment survey (that was administered one-week after the treatment). Panel C reports the number of respondents that answered the post-treatment survey, reported watching the video and whether they answered correctly or incorrectly the attention check question. We used the levels of compliance in panel C to estimate the average compliance causal effects reported in Table 2.18, which we called *Post*. Panel D, reports the numbers of respondents that answered both the Chatbot and post-treatment survey. It also reports the numbers of respondents that answered correctly both question check questions. We used the levels of compliance in panel D to estimate the average compliance causal effects reported in Table 2.18, which we called *Chatbot*. Panel F reports the levels of compliance in the follow-up survey. Panel G reports the levels of compliance for respondents that answered both the Chatbot and post-treatment survey.

Table 2.8 suggests different definitions of compliance: 1) "Chatbot Compliers" are subjects

who responded to the Chatbot survey that appeared right after the video treatment and also responded the Post-treatment survey; 2) Post-Treatment Compliers are subjects who both answered the post-treatment survey, reported viewing the treatment videos and answered correctly to the attention check questions. Tables 2.18, 2.19, and 2.20 report the estimates for Chatbot Compliers and the Post-Treatment Compliers for the *Baseline*, *Negative*, and *Malfeasance* models reported in Table 2.1. We find results that are overall are in line with the baseline found in our baseline model models. The CACE for both measures of compliance tends to be larger than the unconditional treatment effect.

Treatment Diagnosis and Attrition

Diagnostic comparing the demographic profiles of subjects assigned to the different treatment arms will signal the likelihood that missingness is not independent of treatment status. To the extent that we observe any significant attrition in the sample and evidence that it might be correlated with treatment status, we can implement estimation strategies designed to address any bias that might be associated with this attrition.

The randomisation checks suggest that attrition over the three interview waves had relatively minor affects on the distribution of covariate values across our treatment arms. The assumption of course is that any missing observations resulting from attrition are missing at random. We propose two further strategies for addressing possible bias in estimated treatment effects that might have resulted from "missingness" that is correlated with treatment status.

Table 2.9 – Randomisation checks - Pre-treatment variables

	Baseline	Post-Treatment	Follow-up
Intercept	0.661***	0.692***	0.628***
	(0.022)	(0.035)	(0.054)
Gender	0.007	0.005	-0.018
	(0.013)	(0.017)	(0.022)
Income	-0.005	-0.009	-0.009
	(0.014)	(0.018)	(0.031)
Education	0.003	-0.007	-0.010
	(0.015)	(0.022)	(0.029)
Partisanship	-0.006*	-0.010**	-0.015**
	(0.003)	(0.004)	(0.006)
Turnout	-0.014	-0.016	-0.001
	(0.013)	(0.019)	(0.023)
Support incumbent	-0.011	-0.011	-0.010
	(0.014)	(0.021)	(0.036)
Civic knowledge High	0.027	-0.066	-0.184*
	(0.058)	(0.075)	(0.090)
Civic knowledge Medium	0.038	-0.040	-0.160*
	(0.058)	(0.073)	(0.088)
Political information High	0.016	0.032	0.046
	(0.016)	(0.023)	(0.032)
Political information Medium	0.003	0.009	0.019
	(0.013)	(0.015)	(0.021)
Anova F-test Nested (P-value)	0.353	0.07	0.045*
Num.Obs.	6050	3592	2255
R2	0.002	0.004	0.008
R2 Adj.	0.000	0.001	0.003

Note: This table reports the regression estimates from regressing a dummy treatment assignment variable on pre-treatment variables. We observed balance across most pre-treatment variables with the exception of partisanship and civic knowledge in the last wave. Treatment assignment was conducted on the original 6,050 respondents; however, 525 subjects were excluded from the analysis as the CGO did not complete the audit on time. Standard errors are clustered at the municipal level. We also report the p-value for a F-test that compares a nested model that only includes the intercept to the full specification reported in this table for all waves. *$p < .05$, **$p < .01$, ***$p < 0.001$.

Wave	D	P-value
Baseline	0.03	0.22
Post-treatment	0.04	0.06
Follow-up	0.05	0.07

Note: This table report the results of a Two-sample Kolmogorov-Smirnov test for the partisanship variable. Although we find unbalance on partisanship reported in Table 2.9, we did not find statistically significant differences in the cumulative distribution of this variable.

Table 2.10 – Two-sample Kolmogorov-Smirnov test - Partisanship

Inverse Probability Weighting With the complete set of demographic measures collected in the pre-treatment survey we can model attrition quite robustly. Inverse probability weighting is one of the estimation strategies we can implement. It has the value of being very straight-forward – essentially modelling the attrition process as a function of observable covariates (Anderson et al., 2021). The weights are based on the predicted values from a logistic regression of a binary variable indicating whether the observation is missing on the available covariates, where all the available covariates are allowed to interact with the treatment indicators.[23] The weight is simply 1 over 1 minus these predicted probabilities.[24] We would then re-estimate treatment effects on the we subset of the data where outcomes are observed and weight that estimate using these weights. The estimate from this regression is a consistent estimate for the treatment effect assuming the censoring process is observable. Table 2.11 compares the unweighted baseline treatment model coefficients with those that incorporate IPW for the three outcome variables. The prior coefficients for the weighted estimates are significantly higher suggesting that subjects we do not observe in subsequent waves give much more weight to their prior beliefs than is the case for those subjects that we do observe. The estimated treatment effects, on the other hand, are similar in both weighted and unweighted estimations.

Table 2.11 – Regression results of Malfeasance Beliefs - Unweighted and IPW

	Qualitative		Resources		Distribution	
	Unweighted	Unstable	Unweighted	Unstable	Unweighted	Unstable
Prior	0.415***	0.424***	0.426***	0.428***	0.325***	0.291***
	(0.015)	(0.016)	(0.016)	(0.017)	(0.018)	(0.017)
Treat	0.758***	0.734***	0.767***	1.057***	0.351***	0.390***
	(0.071)	(0.073)	(0.079)	(0.084)	(0.026)	(0.027)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3592	3592	3592	3592	3592	3592
R2	0.313	1.000	0.262	1.000	0.272	1.000
R2 Adj.	0.290	1.000	0.237	1.000	0.248	1.000

Note: This table reports the unweighted and weighted estimates for the *Baseline* specification in Table 2.1 for all three outcomes. *Unweighted* estimates correspond to the baseline model without using any weighting scheme on the estimation of the regression coefficients. *Unstable* correspond to regression estimates that incorporate inverse probability weighting. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

[23] An alternative here, that we would consider using, is a propensity score estimation algorithm to fully model any possible nonlinearities – the *twang* package, for example (Ridgeway et al., 2021).

[24] These weights are often characterised as "unstable" – a slightly modified estimation strategy can generate more "stable" weights.

Bounded Treatment Effects An alternative estimation employed by (Lee, 2009) is a bounding estimator that allows us to assess whether results are sensitive to the missingness resulting from attrition. He assumes that attrition is monotonic, which in our setting, implies that any subject who would not be missing if assigned to the control group would also not be missing if assigned to one of the treatment arms.

First, we calculate the proportion of missing and observed data in the treated and control groups. Second we calculate the average malfeasance belief score among the observed treated subjects and the malfeasance score among the observed control subjects. For the bounds, we calculate a weighted average and replace the unobserved outcomes with either the minimum or maximum malfeasance scores (this is 0 and 10 for the qualitative outcome; 0 and 10 for the Resources variable; and 1 and 4 for the Distribution variable). As is generally the case, these bounds are quite wide. In order to trim the bounds we assume, under the bounds, that a subject wouldn't have a missing malfeasance score if assigned to control would also not be missing if assigned to the treatment. These bounds will be narrower than simply imputing minimum and maximum values Anderson et al. (2021); Lee (2009).

We re-estimate the treatment effect regressions imputing missing values with upper and lower trimmed malfeasance scores. Table 2.12 reports the results for the qualitative malfeasance outcome measure. In this case, we did find important differences for the treatment and log malfeasance variables.

Table 2.12 – Regression results of Malfeasance Beliefs - Bounds

	Pooled			Upper bound			Lower bound		
	Baseline	Negative	Malfeasance	Baseline	Negative	Malfeasance	Baseline	Negative	Malfeasance
Prior	0.406***	0.406***	0.406***	0.365***	0.365***	0.365***	0.334***	0.334***	0.335***
	(0.019)	(0.019)	(0.019)	(0.018)	(0.018)	(0.018)	(0.020)	(0.020)	(0.020)
Treat	0.733***	0.530*	0.337	1.969***	1.822***	2.070**	-0.730***	-1.014***	-2.168***
	(0.080)	(0.233)	(0.469)	(0.089)	(0.249)	(0.634)	(0.100)	(0.229)	(0.580)
Negative		-3.009***			1.979***			4.283***	
		(0.314)			(0.170)			(0.159)	
Log malfeasance			0.203***			0.096**			0.350***
			(0.028)			(0.032)			(0.030)
Treat:Negative		0.238			0.173			0.333	
		(0.249)			(0.266)			(0.256)	
Treat:Log malfeasance			0.028			-0.007			0.101*
			(0.034)			(0.044)			(0.040)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	3439	3439	3915	3915	3915	3915	3915	3915
R2	0.319	0.319	0.319	0.310	0.310	0.310	0.200	0.201	0.202
R2 Adj.	0.290	0.290	0.290	0.284	0.284	0.284	0.170	0.170	0.171

Note: This table reports both ITT and Lee bounds estimates for all models reported in Table 2.1. We reported pooled estimates to use as a benchmark against the Lee bounds. The *Upper bound* assumes that all respondents that did not responded the post-treatment survey in the treatment group had the highest outcome, and all attritors in the comparison group had the lowest outcome. Conversely, the *Lower Bound* corresponds when all attritors in the treatment group had the lowest outcome and all attritors in the control group had the highest outcome. We clustered standard errors at the commune level. For all models, we control for commune fixed effects.$+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.14 – Regression results of Malfeasance Beliefs - Resource Outcome

	Post-treatment				Follow-up			
	Baseline	Positive	Negative	Malfeasance	Baseline	Positive	Negative	Malfeasance
	(0.679)	(2.110)	(0.703)		(0.753)	(2.203)	(0.810)	
Prior	0.412***	0.398***	0.414***	0.412***	0.347***	0.426***	0.341***	0.347***
	(0.018)	(0.049)	(0.020)	(0.019)	(0.023)	(0.077)	(0.025)	(0.023)
Treat	0.728***	0.255	0.821***	-0.351	0.469***	0.252	0.490***	-0.004
	(0.091)	(0.179)	(0.098)	(0.537)	(0.120)	(0.358)	(0.130)	(0.776)
Log malfeasance				0.239***				0.282***
				(0.048)				(0.054)
Treat:Log malfeasance				0.075+				0.033
				(0.039)				(0.054)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	498	2941	3439	2143	269	1874	2143
R2	0.268	0.362	0.258	0.269	0.261	0.404	0.254	0.261
R2 Adj.	0.237	0.297	0.227	0.238	0.211	0.290	0.205	0.211

Note: This table reports the regression coefficients of both the *Baseline* and *Malfeasance levels* specifications reported in Table 2.1 for the Resources outcome. We report estimates obtained from both the post-treatment and follow-up. Post-treatment estimates correspond to models within the 3-6 days and follow-up estimates are all estimates obtained a month after. We included controls in all models. Standard errors are clustered at the commune level. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.13 – Regression results of Malfeasance Beliefs - Qualitative Outcome

	Post-treatment				Follow-up			
	Baseline	Positive	Negative	Malfeasance	Baseline	Positive	Negative	Malfeasance
Prior	0.406***	0.301***	0.425***	0.406***	0.378***	0.336***	0.386***	0.378***
	(0.019)	(0.036)	(0.021)	(0.019)	(0.026)	(0.062)	(0.028)	(0.026)
Treat	0.733***	0.533*	0.772***	0.337	0.650***	0.570	0.654***	0.076
	(0.080)	(0.232)	(0.086)	(0.469)	(0.127)	(0.497)	(0.134)	(1.016)
Log malfeasance				0.203***				0.271***
				(0.028)				(0.041)
Treat x Log malfeasance				0.028				0.039
				(0.034)				(0.067)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	498	2941	3439	2143	269	1874	2143
R2	0.319	0.349	0.316	0.319	0.322	0.426	0.315	0.322
R2 Adj.	0.290	0.283	0.287	0.290	0.276	0.316	0.270	0.276

Note: This table reports the regression coefficients of both the *Baseline* and *Malfeasance levels* specifications reported in Table 2.1 for the Qualitative outcome. We report estimates obtained from both the post-treatment and follow-up. Post-treatment estimates correspond to models within the 3-6 days and follow-up estimates are all estimates obtained a month after. We included controls in all models. Standard errors are clustered at the municipal level. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.15 – Regression results of Malfeasance Beliefs - Distribution Outcome

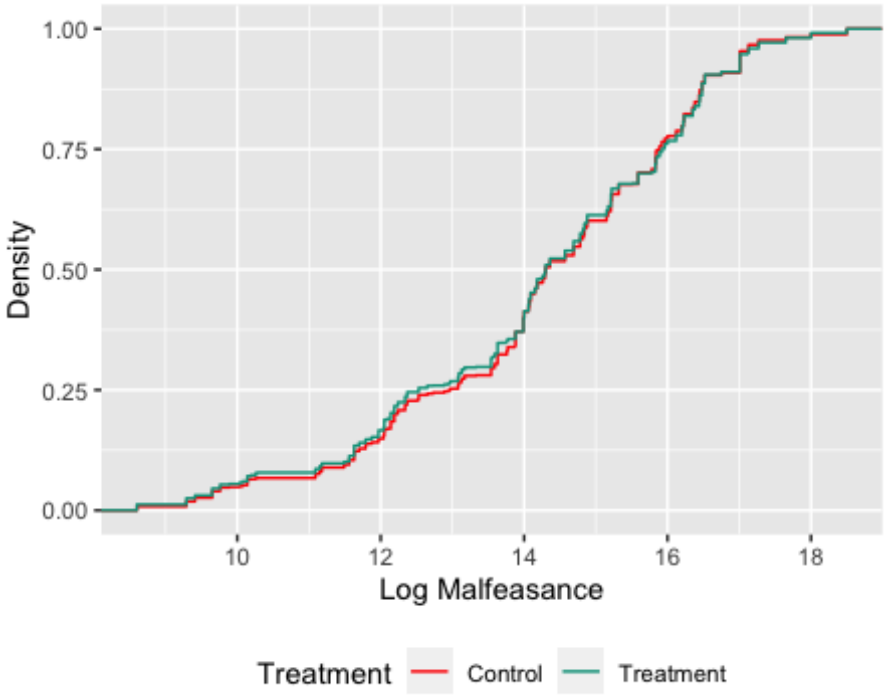
	Post-treatment				Follow-up			
	Baseline	Positive	Negative	Malfeasance	Baseline	Positive	Negative	Malfeasance
Prior	0.320***	0.298***	0.324***	0.321***	0.265***	0.240***	0.268***	0.265***
	(0.021)	(0.056)	(0.023)	(0.021)	(0.026)	(0.046)	(0.030)	(0.026)
Treat	0.342***	0.091	0.386***	-0.455**	0.255***	0.101	0.274***	-0.076
	(0.029)	(0.067)	(0.028)	(0.148)	(0.031)	(0.093)	(0.032)	(0.225)
Log malfeasance				0.093***				0.143***
				(0.010)				(0.014)
Treat:Log malfeasance				0.055***				0.023
				(0.010)				(0.016)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	498	2941	3439	2143	269	1874	2143
R2	0.280	0.349	0.266	0.284	0.271	0.426	0.255	0.271
R2 Adj.	0.250	0.283	0.235	0.254	0.222	0.316	0.207	0.222

Note: This table reports the regression coefficients of both the *Baseline* and *Malfeasance levels* specifications reported in Table 2.1 for the Distribution outcome. We report estimates obtained from both the post-treatment and follow-up. Post-treatment estimates correspond to models within the 3-6 days and follow-up estimates are all estimates obtained a month after. We included controls in all models. Standard errors are clustered at the commune level. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Dosage Instrumental Variable Estimation As we point out in the main text, subjects are randomly assigned to treatment and placebo conditions. But the "dosage" of malfeasance reported for treated subjects is not randomly assigned. It certainly is possible that there are confounding unobservables associated with the communes receiving, for example, high versus low malfeasance reports[25]. Here we implement the instrumental variables (IV) estimation strategy proposed by Angrist and Imbens (1995) instrumenting for the amount of malfeasance reported in each commune using the commune's random assignment to treatment or control. As the dosage cumulative density functions for treatment and placebo in Figure 2.9 indicate there are no significant overall differences in the reported malfeasance dosages for subjects in treatment and control. This essential confirms that there was successful random assignment within each commune.

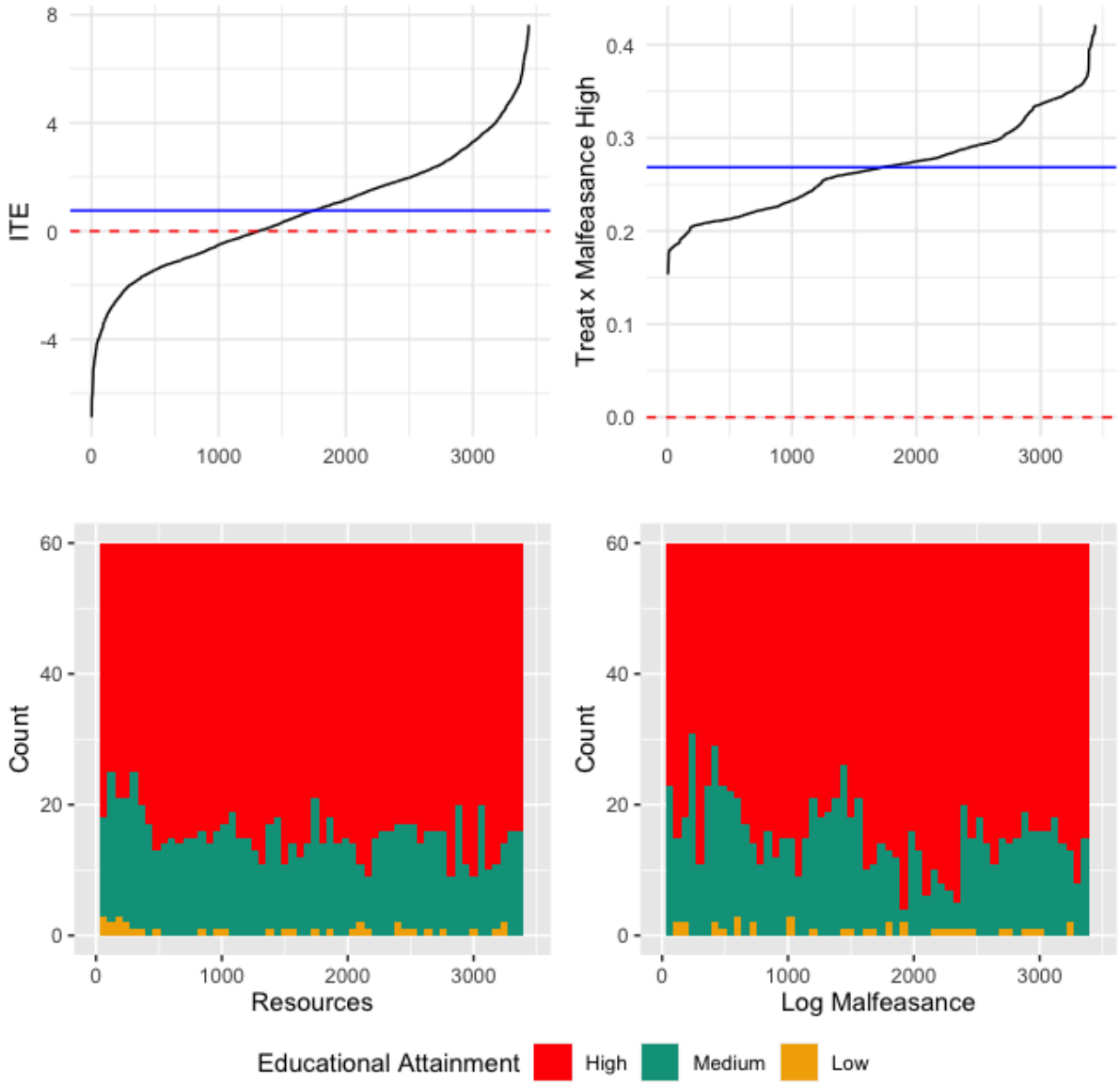
[25]Just to remind the reader, all municipalities had positive levels of malfeasance on their last audit, even municipalities that received a positive audit. The levels of malfeasance vary from one local government to another. Thus, there are some municipalities with high levels of malfeasance and others with low levels of malfeasance. Therefore, there are cases when municipalities got a positive audit, i.e. corruption went down concerning the previous audit, but the level of malfeasance was still high.

Figure 2.9 – CDF of Malfeasance Dosage by Treatment Status



Note: This figure reports the cumulative distribution of the levels of malfeasance reported in the video treatments for all communes. We used the log of malfeasance, which is the log transformation of the amount of malfeasance found in the municipality in Chilean pesos. The log of malfeasance ranges from 8.6 to 18.5, with a mean of 14.1.

Figure 2.10 – Heterogeneous Conditional Average Treatment Effects: Corruption Resources Beliefs



Note: You can find data and scripts to replicate the results of this figure on the following link: https://doi.org/10.7910/DVN/SGTS1K

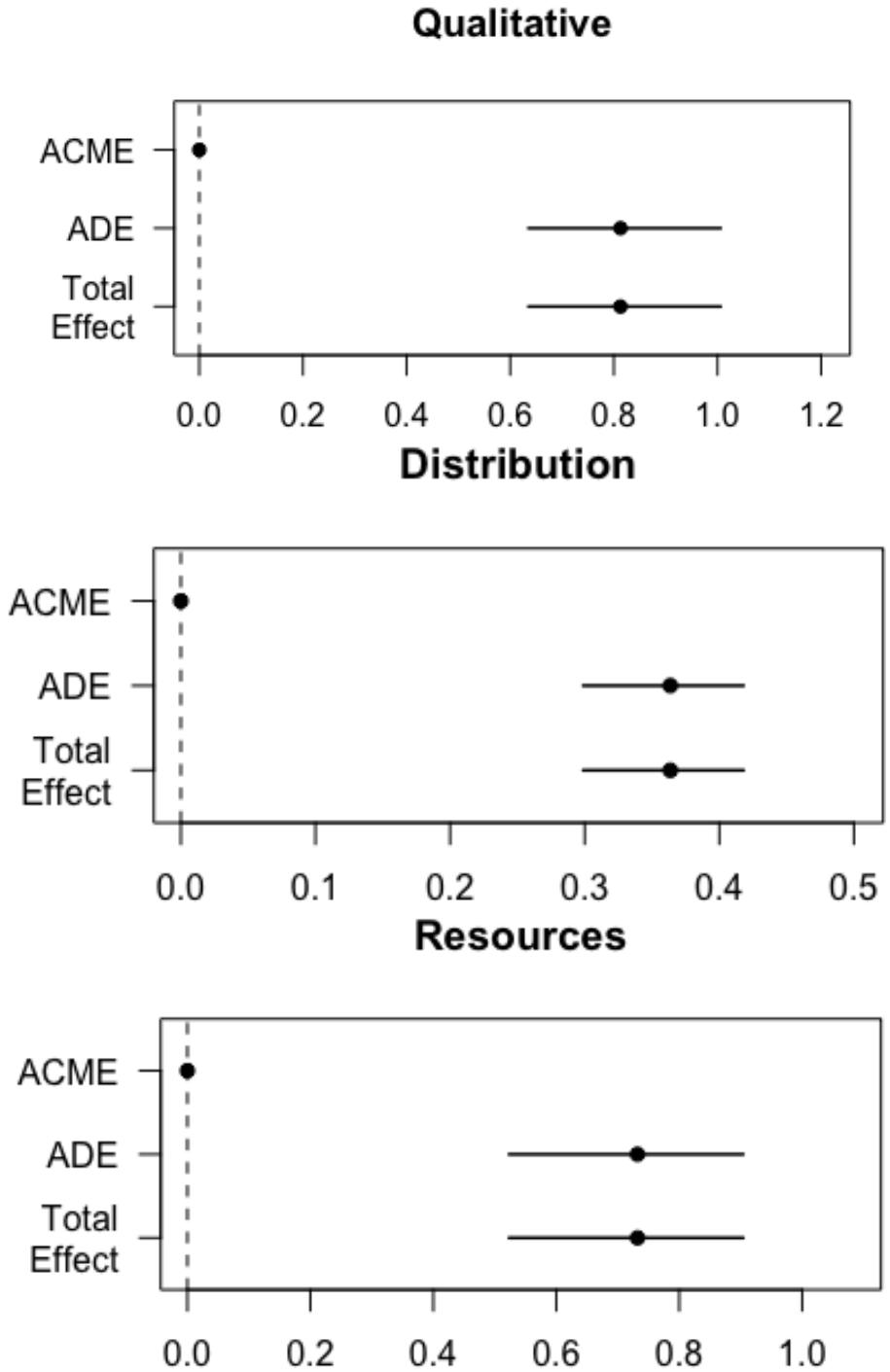
Table 2.16 – Summary information treatments

		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. (2019)	Leaflets				x	x			x		
Arias et al. (2018)	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						

Table 2.17 – Summary studies belief updating

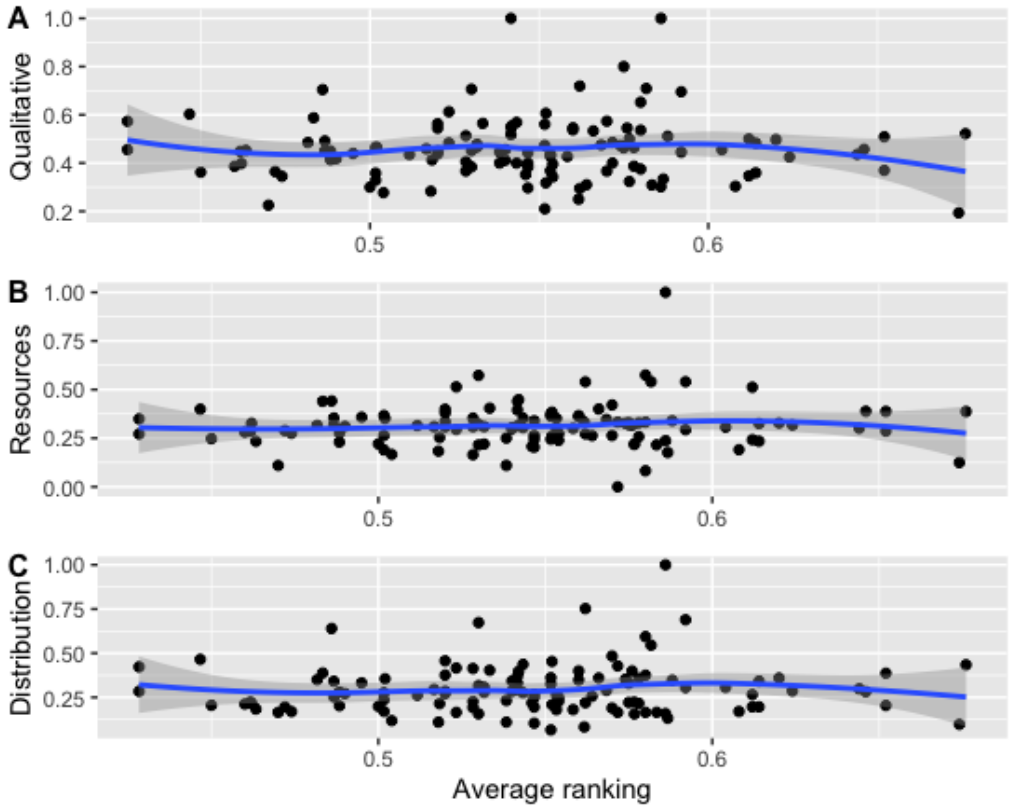
Study	Topic	Outcome	Baseline	Post-treatment	Summary measures of beliefs
Prior papers	Malfeasance	Beliefs about malfeasance	No	Yes	Belief updating is the control group once they are shown the information treatment after the post-treatment survey.
Does corruption Information	Malfeasance	Vote intention	No	Yes	Respondents were asked in a follow-up survey whether disagree or agree that their municipal government used public resources honestly.
SMS corruption	Malfeasance	Vote intention	Yes	Yes	Respondents were asked whether they will be surprised if they learn that there is corruption in their district.
Norm versus Action	Malfeasance	Vote choice	Yes	Yes	Respondents were asked how much surprised much they will be surprised if they learned about corruption in their municipality.
Greater Expectations: A Field Experiment	Performance Capacity	Vote choice/ beliefs performance			Measures learning about government capacity.
Debates: Voting and Expenditure	Political knowledge	Policy platforms candidates Vote choice	No	Yes	Belief updating about political knowledge, political alignment; relative to the control group.
Do Informed Voters Make Better Choices?	Performance	Voter Knowledge	No	Yes	

Figure 2.11 – Mediation Analysis - All Outcomes



Note: This Figure reports whether belief updating mediates subjects' behaviour of making a costly donation to their municipal government for all outcomes. We can observe that for all outcomes that donations are not mediated by belief updating.

Figure 2.12 – Relationship between CGO audits scores and priors



Note: This figure reports the relationship between the subjects' average priors within each municipality versus the average ranking score given by the CGO to each municipality. We observe a very weak relationship between average priors within each commune and historical corruption rankings generated by the CGO for all outcomes.
Source: All data wave

Table 2.18 – Complier Average Causal Effects -*Baseline* model

	Qualitative			Resources			Distribution		
	Baseline	Chatbot	Post	Baseline	Chatbot	Post	Baseline	Chatbot	Post
Prior	0.406***	0.460***	0.450***	0.412***	0.442***	0.452***	0.320***	0.404***	0.393***
	(0.019)	(0.018)	(0.018)	(0.018)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
Treat	0.733***			0.728***			0.342***		
	(0.080)			(0.091)			(0.029)		
CACE		1.036***	1.075***		1.038***	1.052***		0.522***	0.534***
		(0.090)	(0.092)		(0.102)	(0.105)		(0.035)	(0.037)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	No	No	Yes	No	No	Yes	No	No
Num.Obs.	3439	2111	1997	3439	2111	1997	3439	2111	1997
R2	0.319	0.292	0.300	0.268	0.233	0.239	0.280	0.241	0.238
R2 Adj.	0.290	0.283	0.291	0.237	0.224	0.229	0.250	0.232	0.228

Note: The following table reports the ITT and Complier Average Treatment Effects (CACE) for the *Baseline* model in Table 2.1 for all outcomes. We report two measures of treatment compliance. *Chatbot* corresponds to treatment compliance where respondents answered both attention checks questions correctly in the Chatbot survey and subsequently answered the post-treatment survey. We also report the CACE for those respondents that reported watching the video and answered correctly the attention check both in the post-treatment survey (*Post*). For the *Baseline* model, we clustered standard errors at the commune level. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.19 – Complier Average Causal Effects - Negative model

	Qualitative			Resources			Distribution		
	Negative	Chatbot	Post	Negative	Chatbot	Post	Negative	Chatbot	Post
Prior	0.466***	0.458***	0.447***	0.445***	0.441***	0.450***	0.420***	0.402***	0.390***
	(0.020)	(0.018)	(0.018)	(0.019)	(0.021)	(0.021)	(0.026)	(0.021)	(0.022)
Treat	0.575**			0.321+			0.110		
	(0.208)			(0.167)			(0.067)		
Negative	0.311	0.299	0.382+	0.132	0.040	0.073	0.031	0.024	0.081
	(0.212)	(0.194)	(0.205)	(0.174)	(0.221)	(0.234)	(0.084)	(0.074)	(0.081)
CACE		0.641**	0.703**		0.399	0.471+		0.137	0.181*
		(0.228)	(0.233)		(0.261)	(0.267)		(0.088)	(0.092)
Negative x CACE		0.449+	0.425+		0.739**	0.675*		0.446***	0.409***
		(0.248)	(0.253)		(0.283)	(0.290)		(0.095)	(0.100)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	No	No	Yes	No	No	Yes	No	No
Num.Obs.	3439	2111	1997	3439	2111	1997	3439	2111	1997
R2	0.267	0.301	0.312	0.225	0.240	0.247	0.213	0.263	0.264
R2 Adj.	0.261	0.291	0.302	0.218	0.230	0.236	0.206	0.253	0.253

Note: The following table reports the ITT and Complier Average Treatment Effects (CACE) for the *Negative* model listed in Table 2.1 for all outcomes. We report two measures of treatment compliance. *Chatbot* corresponds to treatment compliance where respondents answered both attention checks questions correctly in the Chatbot survey and subsequently answered the post-treatment survey. We also report the CACE for those respondents that reported watching the video and answered correctly the attention check both in the post-treatment survey (*Post*). For the *Negative* model, we clustered standard errors at the commune level. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.20 – Complier Average Causal Effects - Malfeasance Levels

	Qualitative			Resources			Distribution		
	Malfeasance	Chatbot	Post	Malfeasance	Chatbot	Post	Malfeasance	Chatbot	Post
Prior	0.406***	0.460***	0.451***	0.412***	0.443***	0.452***	0.321***	0.404***	0.392***
	(0.019)	(0.018)	(0.018)	(0.019)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
Log malfeasance	0.051***			0.051***			0.024***		
	(0.006)			(0.006)			(0.002)		
CACE		1.159***	1.218***		1.186***	1.203***		0.586***	0.602***
		(0.092)	(0.095)		(0.105)	(0.108)		(0.036)	(0.038)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	No	No	Yes	No	No	Yes	No	No
Num.Obs.	3439	2111	1997	3439	2111	1997	3439	2111	1997
R2	0.319	0.292	0.300	0.269	0.232	0.238	0.283	0.240	0.237
R2 Adj.	0.290	0.283	0.290	0.238	0.223	0.228	0.252	0.231	0.226

Note: The following table reports the ITT and Complier Average Treatment Effects (CACE) for the *Malfeasance* model listed in Table 2.1 for all outcomes. We report two measures of treatment compliance. *Chatbot* corresponds to treatment compliance where respondents answered both attention checks questions correctly in the Chatbot survey and subsequently answered the post-treatment survey. We also report the CACE for those respondents that reported watching the video and answered correctly the attention check both in the post-treatment survey (*Post*). For the *Malfeasance* model, we clustered standard errors at the commune level. The *Malfeasance* variable takes the value of 0 for respondents in the "Placebo" condition and it takes the value of log malfeasance found in the respective municipality for "Treated" units. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.21 – Certainty Qualitative Beliefs

	Treatment	**Control**	**All Sample**
Baseline			
High uncertainty (1-2)	16%	17%	16%
Uncertain (3 - 5)	37%	35%	36%
Certain (6 - 8)	33%	34%	34%
High certainty (9-10)	13%	14%	14%
Post-treatment			
High uncertainty (1-2)	12%	17%	14%
Uncertain (3 - 5)	36%	38%	36%
Certain (6 - 8)	36%	32%	35%
High certainty (9-10)	16%	13%	15%
Follow-up			
High uncertainty (1-2)	14%	15%	14%
Uncertain (3 - 5)	37%	39%	38%
Certain (6 - 8)	35%	32%	34%
High certainty (9-10)	13%	15%	14%

Note: The following table reports certainty about beliefs malfeasance for all three waves. Respondents were asked to report their certainty about beliefs about malfeasance using a 0-10 scale for the Qualitative outcome. In this table, we disaggregated this information by treatment status and also reported the pooled estimates. As we observed, certainty levels remain relatively the same throughout all waves, although we do observe a mild increase in certainty levels between the baseline and post-treatment for treated subjects.

Table 2.22 – Regression Results of Malfeasance Beliefs - Partisan Alignment and Partisanship Strength

	Qualitative			Resources			Distribution		
	Pooled	Strong	Weak	Pooled	Strong	Weak	Pooled	Strong	Weak
Prior	0.378***	0.374***	0.380***	0.408***	0.425***	0.399***	0.299***	0.306***	0.295***
	(0.019)	(0.029)	(0.024)	(0.018)	(0.031)	(0.021)	(0.022)	(0.031)	(0.032)
Treat	0.715***	0.667***	0.780***	0.702***	0.620***	0.743***	0.352***	0.336***	0.366***
	(0.086)	(0.116)	(0.118)	(0.112)	(0.157)	(0.145)	(0.035)	(0.047)	(0.046)
Aligned	-0.598***	-0.416*	-0.645***	-0.250+	-0.337	-0.163	-0.150**	-0.124	-0.142+
	(0.137)	(0.205)	(0.192)	(0.149)	(0.214)	(0.195)	(0.057)	(0.077)	(0.077)
Treat x Aligned	0.026	0.120	-0.075	0.083	0.246	-0.024	-0.046	-0.022	-0.067
	(0.168)	(0.277)	(0.215)	(0.166)	(0.277)	(0.254)	(0.068)	(0.108)	(0.086)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	1360	2079	3439	1360	2079	3439	1360	2079
R2	0.328	0.374	0.343	0.269	0.309	0.289	0.287	0.337	0.296
R2 Adj.	0.299	0.308	0.296	0.238	0.236	0.238	0.256	0.267	0.245

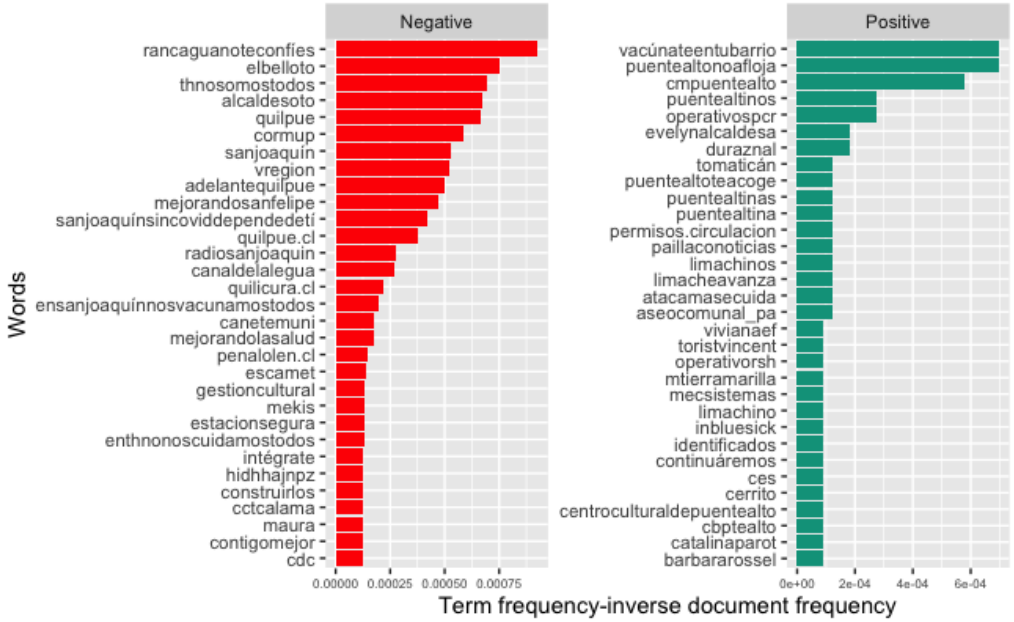
Note: This table reports regression estimates for pooled, aligned and unaligned respondents for all three outcomes. Aligned respondents correspond to subjects that reported a voting intention to support the incumbent, whereas the Unaligned respondents are subjects that reported no intention to vote for the incumbent. We provide a further breakdown by looking at aligned versus unaligned respondents and also based on their partisanship strength. We measured partisanship in our baseline survey using 0-10 scale, where 0 represents left-leaning and 10 right-leaning. We set *Strong* partisanship for all respondents reported a partisanship between 0 and 2 or 8 and 10. Whereas *Weak* partisanship for those that reported between 3 and 7. We clustered standard errors at the commune level. *For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Figure 2.13 – Screen shoots - Placebo treatment



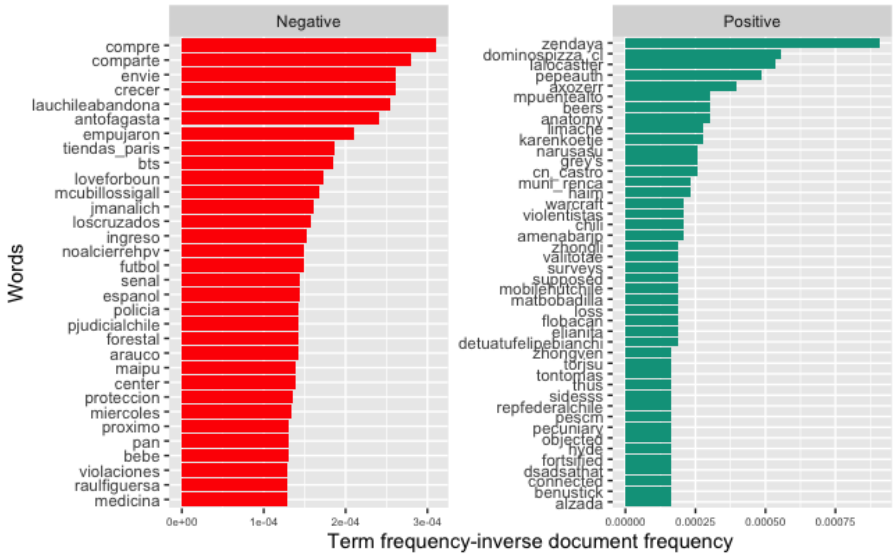
Note: This figure shows example screenshots of the placebo treatment arm. Subjects in the placebo condition received three pieces of information about their municipality. The first screenshot (left) reported the population that resides in the municipality. The second screenshot (middle) reported the population that is above the legal age (18 years old). The third (right) and final piece of information is the size of the commune in terms of squared kilometres.

Figure 2.14 – Term frequency-inverse document frequency - Local governments



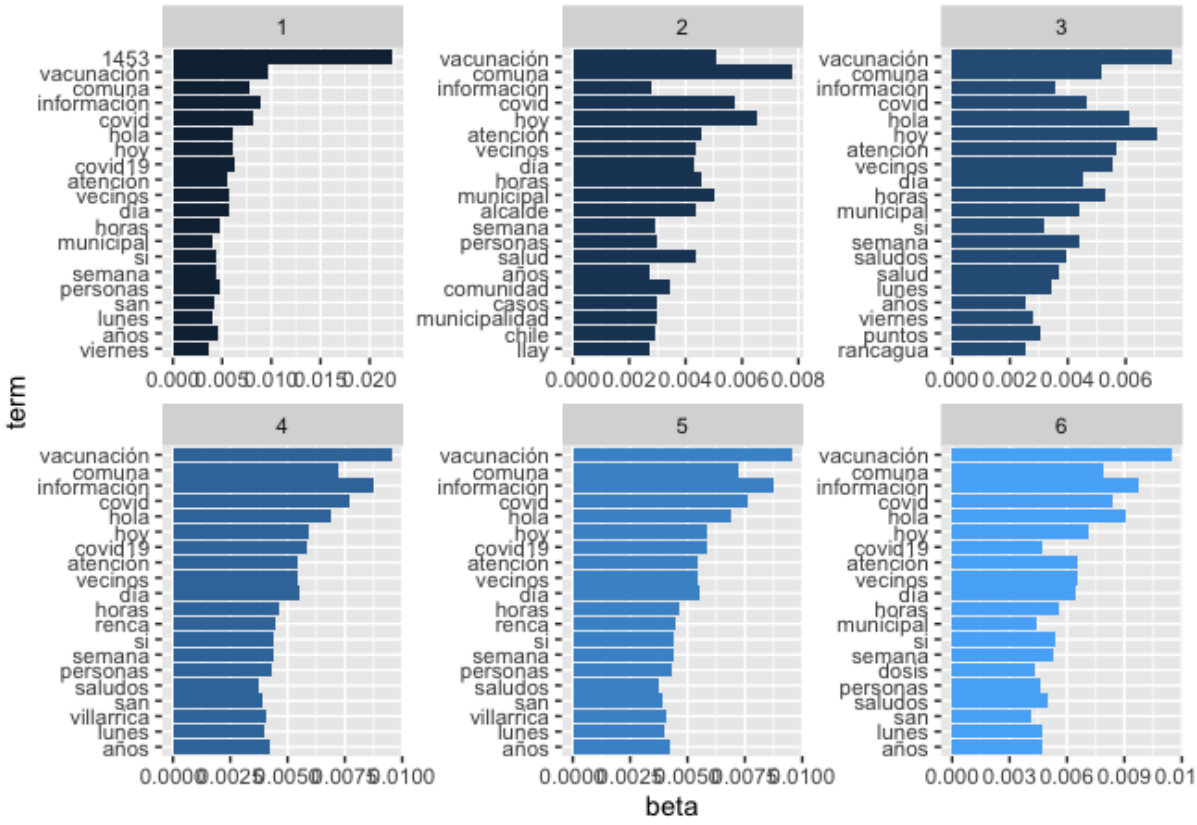
Note: This table reports the top 20 terms using term frequency-inverse document frequency generated from Tweets posted by local governments and mayors.

Figure 2.15 – Term frequency-inverse document frequency - Respondents



Note: This table reports the top 20 terms using term frequency-inverse document frequency generated from Tweets posted by respondents.

Figure 2.16 – Structural Topic modelling



Note: This figure reports the results of a structural topic modelling analysis setting K topics equal to 6. We collected Twitter data from the Twitter accounts of the 116 municipal governments and mayors. We also collected Twitter data from respondents (n = 1155) We found that most of the discussions revolves around health issues related to the COVID-19 pandemic. Some of the words identified in this analysis were for example, vaccination, Covid-19, and dose.

Table 2.23 – Regression Results of Malfeasance Beliefs - One week and One-Month Results

Panel A: Post-treatment											
	Qualitative			Resources				Distribution			
	Baseline	Negative	Malfeasance	Baseline	Negative	Malfeasance	Learning	Baseline	Negative	Malfeasance	Learning
Prior	0.406***	0.406***	0.410***	0.412***	0.412***	0.411***		0.320***	0.321***	0.319***	
	(0.019)	(0.019)	(0.019)	(0.018)	(0.019)	(0.019)		(0.021)	(0.021)	(0.022)	
Treat	0.733***	0.530*		0.728***	0.204		0.738***	0.342***	0.085		0.315***
	(0.080)	(0.233)		(0.091)	(0.186)		(0.095)	(0.029)	(0.069)		(0.031)
Negative		-3.009***			-2.545***				-0.763***		
		(0.314)			(0.281)				(0.105)		
Treat x Negative		0.238			0.613**				0.300***		
		(0.249)			(0.207)				(0.075)		
Corrup diff							-0.417***				-0.359***
							(0.028)				(0.024)
Treat x Corrup diff							0.007				0.061*
							(0.029)				(0.025)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	3439	3439	3439	3439	3439	3439	3439	3439	3439	3439
R2	0.319	0.319	0.339	0.268	0.270	0.290	0.268	0.280	0.284	0.302	0.282
R2 Adj.	0.290	0.290	0.291	0.237	0.238	0.239	0.237	0.250	0.253	0.251	0.251
Panel B: Follow-up											
Prior	0.378***	0.378***	0.349***	0.347***	0.347***	0.349***		0.265***	0.265***	0.270***	
	(0.026)	(0.026)	(0.023)	(0.023)	(0.023)	(0.023)		(0.026)	(0.026)	(0.027)	
Treat	0.650***	0.621		0.469***	0.360		0.395***	0.255***	0.124+		0.247***
	(0.127)	(0.417)		(0.120)	(0.324)		(0.116)	(0.031)	(0.063)		(0.034)
Negative		2.972***			4.074***				0.132***		
		(0.254)			(0.211)				(0.038)		
Treat x Negative		0.032			0.123				0.148*		
		(0.440)			(0.346)				(0.070)		
Corrup diff							-0.319***				-0.275***
							(0.033)				(0.031)
Treat x Corrup diff							-0.051				0.018
							(0.037)				(0.030)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	2143	2143	2143	2143	2143	2143	2143	2143	2143	2143	2143
R2	0.322	0.322	0.281	0.261	0.261	0.281	0.262	0.271	0.271	0.283	0.271
R2 Adj.	0.276	0.276	0.212	0.211	0.211	0.212	0.212	0.222	0.222	0.214	0.221

Note: This table reports the regression coefficients reported in Figure 2.5. This table reports regression estimates for all four models listed in Table 2.1. Panel reports treatment effects and interaction effects one-week after the intervention take place. Panel B reports the same estimates but for one-month after treatment. Confidence intervals were constructed using a 95% confidence level with clustered standard errors at the municipal level. For all models, we control for commune fixed effects. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.24 – Regression Results of Malfeasance Beliefs - *Negative x Malfeasance Interaction*

	Qualitative	Resources	Distribution
Prior	0.406***	0.412***	0.321***
	(0.019)	(0.019)	(0.021)
Negative	3.248***	-2.488***	-0.740***
	(0.152)	(0.281)	(0.100)
Log Malfeasance	0.045*	0.020	0.008
	(0.021)	(0.016)	(0.005)
Log Malfeasance x Negative	0.006	0.035*	0.018**
	(0.022)	(0.017)	(0.006)
Covariates	Yes	Yes	Yes
FE	Yes	Yes	Yes
Num.Obs.	3439	3439	3439
R2	0.319	0.270	0.285
R2 Adj.	0.290	0.238	0.254

Note: Estimates are presented for a fully-saturated model that includes the interaction term, *Negative* and the *Malfeasance IV*, for all three outcomes. We clustered standard errors at the commune level. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.25 – Regression Results of Malfeasance Beliefs - Priors Group

	Qualitative			Resources			Distribution		
	Pooled	High	Low	Pooled	High	Low	Pooled	High	Low
Panel A: Baseline model									
Prior	0.406***	0.372***	0.355***	0.412***	0.295***	0.515***	0.320***	0.289***	0.314***
	(0.019)	(0.035)	(0.027)	(0.018)	(0.039)	(0.053)	(0.021)	(0.053)	(0.049)
Treat	0.733***	0.756***	0.717***	0.728***	0.841***	0.672***	0.342***	0.351***	0.352***
	(0.080)	(0.129)	(0.103)	(0.091)	(0.167)	(0.103)	(0.029)	(0.050)	(0.037)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	1366	2073	3439	1366	2073	3439	1595	1844
R2	0.319	0.305	0.311	0.268	0.203	0.202	0.280	0.274	0.303
R2 Adj.	0.290	0.230	0.262	0.237	0.117	0.146	0.250	0.209	0.248
Panel B: Negative model									
Prior	0.406***	0.372***	0.355***	0.412***	0.295***	0.514***	0.321***	0.291***	0.310***
	(0.019)	(0.035)	(0.027)	(0.019)	(0.039)	(0.053)	(0.021)	(0.053)	(0.049)
Treat	0.530*	0.950*	0.299	0.204	0.729*	-0.023	0.085	0.128	0.058
	(0.233)	(0.380)	(0.352)	(0.186)	(0.318)	(0.237)	(0.069)	(0.116)	(0.070)
Negative	-3.009***	2.869***	-2.523***	-2.545***	-1.426**	0.125	-0.763***	1.433***	0.307***
	(0.314)	(0.226)	(0.144)	(0.281)	(0.511)	(0.139)	(0.105)	(0.086)	(0.053)
Treat:Negative	0.238	-0.223	0.492	0.613**	0.127	0.819**	0.300***	0.258*	0.348***
	(0.249)	(0.404)	(0.370)	(0.207)	(0.360)	(0.263)	(0.075)	(0.127)	(0.080)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	1366	2073	3439	1366	2073	3439	1595	1844
R2	0.319	0.305	0.312	0.270	0.203	0.206	0.284	0.277	0.308
R2 Adj.	0.290	0.230	0.263	0.238	0.117	0.149	0.253	0.211	0.252
Panel C: Malfeasance model									
Prior	0.406***	0.372***	0.356***	0.412***	0.295***	0.663***	0.321***	0.426***	0.311***
	(0.019)	(0.035)	(0.027)	(0.019)	(0.039)	(0.054)	(0.021)	(0.042)	(0.050)
Treat	0.337	1.164	0.058	-0.351	0.156	-0.662	-0.455**	-0.511+	-0.204
	(0.469)	(0.924)	(0.746)	(0.537)	(1.075)	(0.592)	(0.148)	(0.283)	(0.257)
Log malfeasance	0.203***	0.180***	0.224***	0.239***	0.287***	0.036	0.093***	-0.009	0.068***
	(0.028)	(0.039)	(0.033)	(0.048)	(0.072)	(0.033)	(0.010)	(0.015)	(0.014)
Treat x Log malfeasance	0.028	-0.028	0.046	0.075+	0.047	0.093*	0.055***	0.060**	0.039*
	(0.034)	(0.063)	(0.052)	(0.039)	(0.076)	(0.042)	(0.010)	(0.020)	(0.018)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	1366	2073	3439	1366	2073	3439	1595	1844
R2	0.319	0.305	0.311	0.269	0.203	0.147	0.284	0.186	0.305
R2 Adj.	0.290	0.230	0.262	0.238	0.117	0.135	0.254	0.171	0.249
Panel D: Learning model									
Prior				0.422***	0.338***	0.475***	0.373***	0.220***	-0.013
				(0.021)	(0.053)	(0.061)	(0.021)	(0.038)	(0.077)
Treat				0.738***	0.884***	0.675***	0.315***	0.357***	0.317***
				(0.095)	(0.261)	(0.104)	(0.031)	(0.046)	(0.049)
Diff				0.006	0.035	-0.059**	0.014	-0.128***	-0.349***
				(0.018)	(0.055)	(0.020)	(0.016)	(0.033)	(0.073)
Treat x Diff				0.007	0.012	0.033	0.061*	0.101*	0.039
				(0.029)	(0.057)	(0.047)	(0.025)	(0.050)	(0.048)
Covariates				Yes	Yes	Yes	Yes	Yes	Yes
FE				Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.				3439	1366	2073	3439	1595	1844
R2				0.268	0.203	0.202	0.282	0.278	0.303
R2 Adj.				0.237	0.117	0.145	0.251	0.212	0.248

Note: This Table reports regression estimates based on respondents' prior beliefs about malfeasance in their commune. We report estimates for all three outcomes for three subsets of respondents: "Pooled", "High", and "Low". "Pooled" estimates correspond to all respondents. The "High" group are a subset of respondents that reported beliefs that were higher levels of malfeasance than the average subjects within the same commune, for each particular outcome. Conversely, the "Low" group corresponds to respondents that reported beliefs about malfeasance below their respective municipal average. We provide the results for all four models outlined in Table 2.1 in four separate panels. We clustered standard errors at the commune level. $+p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < 0.001$.

Table 2.26 – Regression Results of Malfeasance Beliefs - Partisan Alignment

	Qualitative		Resources		Distribution	
	Aligned	Unaligned	Aligned	Unaligned	Aligned	Unaligned
Panel A: Baseline model						
Intercept	2.061	3.144***	2.877	2.125***	1.058*	1.166***
	(1.115)	(0.393)	(1.885)	(0.533)	(0.401)	(0.153)
Prior	0.424***	0.415***	0.414***	0.435***	0.382***	0.381***
	(0.037)	(0.023)	(0.038)	(0.019)	(0.050)	(0.026)
Treat	0.742***	0.733***	0.788***	0.723***	0.303***	0.357***
	(0.152)	(0.083)	(0.135)	(0.107)	(0.056)	(0.034)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
FE	No	No	No	No	No	No
Num.Obs.	929	2510	929	2510	929	2510
R2	0.195	0.233	0.219	0.215	0.158	0.193
R2 Adj.	0.172	0.224	0.197	0.206	0.134	0.185
Panel B: Negative model						
Intercept	1.483	3.121***	2.471	2.236***	0.926*	1.223***
	(1.151)	(0.417)	(1.864)	(0.517)	(0.388)	(0.178)
Prior	0.424***	0.416***	0.414***	0.435***	0.377***	0.382***
	(0.036)	(0.022)	(0.038)	(0.020)	(0.048)	(0.028)
Treat	0.908	0.362*	0.471*	0.213	0.143	0.080
	(0.498)	(0.139)	(0.186)	(0.251)	(0.100)	(0.101)
Negative	0.710*	0.036	0.494*	-0.106	0.169	-0.055
	(0.292)	(0.234)	(0.207)	(0.246)	(0.106)	(0.113)
Treat:Negative	-0.217	0.430*	0.371	0.590*	0.189	0.320**
	(0.528)	(0.170)	(0.247)	(0.275)	(0.120)	(0.107)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
FE	No	No	No	No	No	No
Num.Obs.	929	2510	929	2510	929	2510
R2	0.204	0.236	0.231	0.217	0.175	0.201
R2 Adj.	0.179	0.227	0.207	0.208	0.149	0.192
Panel B: Malfeasance model						
Intercept	2.033	3.118***	2.837	2.091***	1.038*	1.155***
	(1.111)	(0.397)	(1.890)	(0.531)	(0.403)	(0.154)
Prior	0.421***	0.417***	0.410***	0.435***	0.376***	0.381***
	(0.037)	(0.022)	(0.038)	(0.019)	(0.049)	(0.027)
Malfeasance IV	0.053***	0.054***	0.058***	0.054***	0.023***	0.026***
	(0.010)	(0.006)	(0.010)	(0.007)	(0.004)	(0.002)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
FE	No	No	No	No	No	No
No	No	No				
Num.Obs.	929	2510	929	2510	929	2510
R2	0.198	0.237	0.223	0.219	0.166	0.201
FE	No	No	No	No	No	No

Note: This table reports the regression estimates reported Panel B in Figure 2.6 for the *Baseline*, *Negative*, and *Malfeasance* models for all three outcomes. We break down the sample into two groups: "Aligned" and "Unaligned". "Aligned" respondents correspond to subjects that reported in our baseline survey their intention to vote for the incumbent in the upcoming municipal election. Conversely, the "Unaligned" group comprises all respondents that declared their intention of not supporting the incumbent candidate. Results for the same analysis of the *Learning* model are reported in Table 2.5 panel B. We clustered standard errors at the commune level. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

Table 2.27 – Treatment Assignment by Commune

Commune	Treatment	Control	Total
Alto Hospicio	33	20	53
Ancud	23	23	46
Angol	30	14	44
Antofagasta	32	21	53
Arica	42	20	62
Cabildo	8	4	12
Calama	36	12	48
Caldera	16	6	22
Canete	21	10	31
Cartagena	11	13	24
Cerrillos	54	23	77
Cerro Navia	44	24	68
Chillan	46	27	73
Chillan Viejo	12	14	26
Chimbarongo	26	12	38
Cisnes	1	0	1
Colchane	1	0	1
Colina	35	15	50
Combarbala	8	3	11
Conchali	49	14	63
Concon	25	17	42
Copiapo	29	18	47
Coquimbo	49	25	74
Coronel	37	37	74
Curico	42	26	68
El Bosque	32	14	46
El Quisco	17	7	24
El Tabo	11	5	16
Estacion central	66	18	84
Freirina	8	2	10
Frutillar	16	5	21
Huechuraba	40	11	51
Independencia	43	19	62
Iquique	29	17	46
La Cisterna	43	24	67
La Cruz	17	14	31
La Florida	92	40	132
La Granja	27	17	44
La Reina	37	16	53
La Serena	49	26	75
La Union	26	10	36
Lampa	26	21	47
Las Condes	53	25	78
Lautaro	22	10	32
Limache	45	15	60
Linares	37	19	56
Lo Barnechea	38	11	49
Lo Espejo	32	13	45
Los Alamos	7	3	10
Los Andes	28	20	48
Los Angeles	60	22	82
Los Lagos	7	1	8
Los Vilos	9	5	14
Lota	25	15	40
Macul	38	33	71
Maipu	44	24	68
Melipilla	36	12	48
Molina	29	13	42
Mostazal	12	11	23
Mulchen	22	16	38
Natales	18	4	22
Negrete	0	1	1
Nueva Imperial	21	10	31
Nunoa	92	29	121
Osorno	44	23	67
Ovalle	32	17	49
Paillaco	8	4	12
Paredones	2	0	2
Parral	21	14	35
Pedro Aguirre Cerda	43	21	64
Penaflor	43	16	59
Penalolen	70	21	91
Porvenir	3	1	4
Pozo Almonte	5	0	5
Providencia	79	34	113
Pudahuel	67	27	94
Puente Alto	63	26	89
Puerto Octay	3	1	4
Punta Arenas	62	23	85
Quilicura	62	44	106
Quilpue	33	35	68
Quinta Normal	45	17	62
Rancagua	40	26	66
Recoleta	44	25	69
Renca	36	17	53
Salamanca	9	5	14
San Bernardo	56	24	80
San Clemente	16	11	27
San Felipe	35	19	54
San Fernando	51	14	65
San Ignacio	5	1	6
San Javier	24	10	34
San Joaquin	46	18	64
San Jose de Maipo	19	12	31
San Miguel	59	29	88
San Ramon	41	22	63
San Rosendo	0	3	3
San Vicente	26	18	44
Santa Juana	2	4	6
Santiago	107	65	172
Talca	45	27	72
Talcahuano	43	20	63
Taltal	3	1	4
Temuco	44	29	73
Tierra Amarilla	2	1	3
Tiltil	15	5	20
Tocopilla	11	6	17
Tome	28	16	44
Tucapel	4	2	6
Valdivia	83	36	119
Vallenar	30	14	44
Victoria	22	14	36
Vicuna	13	2	15
Villarrica	48	19	67
Vitacura	35	15	50
Zapallar	5	2	7

Chapter 3

Paper 2: Choice Architecture for Informing Corruption: Evidence from an Adaptive Experiment

3.1 Motivation

Extensive survey 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 overwhelming empirical evidence of the costs to society associated with rent-seeking in public and private sectors (Fishman and Golden, 2017). There is some consensus amongst scholars that corruption is a salient issue for the average individual (Polk et al., 2017; Engler, 2020). It would be reasonable to expect that citizens would pay attention to news or information about corruption.

There is extensive literature claiming that information about malfeasance generated from government audit reports affects the political preferences of the average voter. For example, 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). This accrued evidence liking information with accountability raises several questions about the information chain through which citizens learned about their elected officials' wrongdoing.

Experimental work linking corruption information to voter behaviour relies on some messaging template or strategy. For example, Buntaine and Daniels (2020) study the effect of audit information on vote outcomes; they condense the voluminous findings from the Uganda Auditor-General report into brief comparative information they deliver to voters via SMS text messages. Specifically, they provided information about "how the percentage of unaccounted-for funds in a citizen's district compared with unaccounted-for funds elsewhere." Their bad/good comparative metric was whether a district fell above or below the district with the median number of

irregularities. The authors' approach to convey poor performance was to provide messages reporting the total cost of the irregularities relative to the total budget size. They also expressed irregularities as the number of UGX out of 1000 UGX had issues. This would suggest that there were no sufficient efforts by the authors to tease out effective versus ineffective messaging strategies. They essentially broadcast multiple (12) assessment messages to the treated districts, which could partly explain why the authors yield virtually null effects.

From the previous literature outlined, it is difficult to know whether null (or significant) effects are an artefact of how information is assembled and presented to voters. This raises several issues: What are the 'candidates' of possible corruption messages? Is the range of plausible corruption messages conditioned by context – so the plausible malfeasance messages will almost certainly look different in Nigeria to in Russia? The challenge here is then to identify the content and format of delivery that will be most accessible and persuasive for the average citizen.

It is paramount to think carefully about how to inform citizens, given that there is evidence that providing "noisy" signals can shape citizens' and institutions' preferences and behaviour in different paths. Aoyagi, Fréchette and Yuksel (2021) find that economic agents' decision to sustain cooperation in a repeated prisoner's dilemma varies on whether agents received *perfect* or *noisy* information about others' players actions. Similarly, Fu (2014) evaluates how colleges make costly decisions that rest upon "imperfect" signals of students' ability. The author finds a considerable mismatch between college admission offers and students' abilities and preferences. Thus, information should be digestible but should also should help citizens to update their beliefs in the right direction and reduce uncertainty about the state of the world.

Several countries have set up institutions to increase oversight within the public sector. Many of these agencies have adopted audits as a policy designed to monitor public institutions. Their focus is to evaluate compliance of government officials with norms, statutory rules and financial regulations. These agencies are primarily concerned with the financial consequences of corrupt practices, and this information, is not always, is made public. Therefore, the data generated by these organisations narrows down the informational space of the plausible messaging campaigns.

The rich informational content of these audit reports constitutes the puzzle of this essay. Most field or survey 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 3.1 synthesises the message content or format of the malfeasance information treatments used in these experiments. I identify ten specific dimensions organised into two general categories: metrics and benchmarking. As Table 3.1 suggests, these studies have little consistency in terms of the messaging strategies, contents, and metrics.

Table 3.1 – 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						

In this essay, I aim to evaluate which features of information tend to resonate more with the average citizen. Rather than looking at beliefs or political outcomes. This study seeks to identify which information features tend to be more compelling to the average citizen. The main contribution of this paper is to the growing literature that looks at the effects of information about malfeasance upon political accountability. Specifically, it provides evidence regarding which pieces of information tend to be more compelling in terms of citizens engaging. More broadly, it also provides evidence that links information with electoral accountability, but specifically on producing effective and telling information strategies about political actors and institutions.

The second contribution of this essay is to the research within the field of behavioural science on the effects of framing. This research has broadly studied whether conveying the same information in a loss or gain frame significantly affects individuals' behaviour of interest. In this study, I give empirical evidence to demonstrate whether individuals find more compelling information that incorporates several frames versus standard "default" information messaging strategies.

The third contribution of this essay 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 optimisation and reinforcement learning that have been mainly implemented within the field of computer science. This also provides an example for using adaptive experiments for studies with two competing goals: best-arm identification and drawing inferences from point estimates.

This essay continues as follows. Section 3.2 discusses the literature on information frames, particularly on the topic of corruption. Section 3.3 explains the experimental design implemented in this study. I then elaborate on the treatments included in this study and the selection of outcome measures, and the adaptive component of the experiment. Section 3.5 summarises the empirical findings of the study, using a number of estimation strategies. Finally, in Section 3.6, I summarised the main findings and limitations of this study.

3.2 Literature Review

Education and Choice Architecture An essential piece of information that these 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 I have gathered from the literature (see Table 3.1), the information about malfeasance has been broken down 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 behaviour. The table 3.1 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 that have been used in 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 the malfeasance as foregone expenditures on a particular public good. For example, Arias et al. (2019) 's primary treatment expresses the magnitude of malfeasance in terms of foregone spending on social programmes. Similarly, Larreguy, Marshall and Snyder (2020)'s experiment reports the amount of funds spent on unauthorised projects. All three metrics convey different dimensions and features of wrongdoing mainly at the local level.

Benchmarks. Table 3.1 also highlights those field and survey experiments that incorporate some form of benchmark as part of their information strategy. The reasoning here is that voters can better distinguish whether the levels of malfeasance which their local government are high or low if they have a relevant reference point. These benchmarks signal to individuals the extent to which municipalities deviate 1) from other municipalities ("spatial" benchmark) or 2) from the municipality's levels of malfeasance in previous years ("temporal" benchmark). These benchmarking measurement strategies build on extensive theoretical and empirical literature highlighting the importance of retrospection and comparisons for vote choice (Ferejohn, 1986; Fiorina, 2006; Besley, 2006; Besley and Case, 1995; Duch and Stevenson, 2008; Healy and Mal-

hotra, 2013). Thus, it is the researchers' challenge to bring to light the precise benchmarking mechanism (Healy and Lenz, 2014; Kayser and Peress, 2012; Duch and Stevenson, 2008; Healy and Malhotra, 2013).

Recent efforts to calibrate the impact of malfeasance messaging on political accountability have given careful attention to this notion of benchmarking. The studies in Table 2.16 adopt varying benchmarking approaches, and there is no consistent set of outcomes that would lead us to settle on a single benchmarking information frame. Lierl and Holmlund (2019) introduce a temporal frame by comparing current performance to that of the previous municipal government. Additionally, their spatial information frame is to compare the subject's municipality to an overall national benchmark. Neither of the frames affected their outcome variables: voter turnout and voter preference. Bhandari, Larreguy and Marshall (2019), find that temporal benchmarking (comparing an incumbent's current performance to that of the district's previous elected officials) affects updating. In contrast, they report that spatial benchmarking (comparing an incumbent's current performance with the performance of other, currently elected officials in other districts) has a weak effect on accountability.

We can also link benchmarking also links to literature highlighting reference points. Here, individuals evaluate outcomes and express preferences relative to an existing reference point. Kahneman and Tversky (1979)'s prospect theory states that the carrier of utility changes in outcome relative to a reference point, rather than the absolute value of those same outcomes. There are several proposals that individuals can use as their reference point(s). These reference points can include the status quo (Kahneman, 1992), expectations (Augenblick and Rabin, 2021), to even social comparisons within and between groups (Hargreaves Heap, Ramalingam and Rojo Arjona, 2017). Thus, it would be reasonable to expect that information treatment with benchmarks may change the salience of some reference points over others.

Experimental Approach. Often, scholars will utilise a singular approach when testing multiple treatments in one experiment (Muralidharan, Romero and Wüthrich, 2019). One of the biggest limitations of these designs is that scientists need to acknowledge that there might be the possibility of assigning experimental units to less promising interventions (Scott, 2010). In this way, allocating to a 'poor' or 'less informative' treatment arms. Thus, there is a need to establish a balance between drawing inference of the effect of each treatment arm, but also maximising efficiency.

Scientists can implement a design-based approach to minimise squandering units in optimal treatment arms by sequentially changing treatment assignment probabilities based on the study's goals. Tech firms and scholars have been implementing *adaptive experiments* that allow to adjust features of the experiment during its roll-out in order to maximise the researcher's goals. Thus, one of main benefits of adaptive designs is that they allow the user to find the most suitable treatment arm(s) according to their goals, but with fewer observations, which therefore makes it more efficient Pallmann et al. (2018).

3.3 Experimental Design

To evaluate what frames and content of information metrics are more compelling, I conducted an online survey experiment that has three critical components: 1) The medium to deliver the information treatment are short 50-second videos. 2) An adaptive sampling strategy that sequentially modifies treatment assignment shares based on the performance of each arm. 3) A behavioural information-seeking outcome that seeks to differentiate which frames and pieces of content about corruption are evaluated to be more compelling.

Video Treatments The information treatments are designed using historical data from audit reports conducted 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 found in local governments, such as the amount of malfeasance and the number of irregularities found. 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.

Field experiments that include information treatments notifying citizens about malfeasance have found minimal or zero effects on belief updating or voting outcomes Dunning, Grossman, Humphreys, Hyde, McIntosh and Nellis (2019). The challenge with these information treatments is that the average citizen may have little interest in or understanding of many malfeasance aspects of malfeasance. Thus, this study aims to address this challenge by evaluating a reasonable set of malfeasance themes and framing strategies that are most likely to be persuasive.

Table 3.2 summarises the malfeasance message content and format included in the experiments. There will be two dimensions implemented in the experimental design: benchmarks and metrics. The first dimension is *benchmarks*, which has two levels. In the *spatial* level, respondents will learn the amount of malfeasance of the hypothetical municipality compared to those of 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 3.10 and 3.12.

The second dimension, *metrics*, has three levels: The first is *Individual*, which expresses the cost of malfeasance as a the share of every $1,000 Chilean pesos of the municipality's spending budget. The second metric is *Foregone Loses*, 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.

As mentioned previously, the basis of the information included in the treatments is data from audit reports generated by the CGO. These reports are typically over 40 pages long and have highly detailed and complex information about the irregularities found in local governments. Because of the extent of these audit reports, local NGOs typically report simpler and summarised versions of these reports. The intuition here is that these summarised versions are more accessible to the a layperson and facilitate learning. Given that the implementation of this study was

in partnership with the local NGO Chile Transparente (a non-partisan NGO that focuses on corruption in Chile), I incorporated some design components into the video treatments.

Table 3.2 – Description of Treatment Factors

	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 in a 50-second video that reported the following information in this order: 1) Study participants first learnt that there was corruption in a municipality, accompanied by the total amount of malfeasance expressed in Chilean pesos for the *Individual* and *Foregone* treatments. For the *Standard* treatments, the first statistic provided to subjects was the total number of irregularities found.

In the second part of the video, respondents received the benchmark component of the treatment. For the spatial level, the video compared the amount of malfeasance found in the hypothetical municipality to other municipalities in the same region. For the Temporal level, the video reports a time series of the total malfeasance found in the hypothetical municipality. The video showed the amount of malfeasance found in the latest audit versus the previous one.

At the end of the video, respondents saw the *Metrics* component of the treatment. Respondents assigned to the 'Individual' metric level received the amount of malfeasance represented as the amount out $1,000 spent by the municipality. Subjects assigned to the 'Resources' were informed of the amount of malfeasance expressed as the number of influenza vaccines lost due to corruption. As Table 3.1 summarises, the literature has not generated any strong priors regarding what are effective messaging strategies.

An increasing number of online and field experiments have used videos as the medium to deliver their information interventions. For example, Dunning, Grossman, Humphreys, Hyde, McIntosh and Nellis (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 of choosing this medium is two-fold: Firstly, from a previous survey, Chileans reported getting political

information using platforms such as WhatsApp, and mainly got this information in a video format[1]. Secondly, Wittenberg et al. (2021) find that videos are modestly more persuasive than written material when it comes to communicating political issues to citizens.

I recruited respondents from the Nuffield Centre for Experimental Social Sciences's panel in Chile. This panel consists of approximately 57,708 respondents in Chile that had been recruited via Facebook, Twitter, Instagram Ads, and other traditional forms of subject recruitment such as posters and leaflets. Respondents were randomly sample from this pool. A number of covariates were collected in the survey in order to take into consideration respondents' demographic information such as gender, age, education, and income.

Behavioural and attitudinal outcomes In this study, I sought to evaluate which frames and metrics of information about malfeasance are more compelling to the average citizen. It is reasonable to think that a particular messaging strategy could hold the interest and attention of citizens if it is sufficiently persuasive and clear. Within the literature on political persuasion, Druckman (2022) outlines that persuasion is a rather complex and multifaceted concept that relies on several factors. For example, who are the speakers and receivers of the message, what is the context of the message, and what are the speaker's goals[2].

Several studies have used information seeking-behaviour as a behavioural measure to capture the strength of persuasiveness a particular information treatment. The logic of using this outcome is that a sufficiently strong stimulus generates interest in the citizen, prompting them to learn more about the political subject in question. In this experiment, I measured how compelling each frame and metric is by using a binary outcome that captured individuals' interest in learning more about corruption in their local constituency. Respondents were asked at the end of the questionnaire whether they wanted to finish the survey or whether they wanted to know about the irregularities found 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 particular stage of the study aimed to replicate the scenario where citizens may (or may not) be interested in learning the details of corruption scandals covered in different media outlets. For example, citizens may be interested in acquiring more information, such as the magnitude of these scandals, to know the political actors involved (Di Tella and Franceschelli, 2011; Stanig, 2015; Le Moglie and Turati, 2019). Figure 3.16 shows how I embedded this outcome into the survey. Figure 3.17 displays screenshots of the information shown on the Audit agency's website.

Similar experimental studies have used click-through rates (CTRs) and other information-seeking behaviour as the primary behavioural outcome. Singh and Roy (2018) conducted an online survey experiment where they looked at how respondents changed their information-gathering behaviour after being informed of results from opinion polls. In their study, respondents were allowed to browse various information boards displaying the parties' positions on the economy and environment. The author measured the level of engagement by assessing whether

[1]This comes from the baseline survey that I conducted for paper 1

[2]For instance, the speaker's goals can be to alter the receiver's motivations, matching the receiver's goals, or shape the receiver's understanding of topics.

respondents clicked on these websites and how much 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 as a way to measure engagement [3].

Other research areas within the field of political science have also used CTRs of websites or social media posts as their primary behavioural measure. Vosoughi, Roy and Aral (2018) analysed 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. The research on political participation and political engagement has also extended beyond conventional outcomes such as voting and party participation. This literature has broadened its range of measures of political engagement to more 'passive' or arm-length forms of participation, such as signing petitions or subscribing to newsletters (Porten-Cheé et al., 2021).

Choosing the right outcome is an essential step in adaptive experiments. This is because treatment assignment probabilities are a function of the success rate of each treatment arm. In this experiment, the primary outcome is click-through rates. Thus, the messaging strategies with higher click-through rates yield higher treatment assignment probabilities. A more suitable approach to measure persuasiveness may be using a composite index that combines behavioural and attitudinal outcomes. However, most advances within the literature on adaptive experiments prioritise 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 only use a binary outcome.

In addition, respondents answered a short number of questions immediately after they watched the videos. They were asked whether they agreed or disagreed if the video treatments were: 1) convincing, 2) trustworthy, 3) reliable and 4) and clear. These attitudinal outcomes have been used within the field of political communication as a way of examining the persuasiveness of different messaging strategies. O'Keefe (2021) points out that one approach to measure 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*. The author also finds that attitudinal measures help assess how compelling the different messages are. He compiled data from several meta-analyses that studied the persuasiveness of information strategies that included attitudinal and behavioural outcomes. The author concludes that these two outcomes tend to be strongly correlated when it comes to the direction of the effects, the magnitude of their effect sizes, and whether both yield statistical significance.

In terms of the items included in studies that examine perceived persuasiveness, Thomas, Masthoff and Oren (2019) find that several studies include questions that assess *whether the content is convincing*, *important to me* or *believable*. As pointed out before, I included four items with two rating scales, i.e. whether they *Agree* or *Disagree* on the statements listed before.

[3]Within the literature of digital marketing 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

Although this rating scale simplifies respondents' assessment of the information treatments, there is a loss in capturing a higher degree of measurement precision of the assessment of the information treatments [4].

Along with the behavioural and perceived measures, I also included two stated preferences type-of-questions. Respondents were asked, if the video contained 'actual' information about their local government, whether 1) they would be willing to share the video with friends, family or colleagues, and 2) they would be willing to post on any social media platform that they use.

After these questions, subjects were asked three attention check questions. The first attention-check question asked about the main topic of the video. The second and third questions sought to identify whether individuals were able to recall the bench-marking and the metric components of the videos. Table 3.14 in the Appendix summarises the proportion of study participants that answered correctly to each of the questions and to all three questions.

Adaptive Design A limitation that researchers face in this type of studies is that they can rarely test the wide but finite range of factors they are interested in. Researchers thus face a trade-off between exploring as many factors as possible versus providing well-powered and precise estimates of each treatment arm. In this scenario, adaptive experiments may ease this search and efficiently discern the best-performing treatment arm(s). In this design, treatment assignment is 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).

Recently, the literature on adaptive experiments builds from the research developed within the field of *Reinforcement learning*. This exploration/exploitation trade-off is often referred to as the "multi-armed bandit problem", where a researcher/policy-maker tests multiple interventions but only sequentially observes noisy signals of the quality of each policy (Russo, 2020). Scientists within this field have put forward different algorithms that balance exploration and exploitation of the different policies/interventions.

There has been a relatively small but increasing number of adaptive examples within economics and political science. Bahety et al. (2021) closely relates to this study. They conducted a ten-batch experiment sending SMS texts that sought to encourage people to comply with social distancing and hand-washing in India during the Covid pandemic.[5]. Esposito and Sautmann (2022) conducted a two-wave adaptive experiment testing six different forms of automated calls in Kenya to encourage parents to read to their children at home. Caria et al. (2020) carried out an adaptive experiment to determine the welfare-maximising employment policies to help Syrian refugees and local job-seekers to find salaried employment.

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

[5]Their experimental design is very similar to this study, their project comprised 10 information arms divided in 10 waves

Algorithm selection Within the multi-arm bandit literature, there exist a wide range of algorithms that deal with the trade-off based on the researcher's goals[6]. Some of these goals can be maximising treatment effects estimates, minimising regret[7] to maximise participants' welfare gains. These algorithms find equilibrium in different ways how confidently they identify the arm(s) that are optimal to the researchers' goals, while at the same time, exploring the remaining competing arms.

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 is that it explores all treatment arms to gather more information about their quality. But 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 that arm is the best.

We can formalise how the Thompson sampling algorithm works as it reported in Offer-Westort, Coppock and Green and Russo expressed in equations 3.1 to 3.3. 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 $FX|(\theta^k)$.

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta, a)p(\theta) \tag{3.1}$$

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:

$$f_{\Theta|\theta_{n^k_{k,t}}}(\theta_k|x_{n^k_{k,t}}) \tag{3.2}$$

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 formalise the selection of the highest probability of success for each treatment as follows:

$$P\left[\Theta_k = \max_k\{\Theta_1, ...\Theta_k|(X_1^{n_{1,t}}, ..., X_K^{n_{k,t}})\right] \tag{3.3}$$

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

[6]There are several other well-known methods and heuristics such as Equal allocation, Greedy, Softmax learning, Upper confidence bound, Gittins index, and top-two Thompson sampling

[7]*Regret* refers in the computer science jargon to the difference between the expected reward of the *best-arm* over the actual assignment conducted in the experiment. The *Regret* indicator is an indicator of performance that helps assess the algorithm's efficiency in trade exploration and exploitation.

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 3.3.

Table 3.3 – 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 $=_{\epsilon} 1...K*$ 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

Based on the previous evidence from the simulations and pilot tests reported in the Appendix, I used a modified version of the Thompson sampling algorithm proposed by (Kasy and Sautmann, 2021) called *Exploration Sampling*. This algorithm sufficiently balances exploration/exploitation by gradually switching sampling away from those treatment arms that are unlikely to be optimal [8]. Kasy and Sautmann 2021 show that their proposed algorithm converges at the same rate as their posterior probability. If a treatment converges faster, the algorithm reduces or stops assigning units to that treatment, allowing competing treatments to catch up. This feature increases expected welfare; it limits assigning more than the fiftieth per cent of experimental to the highest treatment arm. Furthermore, this algorithm avoids stopping early from learning from sub-optimal treatments. In their modified version, $q_{k,t}$ shares of experimental units are assigned to each treatment arm k, where equations 3.4 and 3.5 show how $q_{k,t}$ is computed.

$$q_{k,t} = S_t \cdot p_{k,t}(1 - p_{k,t}) \tag{3.4}$$

$$p_{k,t} = P_t\left(k = \arg\max_{k'} \theta^{k'}\right) \text{ and } S_t = \frac{1}{\sum_k p_{k,t} \cdot (1 - p_{k,t})} \tag{3.5}$$

Batch size and number of batches. The experiment was conducted in eleven batches. The first batch was comprised of 200 observations, whereas the remaining ten batches consisted of 100 observations. The total sample size of the study was 1,200. I increased the size of the first batch based on the simulations reported in Figure 3.6 and the results from the pilot conducted before the study. Having a larger first batch size should yield more accurate estimates of the quality/performance of each arm from the outset. However, there is a compromise of having a

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

larger first batch size as optimal treatments that may emerge in the last batches, but they may not have sufficient observations. More details of the first batch size and RMSE are reported from Figures 3.7 to 3.9. Furthermore, results from the pilot show that poor-performing treatment arms in the first batch are severely punished with treatment assignment shares for the second batch, getting treatment assignment probabilities below 1%. Table 3.16 compares the assignment shares using a first batch of 100 observations versus 200 observations from a pilot that I conducted before carrying out the experiment.

Bayesian updating and optimisation. After each batch was completed, posterior probabilities were updated. In the first batch, treatment assignment probabilities for each arm were equal to one-sixth. This is equivalent to imposing an uninformed prior that all treatment arms have the same posterior probability of being the best. I used Monte Carlo sampling to approximate the actual value of the posterior probability $p_{\alpha_k,t+1}$. Finally, experimental units were randomly allocated across arms using exploration sampling shares $q_{k,t}$ [9].

3.4 Treatment Effects Estimation and Inference

Estimating treatment effects for adaptively collected data may require making necessary estimation adjustments to obtain unbiased estimates of the difference in means estimator (Shin, Ramdas and Rinaldo, 2019; Rafferty, Ying and Williams, 2019). This is because treatment assignment probabilities are updated based on previous outcomes (Nie et al., 2018; Pallmann et al., 2018). Furthermore, adaptively collected data can generate non-normal and heavy-tailed sampling distributions (Hadad et al., 2021).

However, one of the advantages of using *Exploration sampling* is that the sampling bias generated from adaptively collected data using this algorithm should be negligible. As Kasy and Sautmann point out, treatment assignment probabilities for sub-optimal treatment 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 inference of treatment effects. However, as Bahety et al. 2021 indicate given that some treatment conditions may end up with fewer observations, randomisation inference would be a suitable approach to dealt with possible high-leveraging observations Caria et al. 2020 follow this same approach of providing randomisation-based p-values under a sharp null hypothesis of no treatment effects [10].

Inference for adaptive experiments is a relatively new and active area of study. Zhang, Janson 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 their BOLS estimator follows an asymptotically normal distribution. Further-

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

[10] high leveraging observations affect the variance of the estimate. A second approach is to use some weighting technique, followed by bootstrapping.

more, their simulations show that their estimation strategy significantly outperforms similar approaches in reducing a type-I error[11].

Along with providing estimates from the previously outlined estimators, I 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. [12].

Equation 3.6 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 effect for each treatment arm.

$$Y_i = \beta_{0t} + \sum_{k=1}^{5} \beta_j T_{ij} + \omega X_i + \epsilon_i \tag{3.6}$$

The outcome variable Y_i is a binary outcome of whether respondents did clicked or did not clicked to get information regarding the levels of malfeasance within their local government. Finally, X_i is a vector of covariates such as age, gender, and education. Given that this design did not incorporate a pure control, the reference condition is the *Standard Spatial* arm. Although this treatment condition is amongst the most-promising treatment arms, this is based on my prior belief that this condition would yield the lowest CTR as it is the least sophisticated messaging strategy. Furthermore, this is the reference category set in the pre-analysis plan[13].

3.5 Results

In this section of this essay, I discuss the performance of the different information treatments for behavioural information-seeking, perceived persuasiveness and stated preferences outcomes. I present the results by looking at basic statistics and treatment effect estimates from unweighted and weighted estimators tailored for adaptively collected data. Based on these results, I discuss further these findings on whether I find systematic evidence to indicate whether the framing and the content of information about malfeasance matters. I then elaborate on the algorithm's performance in allocating experimental units across treatment arms and the level of convergence to the posterior probability of each arm being the performing best arm.

As I mentioned earlier, the central premise of this study is that the content and benchmarks used to disseminate information about malfeasance do indeed matter when it comes to informing citizens. The critical assumption is that more compelling and straightforward information strategies would trigger a sufficient interest within the target audience for them to learn more about corruption. I measured this information-gathering behaviour using CTRs for each

[11]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.

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

[13]You can find the pre-analysis plan on: https://www.socialscienceregistry.org/trials/7233

treatment. I complemented this analysis using items tailored to capture messaging strategies's PP.

Table 3.4 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, I find 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 benchmarking information, the *Spatial* framing shows to be more compelling than a *Temporal* frame, but this difference is small; of two percentage points. We also note that more promising treatment arms obtained smaller standard errors. This result may be because the best-performing arms received a larger share of experimental units. Table 3.15 summarises the number of respondents assigned to each treatment condition.

Table 3.4 – 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*.

In terms of PP outcomes, the results deviate from the behavioural outcome to some degree. Table 3.5 shows the proportion of people that answered 'yes' to each of the four binary items included in this battery of questions. I included a "composite Index", which is the mean of all four items [14]. This analysis suggests that once you pool treatment arms together by their metric, the *Individual* metric yields the highest rate of persuasiveness versus the other two metrics. This

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

result is somewhat puzzling, as this same condition attained the lowest CRT. We also notice that subjects appraised highly in terms of PP once you introduce temporal benchmarks, where there is a difference of 7 percentage points between the *Spatial* and *Temporal* frame for the Index outcome. As for the different metrics, the results are more consistent between the behavioural and the PP outcomes for the *Foregone* metric. This treatment factor is rated highly in PP and has one of the highest CTRs.

Table 3.5 – Perceived Persuasiveness by Treatment, Frame and Metric

Condition	CTR	Convincing	Trustworthy	Clear	Precise	PP Index
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 PP items. The table also reports these same indicators by *Frame* and by *Metric*.

The evidence is also mixed regarding the stated preference outcomes of sharing or posting the videos. Table 3.6 summarises the proportion of individuals who reported being willing to share and/or post the videos. Again, one finding that sticks out from this analysis is that about 92% of respondents assigned to the *Individual Temporal* treatment video claimed that they would post it on a social media platform, even though it has the lowest CTR. As we have seen previously, the *Foregone Spatial* yields amongst the highest proportions of respondents willing to share or post this video if it contained information about their local government malfeasance. This evidence buttresses previous findings that this treatment arm is amongst the optimal information strategies.

Table 3.6 – 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 summarises 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 results of the multivariate are in accordance with the findings obtained from the descriptive analysis above. Still, they demonstrate a more nuanced understanding of which factors are most relevant when informing citizens about corruption. Table 3.7 provides pooled estimates with and without covariates using the model reported in Equation 3.6. This estimation approach ignores the adaptability of the experimental design and leverages that *Exploration sampling* exponentially converges each treatment arm's posterior probability. The baseline category in this estimation is *Standard Spatial*, which is one of the best-performing arms in terms of CTRs, but under-performs on PP outcomes. This is why most point estimates for the first outcome are negative for the CTR outcome and positive for PP outcome.

From these results, we can discern that both *Individual* arms fails in being sufficiently compelling in triggering respondents to search for more information about corruption. One dimension that both *Individual* arms fall short of is its level of *Trustworthiness*. This evidence suggests that there may be specific attributes of this messaging strategy that undermine its credibility and put off respondents from learning more about corruption. Another possible explanation is the low levels of respondents' attentiveness assigned to this treatment condition. Table 3.14 summarises the results of three attention-check questions included in the study, illustrating a significant drop for the *Individual Spatial* arm on the number of people that correctly answered the question that asked about the framing used in the videos. Thus, there could be issues with treatment compliance. A third plausible explanation is that the indicators included in this

study may not capture all crucial dimensions that would help identify and differentiate which information strategies are more compelling to others.

Table 3.7 – Baseline Results - Clicked and PE Index Outcomes

	CTR		PP Index	
	(1)	(2)	(3)	(4)
(Intercept)	0.54^{***}	0.24	0.66^{***}	0.74^{***}
	(0.02)	(0.12)	(0.02)	(0.09)
Standard Temporal	-0.02	-0.04	0.10^{***}	0.10^{***}
	(0.05)	(0.05)	(0.03)	(0.03)
Individual Spatial	-0.13	-0.12	0.14^{***}	0.12^{**}
	(0.07)	(0.07)	(0.04)	(0.04)
Individual Temporal	-0.15^{*}	-0.17^{*}	0.23^{***}	0.23^{***}
	(0.07)	(0.07)	(0.04)	(0.04)
Foregone Spatial	0.00	0.00	0.14^{***}	0.14^{***}
	(0.04)	(0.04)	(0.02)	(0.02)
Foregone Temporal	-0.01	-0.02	0.16^{***}	0.16^{***}
	(0.04)	(0.04)	(0.03)	(0.03)
Covariates	No	Yes	No	Yes
R2	0.01	0.04	0.05	0.06
Adj. R^2	0.00	0.03	0.04	0.05
Num. obs.	1200	1179	1200	1179

Note: This table reports the pooled 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* are more consistent across all outcomes. In terms of CTR outcome, there is no statistically significant difference once compared to the *Standard Spatial* arm. However, this arm stands out on the PP Index outcome as it yields a sizeable and positive treatment effect[15].

As pointed out in section 3.4, to rule out possible sampling bias we can use Inverse Probability of Treatment Weighting (IPTW) to correct the sampling bias introduced by the adaptability of the data collection process. Table 3.8 reports the average treatment effects of both of these estimators. Point estimates and standard errors are shown to be similar to the pooled unweighted estimates in Table 3.7. Yet, for the *Individual Temporal*, the results are no longer statistically significant at the conventional 5% significance level.

[15]The average treatment effect for the Index outcome is around 0.57 standard deviations above the mean index of the baseline category.

Table 3.8 – IPTW Estimates - Clicked and PE Index Outcomes

	CTR		PP Index	
	(1)	(2)	(3)	(4)
(Intercept)	0.54***	0.10	0.66***	0.88***
	(0.02)	(0.10)	(0.02)	(0.05)
Standard Temporal	−0.04	−0.04	0.10***	0.10***
	(0.05)	(0.04)	(0.03)	(0.03)
Individual Spatial	−0.13	−0.13	0.12**	0.12**
	(0.07)	(0.07)	(0.04)	(0.04)
Individual Temporal	−0.15	−0.14	0.22***	0.22***
	(0.08)	(0.08)	(0.04)	(0.04)
Foregone Spatial	0.01	0.01	0.13***	0.13***
	(0.04)	(0.04)	(0.02)	(0.02)
Foregone Temporal	−0.01	−0.01	0.16***	0.16***
	(0.04)	(0.04)	(0.03)	(0.03)
Covariates	No	Yes	No	Yes
R2	0.01	0.04	0.05	0.06
Adj. R2	0.00	0.03	0.04	0.05
Num. obs.	1200	1200	1200	1200

Note: This table reports the pooled 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. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

I evaluated treatment effects by pooling together treatment arms within frames and metrics and we can see the results of this analysis set out in Table 3.9. This analysis demonstrates that there is no statistically difference on benchmarking information once we look at the behaviour outcome CRTs, but there is for PE 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 PP index, the difference is quite substantial, which is about 0.18 standard deviations of the baseline category. Once you pooled arms by metric, the results are somewhat inconsistent. The *Individual* metric under-performs in CTRs, compared to the *Standard* metric, but yet it obtains higher evaluations on the PP index. There is a more consistent relationship for the *Foregone* metric. This factor generates positive treatment effects for the CTR outcome, although it is not statistically significant. However, for the PP outcome, it yields large and statistical significant treatment effects.

Table 3.9 – Estimates by Frame and by Metric Unweighted - Clicked and PE Index Outcomes

	Frame				Metric			
	CTR		PP Index		CRT		PP Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	0.53***	0.23	0.73***	0.82***	0.53***	0.23	0.69***	0.76***
	(0.02)	(0.12)	(0.01)	(0.09)	(0.02)	(0.12)	(0.01)	(0.09)
Temporal	−0.02	−0.04	0.08***	0.08***				
	(0.03)	(0.03)	(0.02)	(0.02)				
Individual					−0.13**	−0.13**	0.15***	0.14***
					(0.05)	(0.05)	(0.03)	(0.03)
Foregone					0.00	0.01	0.12***	0.12***
					(0.03)	(0.03)	(0.02)	(0.02)
Covariates	No	Yes	No	Yes.	No	Yes	No	Yes
R2	0.00	0.03	0.01	0.03	0.01	0.04	0.04	0.05
Adj. R2	0.00	0.03	0.01	0.02	0.00	0.03	0.03	0.04
Num. obs.	1200	1179	1200	1179	1200	1179	1200	1179

Note: This table reports the pooled unweighted estimates 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 PP Index. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 3.10 reports the weighted estimates using IPTW for the *Frame* and Metric factors. This table shows that the results are smaller than the previous estimations. However, in terms of statistical significance, the results are consistent. For the *Frame* factor, point estimates for the PP outcomes drop from 0.08 to 0.05, but they are still statistically significant at a 99 per cent confidence level. We also see a drop in all estimates for the *Metric* factor. There is a considerable drop in the PP Index for the *Individual* level. Point estimates drop from 0.15 to 0.08 for the baseline model with covariates and from 0.14 to 0.08, including covariates. But these results are still statistically significant.

Table 3.10 – IPTW estimates by Frame and by Metric - Clicked and PE Index Outcomes

	Frame				Metric			
	CTR		PP Index		CRT		PP Index	
	(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)
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 PP Index. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Inference One of the benefits of using *Exploration sampling* is that sampling averages of each treatment arm should be consistent and asymptotically unbiased (Caria et al., 2020). This means we can conduct inference by ignoring the adaptability of the design. However, one approach to rule out any possible sampling bias is by performing randomisation inference, as this technique does not relies on the assumption previously mentioned. Table 3.11 reports the F-test and p-values for each treatment, frame and metric. Based on these results, we fail to reject the sharp null of no differences across treatments and frame, thus finding no evidence of a unique best arm for the CTR outcome. The results are the different once we look at *frame* as we find that there is a statistically significant difference for the CTR outcome.

Table 3.11 – Randomisation Inference and Anova F-test

	CTR		PP Index	
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 randomisation 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 randomisation 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 'PP' index variable and the predictor is the treatment variable.

I extended this inferential analysis using a Batch OLS hypothesis testing procedure. This inference method is particularly suitable for batch adaptive experiments as this one [16]. This procedure incorporates randomisation 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 for the overall effect in a meta-analysis [17]. Here, I imposed the sharp null that there are no differences with reference categories.

Table 3.12 reports the p-values for both outcomes by treatment. The results are consistent with the estimated p-values from the conventional hypothesis testing approach. For the CTR outcome, none of the treatment arms is statistically significant from the reference category. We find that the Foregone arm and the *Standard Temporal* are significant for the PP Index.

[16] 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.

[17] 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 3.12 – Batched OLS Inference - Treatments

Treatment	P-values	
	CTR	PP Index
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**

Looking at the frames and metrics factors, Table 3.13 summarises the results from the BOLS hypothesis testing. Again, both factors are statistically significant for the PP index but not for the CTR outcome. The *Foregone* has shown to be a consistent metric throughout.

Table 3.13 – Batched OLS Inference - Frame and Metrics

Factor	P-values	
	CTR	PP Index
Frame		
Temporal	0.968	0.003*
Metric		
Individual	0.205	0.042*
Foregone	0.915	0.000**

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 3.1, 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 [18]. 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%.

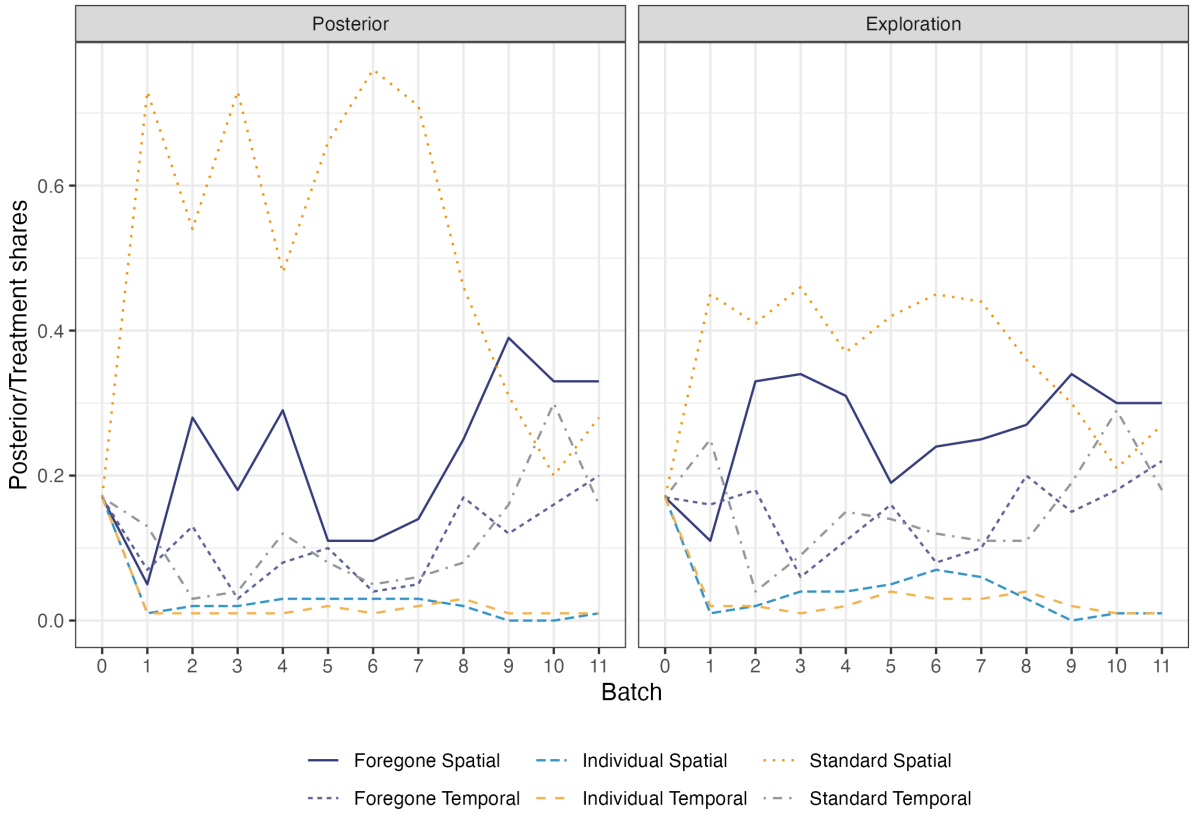
I also find that the algorithm continuously allocates units to second-best arms such as *Standard Temporal* and *Foregone Temporal*, as both treatments gradually converge to their posterior

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

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 3.17 in the Appendix summarises 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 3.3 complements this analysis by depicting the posterior distribution of each arm by batch and by treatment.

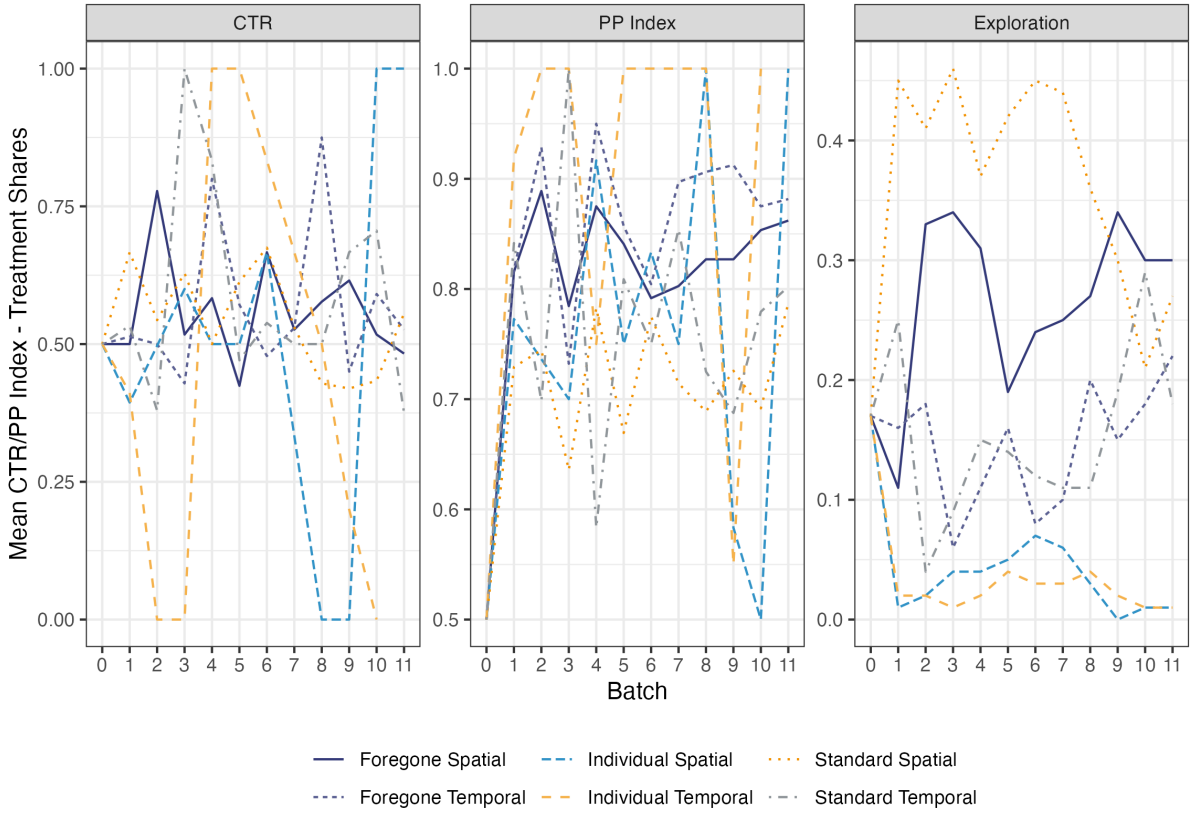
Figure 3.1 – Posterior Probability and Treatment Shares from Exploration Sampling



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%.

One interesting finding that we can find in Figure 3.2 is the low variance of the *Foregone Spatial* and *Foregone Temporal* arms for both outcomes. For both treatment conditions, for the CTR outcome are move between 0.45 to 0.8. Whereas for the PP Index, between 0.75 and 1. Conversely, we observe that treatment conditions such as *Individual Temporal* and *Individual Spatial* have significant variance on its CTR and PP index.

Figure 3.2 – CTR, PP Index and Assignment Shares from Exploration Sampling



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

3.6 Discussion

In this essay, I sought to enhance our understanding of what people deem compelling and informative when they learn about corruption in their local constituencies. This research aimed to contribute to the recent number of studies examining the relationship between informing citizens about elected officials' wrongdoing and voting behaviour. In particular, it focuses on the information architecture of the signal. In this study, I tried to tease out what pieces of information or content are more likely to bring the average citizen's interest. I tested several factors, such as framing and content, used in online and field experiments examining the relationship between information about malfeasance and voting behaviour.

For this, I conducted an adaptive experiment, a powerful tool to identify the most promising treatment arms. In this version of the adaptive experiment, I sequentially changed treatment assignment probabilities, seeking to identify this arm with high certainty and exploring the second-best competing arms.

The results from this experiment identifies two very different information strategies to be somewhat equally compelling. On one end, I find that *Standard Spatial*, which is the most

straightforward and less sophisticated messaging strategy, performs well for the behaviour outcome (CTR) but poorly for the attitudinal outcomes (PP). One 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.

The evidence shows that *Foregone Spatial* is an optimal information strategy. This treatment systematically yields higher CTRs and a high PP Index score. This treatment gradually obtains the highest posterior probability of being the best arm based, with a considerably smaller variance. However, there is no sufficient evidence to claim this is the unique best arm. This finding also buttresses the extensive literature that argues that conveying information that emphasises undesirable outcomes tends to be stronger. This treatment gave much richer content in terms of the complexity of the information provided 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 PP index but poorly for the CTR behavioural outcome. As raised before, these results can be explained due to lower levels of attentiveness found mainly for the *Individual Spatial* treatment. This issue compounded with the fact that this arm received fewer observations[19], which would explain the inconsistent results I 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% respondents to the best arm, even though this arm in the first batch received only 24 observations. At the same time, underperforming arms received near-zero treatment assignment probabilities. Such treatment assignment shares may not be appealing to researchers. An optimal strategy would be to favour 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 combining behavioural and attitudinal measures of persuasiveness rather than having a single metric. Although using binary outcomes simplifies the estimation of the best-performing arms' posterior probabilities, there could 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 implementing this experiment in the field with the information of their local constituency and using a composite index. In this design, we cannot make claims about whether these information

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

treatments shape individuals' political preferences or change beliefs. Still, it tells us that some information features are more compelling than others.

3.7 Appendix

Descriptives

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

Table 3.14 – 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 summarises 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 summarises the portion of respondents that answered all three questions correctly.

Table 3.15 – 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 3.16 – 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 sizable differences for the *Foregone Temporal* and *Individual Spatial*.

Simulations

I conducted several simulations that informed this study's design and sampling algorithm selection. I 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, I only reported the simulations with 6 conditions. Figure 3.4 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 I collected 1,276 evaluations of the four treatment arms of interest. Using this data, I computed the posterior probability of being the best arm for each arm. The results of this analysis are reported in Table 3.18.

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 3.19 reports the simulation parameters for these three scenarios.

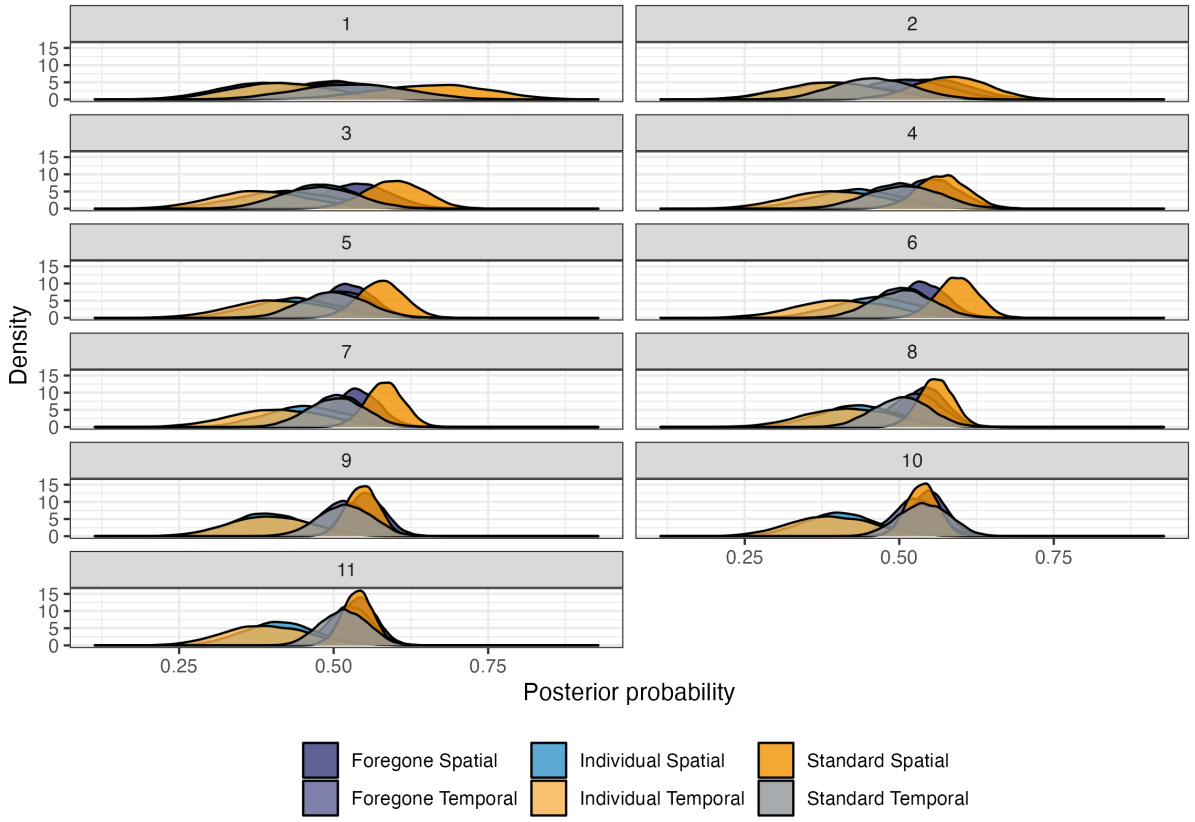
Table 3.18 – Posterior Probability Best-performing Arm - Results from the Pilot

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

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 3.17 – Posterior probability, standard deviation posterior probability and exploration sampling share by treatment arm

Figure 3.3 – Posterior Probability by Batch



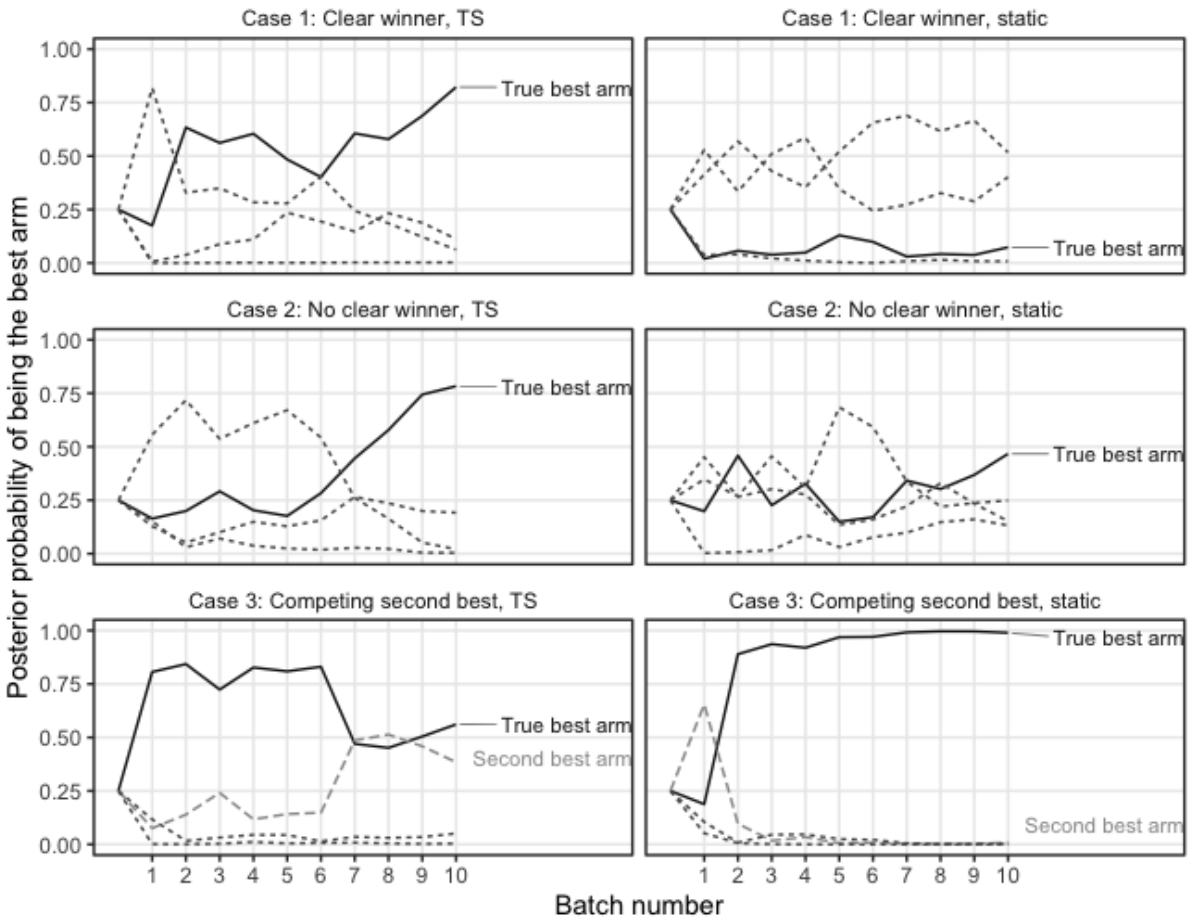
Note: This figure depicts the posterior density of each arm being the best from batch 1 to batch 11. It is important to highlight that the posterior probability of being the best is not the mean of posterior probability density, but is the calculated based on many times each arm attains the highest posterior 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.

Table 3.19 – 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

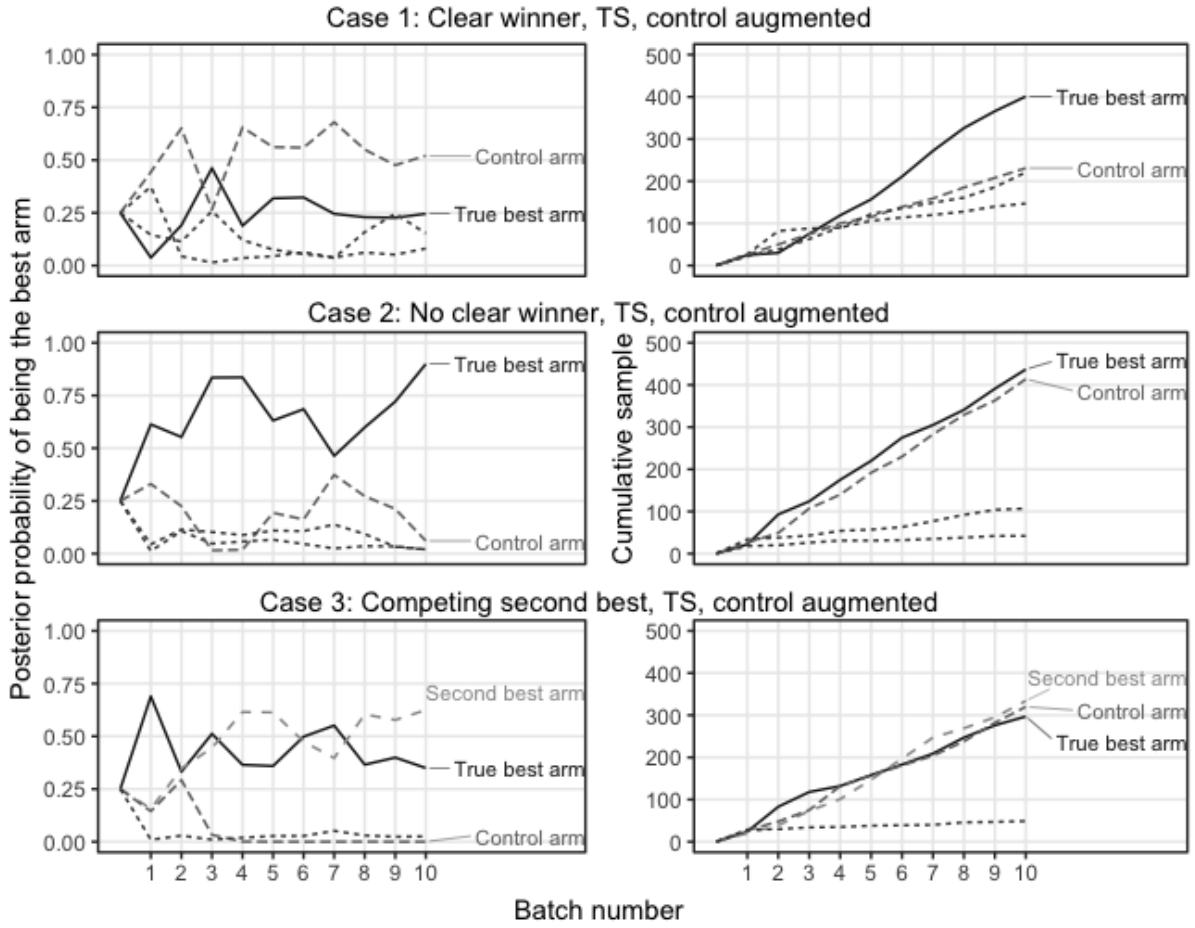
I compared the results of these scenarios using a Thompson sampling algorithm versus a static design where treatment assignment is equal to $1/K$, and K is the number of arms. As shown in Figure 3.4, the Thompson sampling algorithm successfully identifies the best-treatment arm in the *Clear winner* 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 3.4 – 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 3.5 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 3.5 – 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 3.6 – Root mean Square Error by Number of Batches - 6 arms/1200 observations

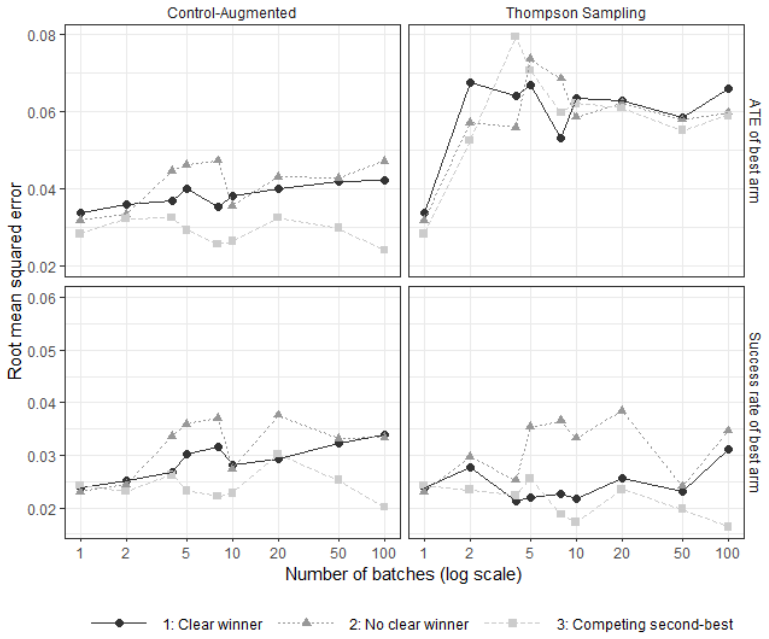


Figure 3.7 – Arm Coverage by Number of Batches - 6 arms/1200 observations

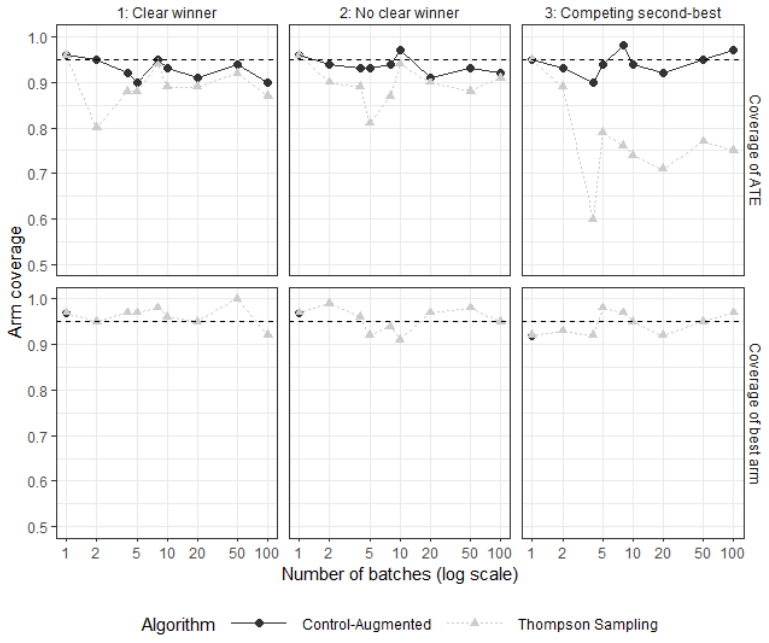


Figure 3.8 – Root Mean Square Error by First Batch Size - 6 arms/1200 observations

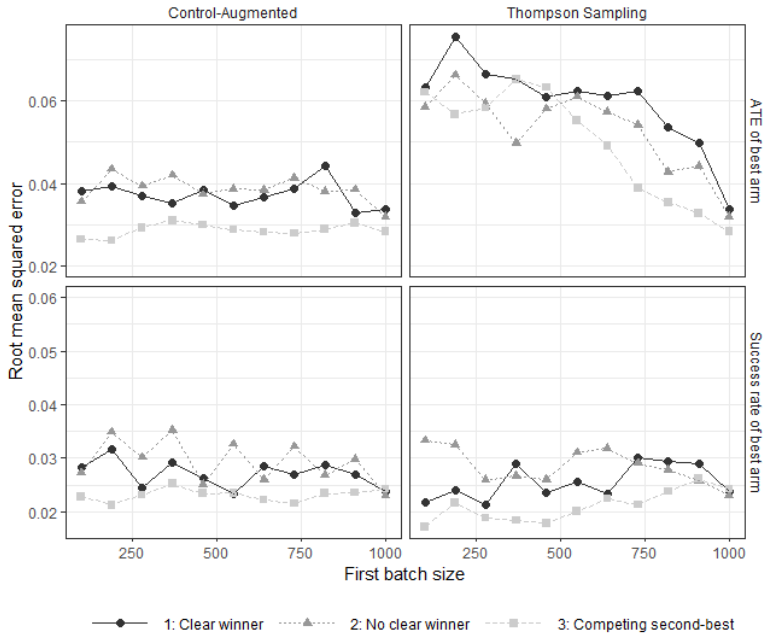
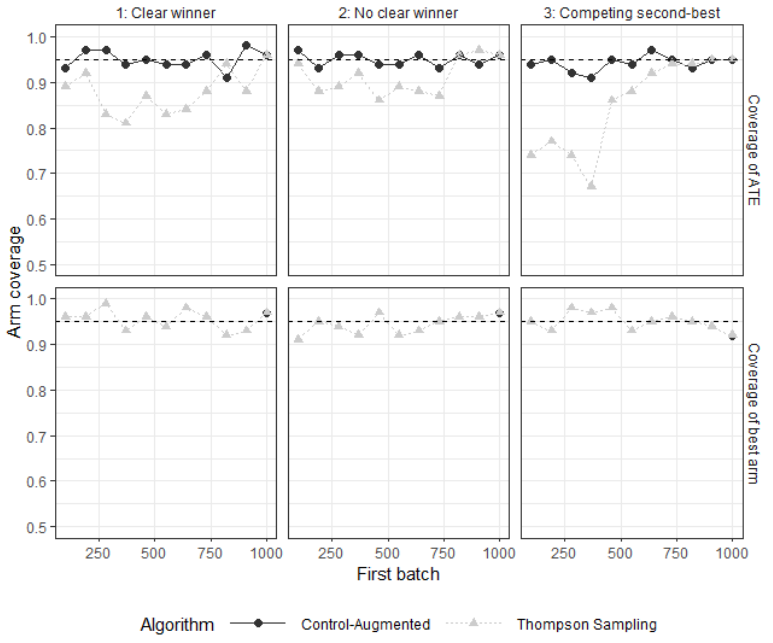


Figure 3.9 – Arm Coverage by First Batch Size - 6 arms/1200 observations



Additional simulations I 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, I 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. 3.21 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, I report in Table 3.22 the mean of each using two weighting techniques: IPTW and Hájek-type stabilising weights.

Table 3.20 – 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

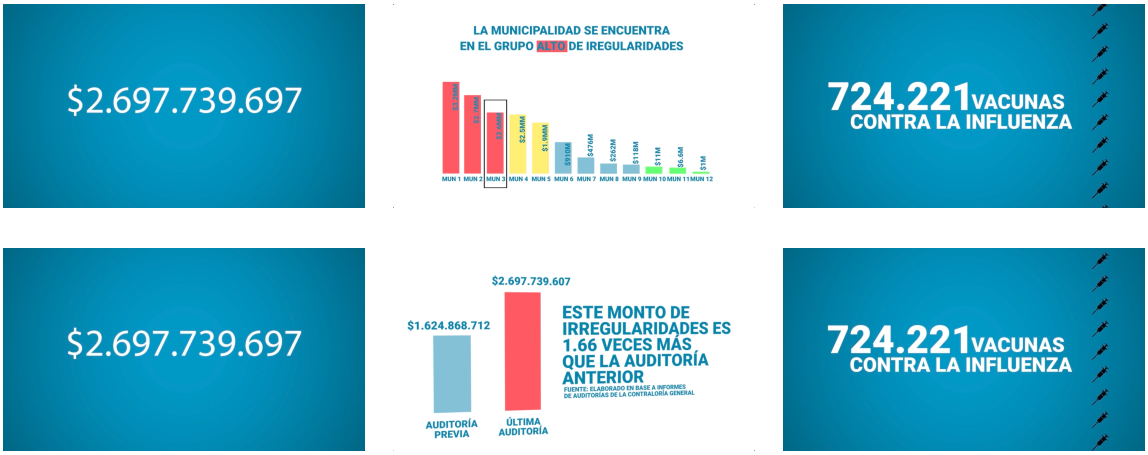
Table 3.21 – Results Exploration Sampling Kasy et al's Exploration Sampling

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 3.22 – Unweighted and IPW means Thompson Sampling

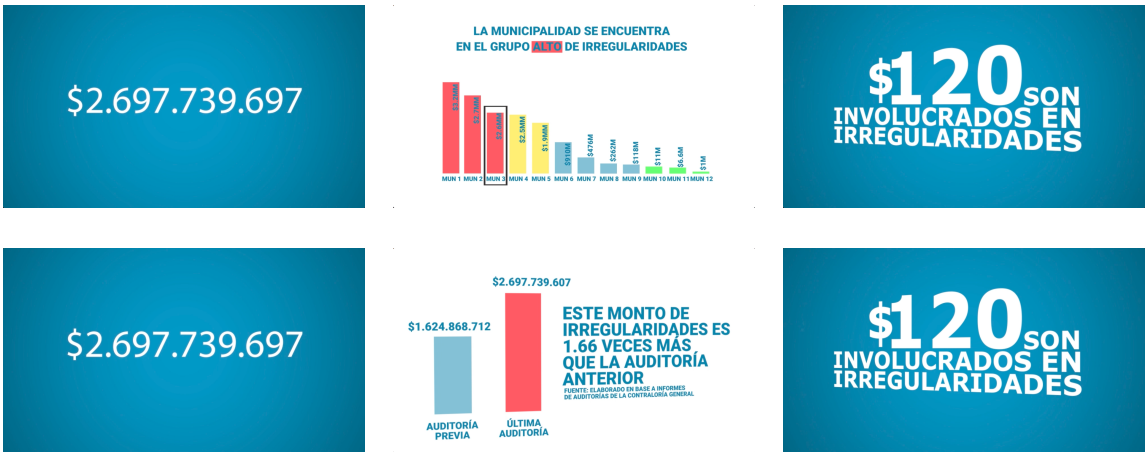
Treatments	Unweighted	IPW	Hajek
Foregone Temporal	**0.208**	0.204	**0.208**
Foregone Spatial	0.160	0.210	0.160
Individual Temporal	0.157	0.101	0.193
Individual Spatial	0.193	0.200	0.157
Standard Temporal	0.145	**0.260**	0.145
Standard Spatial	0.061	0.070	0.061

Figure 3.10 – Screen shoots - Foregone Treatment



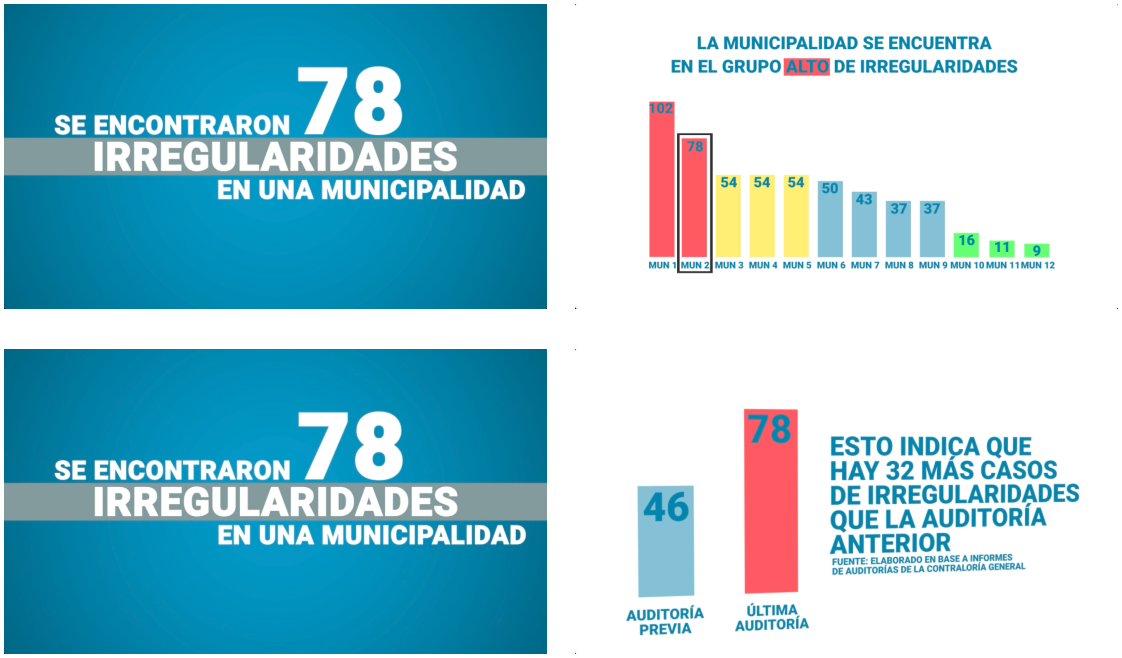
Note: This figure shows screenshots of the video information treatments that subjects received in this experiment. The top 3 images show screenshots of the *Foregone Spatial* treatment. The second row of screenshots corresponds to the *Foregone Temporal* condition. The first frame for both treatments 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 3.11 – 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 3.12 – 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.

Results pilot

In this section, I present the results of a pilot conducted in March 2021 and included 666 subjects. Each respondent evaluated six videos out of twenty-four information treatments. The twenty-four treatments stem from a 2 x 2 x 6 factorial design. Table 3.23 describes the factors and their levels. This pilot differs from the design proposed in this essay in the following features: 1) It had one more factor, which was the levels of malfeasance reported: High or Low. 2) It included four additional metrics.I collected 3,996 observations in this pilot, equivalent to around 166 evaluations per video. Table 3.26 provides details of the sample per arm. Below, I describe the sampling and analyses conducted with the data obtained from this pilot. Respondents were from the same panel of the main study and consisted of approximately 57,708 respondents. From this pool, I randomly sampled 666 eligible respondents.

	Factor: Benchmarking
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

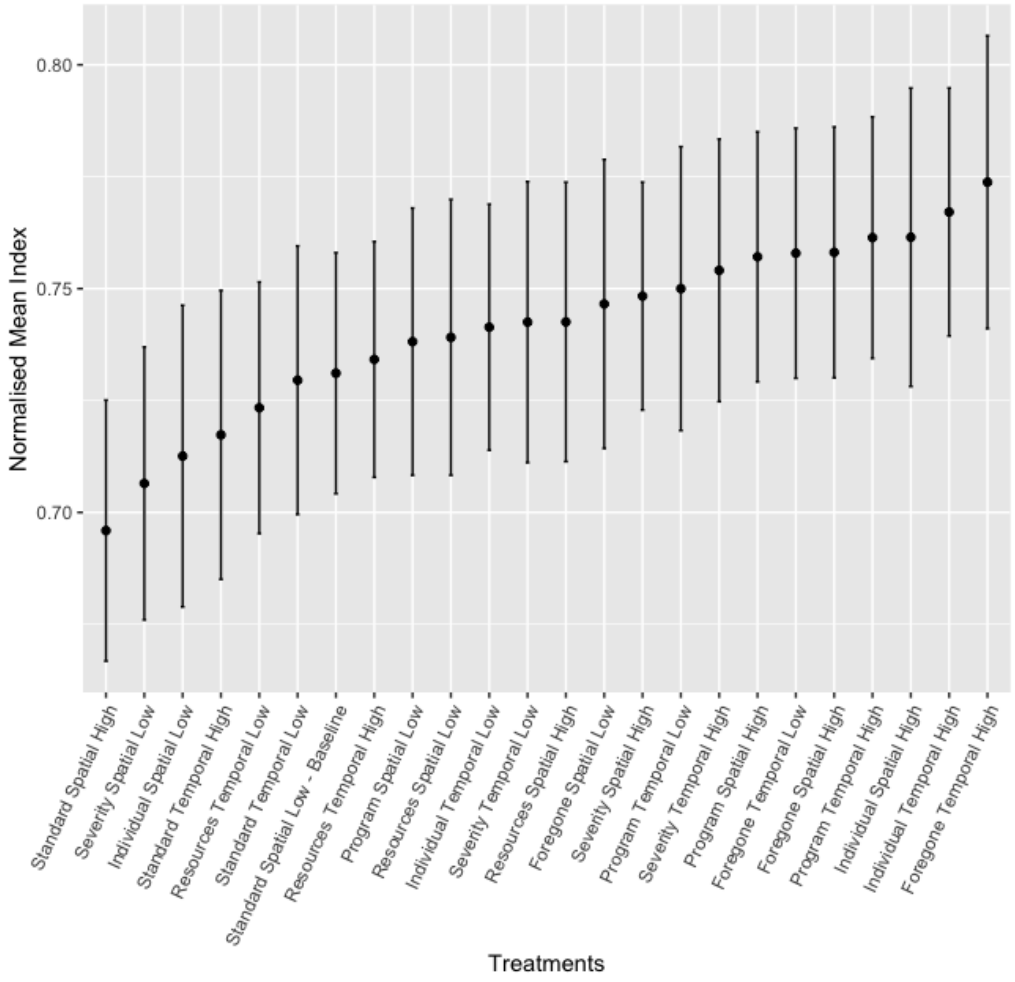
Table 3.23 – Description of treatment factors

Outcome measures. Subjects viewed and evaluated six randomly assigned videos. As described earlier, I used four common messaging evaluation questions that subjects were asked after each video treatment: credibility, convincing, reliability and accuracy. Rather than asking respondents whether they agreed or disagreed with the evaluation statements, I used a 5-point Likert scale from *Strongly Agree* to *Strongly Disagree*. The second set of questions tried to elicit

respondents' beliefs about the levels of malfeasance in the hypothetical municipality.

Figure 3.13 plots the treatment effects for all 24 video treatments. The outcome is a normalised measure of all evaluations items. The Foregone-Temporal and Individual Temporal treatments are the top-ranked conditions. Within the framing factor, the temporal frame slightly outperforms the spatial frame. The multivariate analyses reported in Tables 3.24 and 3.25 confirm the same patterns shown in Figure 3.13, even controlling for relevant covariates and order effects.

Figure 3.13 – Reliability Index (normalised) by Treatment Arms



These results provide evidence that information frames can be crucial in persuading individuals when they learn about corruption. While these estimations are underpowered, Programme, Individual and Foregone are the possible candidates to be the most-promising arms. Almost all three have larger treatment effects on the temporal and spatial versions. We can observe that the Foregone level is the most promising candidate because of the magnitude of its positive coefficient – but again, these coefficients are not very precisely estimated.

The *Benchmark* model includes a dummy variable for the temporal treatments. This model confirms that a temporal bench-marking is a slightly more persuasive messaging strategy than

spatial bench-marking. However, in this case, the difference in means between these two levels is statistical indistinguishable.

In model *Metrics*, I regressed the normalised evaluation index on six dummy variables representing the metric treatments – Program, Individual, Resources, Severity, and Standard. The reference category here is the Standard metric. Again, there is evidence that the Foregone is the most highly evaluated messaging treatment. All the other messaging treatments have positive coefficients but are smaller.

Table 3.25 reports estimates 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 3.24 – Regression Results - Information Treatments

	Dependent variable: Normalised 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. $+p < .10$, $*p < .05$, $**p < .01$, $***p < 0.001$.

It is reasonable to think that the average citizen prefers reducing malfeasance in govern-

Table 3.25 – Regression results to Information Treatments - Including Controls and Order FE

	Dependent variable: Normalised 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	3336
R2	0.022	0.018	0.022	0.031	0.026	0.030

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

ment. The challenge is informing citizens of the prevalence and magnitude of corruption in government. The challenge is not particularly surprising. We know from behavioural economics and psychology that individuals can find it quite challenging to digest information necessary to make welfare-enhancing decisions. This is salient in the case concerning choices in the financial realm – those related to retirement, savings, investment, and insurance (Duch, Granados, Laroze, Lopez, Ormeño and Quintanilla, 2020). Recent evidence suggests that choice architecture, framing information in a digestible fashion for the average individual, can promote welfare-enhancing decision-making. This research design presents an essential initial step in designing optimal malfeasance messaging.

It essentially seeks to crowd-source many evaluations of 24 potential malfeasance messaging videos. Random assignment to video treatments, followed by a set of evaluation questions administered via ChatBot, provided insights into how to frame the malfeasance information generated by audit reports. Individuals were more persuaded by a temporal as opposed to a spatial comparison. The crowd-sourced evaluations also identified what I think is an optimal narrative about the cost of malfeasance. Respondents rated most highly the video treatment that presented the cost of malfeasance in terms of foregone losses on a crucial public good – in this case, I used foregone expenditures on influenza vaccines.

Finally, in Figure 3.14 I reported the posterior probability of each arm being the best-performing treatment arm. I conducted this analysis by making several modifications to the pilot. 1) I included only the four arms of interest 2) I split the data into equal-size batches of 100 observations. 3) I dichotomise the "convincing" outcome to simplify the posterior estimation of each arm. 4) I 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 3.14 – Posterior Probability of Best-Performing Treatment Arm

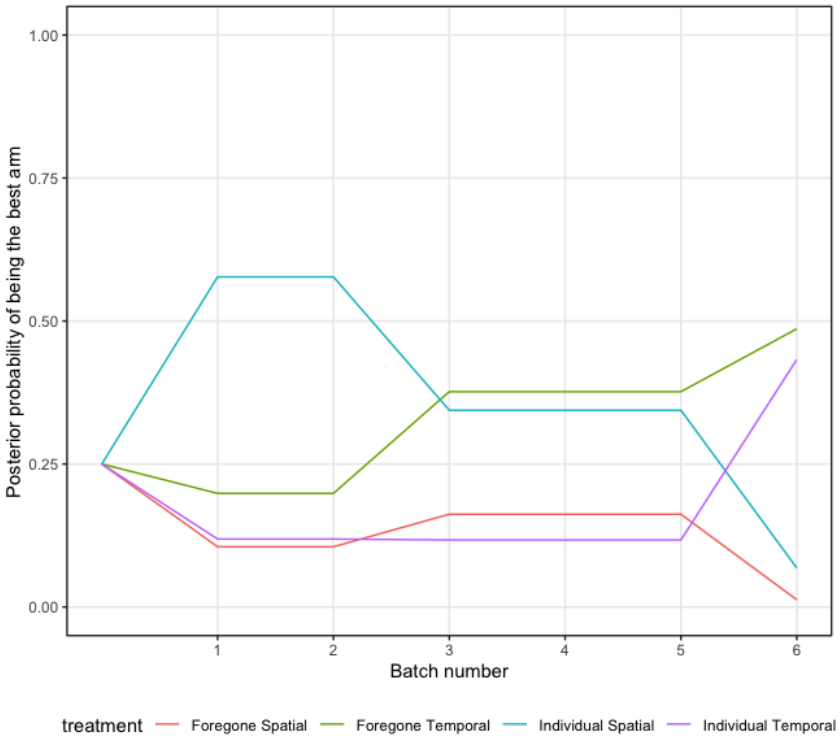


Table 3.26 – Sample per Treatment Arm - Pilot

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 3.27 outlines the steps of this algorithm.

Table 3.27 – 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[20].

[20]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 formalise 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:

$$\hat{\Delta}_t^{BOLS} = \frac{\sum_n^i (1 - A_{t,i}) R_{t,i}}{\sum_n^i 1 - A_{t,i}} - \frac{\sum_n^i A_{t,i} R_{t,i}}{\sum_n^i A_{t,i}} \tag{3.7}$$

In equation 3.7, $A_{t,i} \in \{0, 1\}$ 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.

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{\frac{\sum_n^i (1 - A_{t,i}) \sum_n^i A_{t,i}}{n\sigma^2}} (\hat{\Delta}_t^{BOLS} - c) \tag{3.8}$$

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:

$$\hat{\Delta}^{AW-AIPW} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}} Y_{t,1}}{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}}} - \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}} Y_{t,0}}{\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}}} \tag{3.9}$$

In equation 3.9, $\sqrt{\pi^{(n)}}$ 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:

$$\hat{V}_1 := \frac{\sum_{t=1}^{T} \sum_{i=1}^{(n)} \pi^{(n)} (Y_{t,1} - \hat{\beta}_1^{AW-AIPW})^2}{\left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{\pi_t^{(n)}} \right)^2} \text{ and } \hat{V}_0 := \frac{\sum_{t=1}^{T} \sum_{i=1}^{(n)} \pi^{(n)} (Y_{t,0} - \hat{\beta}_0^{AW-AIPW})^2}{\left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sqrt{1 - \pi_t^{(n)}} \right)^2} \tag{3.10}$$

$$\hat{C}_{0,1} := \frac{\sum_{t=1}^{T}\sum_{i=1}^{(n)}\sqrt{\pi^{(n)}(1-\pi^{(n)})}(Y_{t,1}-\hat{\beta}_1^{AW-AIPW})(Y_{t,0}-\hat{\beta}_0^{AW-AIPW})}{\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{\pi_t^{(n)}}\right)\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sqrt{1-\pi_t^{(n)}}\right)} \tag{3.11}$$

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

$$\frac{\hat{\Delta}^{AW-AIPW}}{\left(\hat{V}_1+\hat{V}_0\right)} \xrightarrow[T\longrightarrow\infty]{d} \mathcal{N}(0,1) \tag{3.12}$$

Figure 3.15 – Screenshot survey

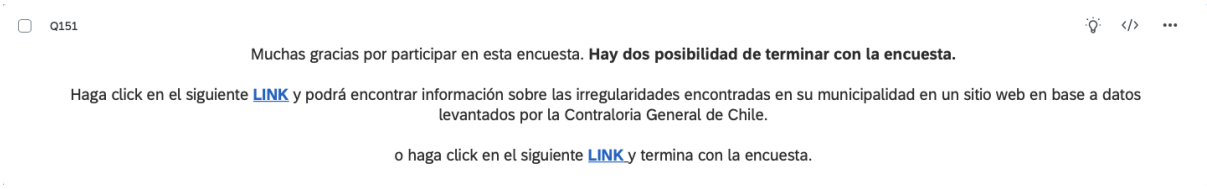
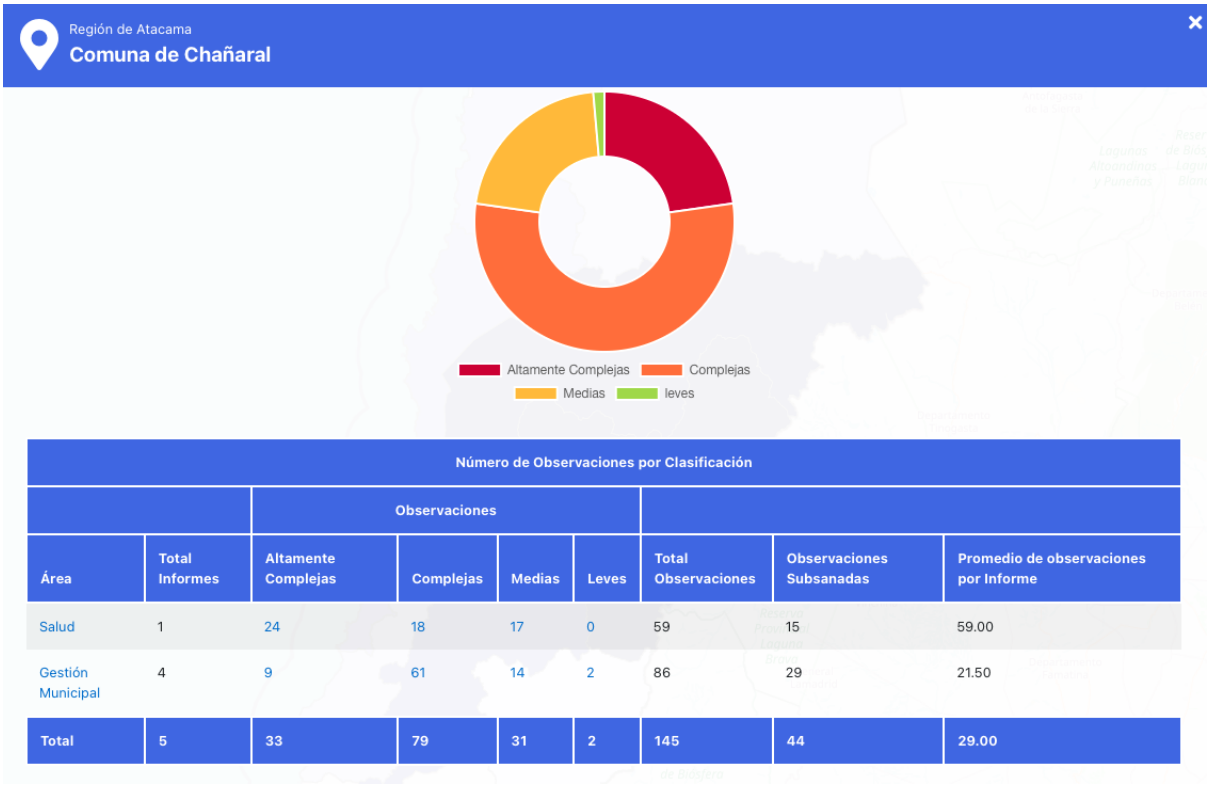


Figure 3.16 – Screenshot survey

Figure 3.17 – 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

Research Questions and hypotheses

Research Questions - The first research question investigates how using different benchmarks makes the information more or less compelling. As pointed out before, using temporal frames tends to have a more substantial effect when voters assess their elected officials' performance (Bhandari, Larreguy and Marshall, 2019). At the same time, there is evidence that spatial benchmarks are particularly good reference points when voters want to evaluate elected officials' economic policies (Kayser and Peress, 2012; Arel-Bundock, Blais and Dassonneville, 2021). Thus, the first research question aims to provide additional evidence in identifying which of these two benchmarks resonates strongly with voters.

- *Research Question 1:* Which benchmark, either spatial or temporal frame, resonates more with the average citizen when evaluating their local government levels of malfeasance?

The second area of interest is the content that citizens receive when they are informed about malfeasance. It is reasonable to expect voters will respond differently to different information contents. However, it is also sensible to expect that treatments that convey losses in a digestible way would be more persuasive than just general descriptive information. Thus, the second question aims to answer what corruption metrics are more compelling to the average voter.

- *Research Question 2:* Which information content about malfeasance resonates more with voters?

As alluded to before, using conventional experimental designs to test multiple treatments can be problematic and wasteful. Thus, this last question aims to answer whether adaptive experiments can be a more efficient approach to identifying the best-performing treatment arm(s).

- *Research Question 3:* Are adaptive experiments more efficient than conventional static designs? What are the efficiency gains from conducting this type of experiment?

Hypotheses Based on the previous theoretical discussion reported in Section 3.1, I expect messages that 1) provide bench-marking information and 2) calibrate malfeasance in metrics an accessible to the average citizen will be more persuasive to the average citizen. Table 3.2 identifies the specific treatments I intend to implement. My hypotheses regarding the specific treatments are:

As mentioned, the evidence leans towards temporal comparisons to be more persuasive and digestible to individuals. This is formalised in *Hypothesis 1.*

- *Hypothesis 1*: Temporal benchmarking frames will generate higher evaluations than spatial frames;

Similarly, extensive theoretical and empirical evidence shows that information expressed as foregone losses tends to be more persuasive. Thus, we should expect a larger treatment effect for this treatment factor. This is formalised in the following hypothesis:

- *Hypothesis 2*: Information about malfeasance reported as foregone expenditure will be more compelling than other corruption metrics.
- *Hypothesis 3*: From the 6 metrics shown in 3.2 the most effective will be the programme treatment.

Chapter 4

Paper 3: Measuring Corruption Using RIRT

4.1 Introduction

A critical assumption of many studies that use survey-based measures is that respondents will answer truthfully. Nevertheless, this can be particularly challenging in the case of survey questions concerning illicit or socially unacceptable behaviours. For example, when scholars seek to obtain measures of attitudes towards race, religion, or extreme political ideologies. In these cases, respondents are unlikely to provide truthful answers to sensitive questions due to social desirability, fear of retribution or even legal consequences (Blair, Imai and Lyall, 2014; Yusaku Horiuchi Horiuchi, Markovich and Yamamoto, 2021; Singh and Tir, 2022; Smallpage et al., 2022).

This is also the case for the case of corruption, where policymakers and scholars have heavily relied on surveys to measure levels of corruption at any given time. (Fisher, 1993; Jo, 2000). There are a number of surveys globally that are important to reflect on when measuring levels of corruption. Fishman and Golden (2017) both outline several surveys, such as the World Bank Business Environment Survey and the Latin America Public Opinion Project, as some of the international attempts to measure corruption. In this case,Wysmułek (2019) finds a remarkable increase in the number of items asking about corruption, from an average of 5 up to 10 items per survey. All of the above demonstrates the an increasing interest in obtaining estimates of corruption globally.

Obtaining unbiased estimates is vital when measuring levels of corruption in any field of study. Ensuring a study has accurate estimates of subjective or experienced-based corruption can play two critical roles whilst investigating this topic. First, when the survey-based measure is an outcome. To illustrate this, some social scientists may be interested in changes over time or differences across groups in the prevalence of corrupt behaviour. Secondly, at the same time, some researchers may want to control for this confounder when the survey measure is a relevant covariate. Consequently, yielding precise estimations would minimise measurement error either in the outcome or the predictors.

Introducing indirect questioning techniques may help address social desirability bias (Fisher, 1993; Jo, 2000). Such methods of questioning have been used to measure the prevalence of corruption and vote-buying (Corbacho et al., 2016; Oliveros and Gingerich, 2019; Gingerich et al., 2016; Gonzalez-Ocantos et al., 2012; Gonzalez-Ocantos, de Jonge and Nickerson, 2014), to measure support for combatants (Blair, Imai and Lyall, 2014). Some of the popular methods are the use of lists, endorsements, as well as randomised response experiments. These techniques conceal an individuals' response by aggregating answers or introducing random noise [1].

Nevertheless, one main limitation is that the conventional form of these methods only provides population or group-level estimates of the sensitive behaviour(s). These techniques do not allow for sub-group or individual-level estimates that researchers may be interested in capturing. This means that researchers cannot identify whether, for example, subject i has or has not engaged in the behaviour in question. Thus, researchers must out-weight the benefits of implementing these survey techniques against these costs [2].

Recent advances in the literature on Randomise Response (RR) incorporate insights from Item Response Theory (IRT) that can help us sidestep some of these limitations raised before. In the following manuscript, I present the results of a Randomised Item Response Technique (RIRT), where I provide estimates of both population and individual levels of corrupt behaviours in Chile. The main goals of this study are four-fold. Firstly, to estimate the prevalence of corrupt behaviour of citizens and whether these results are robust. Secondly, provide bias-corrected estimates due to inattentive respondents. Thirdly, to obtain individual-level estimates of the latent trait of corrupt behaviour. Finally, I compare population estimates to estimates obtained from a direct questioning format from an opt-in and a representative sample.

This manuscript aims to make several contributions to the strands of research that examine survey methods and aim to measure corruption. The first main contribution is in the field of survey methods for sensitive topics. It provides empirical evidence on an extension of the randomised response technique that incorporates Item Response Theory. It also contributes to the research that seeks to measure corruption, where I provide population-level and individual estimates of a vast number of corrupt behaviour. To the best of our knowledge, this is the first essay that provides subject-level estimates of corrupt behaviours using an indirect questioning technique.

A third empirical contribution is to the literature investigating survey inattentiveness. I conducted several novel strategies to identify respondents that do not comply or pay attention to the survey instructions. I provided bias-corrected estimates due to inattentiveness at the individual and group-level estimates.

Finally, a secondary contribution is that I conducted a comparison of a particular set of corrupt behaviours by comparing the results obtained from RIRT to a direct questioning format both from a opt-in and representative sample. I examined estimates differences both at the individual and population level.

[1] There is an extensive literature on this topics beyond the field of political science

[2] They are also less efficient, as the sample size required to implement these techniques must be larger (even twice). Furthermore, list experiments can provide partial or complete protection, which depends on the existence of floor and ceiling effects (Ahlquist, 2018)

The essay continues as follows. Section 4.2 introduces the Randomised Response and the Randomise Item Response Technique. Section 4.3 provides details about the sampling and survey waves, the instrument used, and the implementation of this study. Section 4.4 discussed the estimation approaches for RIRT designs. Then, Section 4.5 reports and elaborates on the results obtained from this technique. It elaborates on population and individual estimates. Then, it shows the results from the different strategies to flag inattentiveness. Finally, section 4.8 summarises the findings of this study and its limitations.

4.2 Randomised Responses

Randomised Responses The RR method was developed by Warner (1965) and its most basic level introduces random noise that would help delink study participants' responses to a sensitive question. In practice, RR designs introduce noise using an instrument such as a die or a coin. Next the respondents are asked to throw a die or flip a coin, and in its basic design. The outcome from this device determines which question the respondent has to answer or which forced answer the respondent has to give. However, the researcher knows the probability distribution that describes the likelihood of all possible outcomes induced by the instrument. This known probability distribution then allows estimating the ratio of affirmative answers. It is important to stress that the RR design provides complete confidentiality, as neither interviewer nor researchers know the outcome of the device. This means that there is no deception in this setup [3].

A concrete example of this is the experiment conducted by Gingerich (2010) where he used the "Mirrored" RR to measure the levels of corruption in the public sector in Brazil, Bolivia, and Chile. In this design, the randomisation device used was a spinner. In turn, respondents were asked to whirl the spinner without revealing the outcome to the interviewer. The translated version of the study instructions was the following:

"For each of the following questions, please spin the arrow until it has made at least one full rotation. If the arrow lands on region A for a particular question, respond true or false in the space indicated only with respect to statement A. If the arrow lands on region B for that question, respond true or false in the space indicated only with respect to statement B. Do not make any marks to indicate which region the arrow fell for each question. Please remember that if you respond falsely to a statement in its negative form that means that the positive form of the statement is true."

If the spinner landed in the A region, respondents were asked to respond "yes" or "no" to the following question:

"A: I have never used, not even once, the resources of my institution for the benefit of a political party."

[3]It is important also to mention that there is no deception in the RR designs. However, researchers can use some information provided by the respondent, such as their month of birth, to produce an unrelated question; for example, the unrelated question could be: "I was born in the three months of the year".

If the spinner landed in the B region, respondents were asked to respond "yes" or "no" to the following question:

"B: I have used, at least once, the resources of my institution for the benefit of a political party."

In this setup, π corresponds to the unobserved probability of giving a positive response to the sensitive question. $Y = 1$ is the answer given by the respondent, and p is the known probability of selecting a particular question. Then, we can compute the probability of answering affirmatively as follows:

$$P(Y = 1) = \lambda = p\pi + (1 - p)(1 - \pi) \tag{4.1}$$

By rearranging equation 4.1, we can obtain the prevalence of the sensitive behaviour, where $\hat{\lambda}$ is the observed proportion of respondents selecting option A.

$$\hat{\pi} = \frac{\hat{\lambda} + p - 1}{2p - 1}, \text{ where p } \neq 0.5 \tag{4.2}$$

We can also obtain the variance of the estimated value of the prevalence behaviour, which is a function of the variance that comes from sampling and the introduction of random noise.

$$Var(\hat{\pi}) = \frac{\hat{\lambda}(1 - \hat{\lambda})}{n(2p - 1)^2} = \frac{\hat{\pi}(1 - \hat{\pi})}{n} + \frac{p(1 - p)}{n(2p - 1)^2} \tag{4.3}$$

The evidence of the effectiveness of RRs in minimising social desirability bias is overall positive. Lensvelt-Mulders et al. (2005) concluded from their meta-analysis that this technique was effective in yielding robust prevalence estimates of sensitive topics once compared to the direct questioning format[4]. Similarly, Höglinger, Jann and Diekmann (2016) conducted a study comparing four different RR variants to direct questioning by looking at student misconduct in Universities in the Netherlands. The authors found evidence that all RR designs outperformed the direct questioning version in at least three of five sensitive questions included in the study. As a result, the RR design has gain reputation to be a reliable technique to yield unbiased estimates of the sensitive in question.

Indeed, the evidence of empirical validation of RRs against official records is small yet promising. Rosenfeld, Imai and Shapiro (2016) found that RR performed better than other techniques[5] in estimating the proportion of people voting *'No'* in the Mississippi anti-abortion referendum in 2011. The authors also noted that RR produced robust estimates with significantly lower variance than the other indirect questioning techniques[6].

[4]Their meta-analysis comprised two meta-analyses: one meta-analysis included only six individual validation studies, whereas the second meta-analysis contained 32 comparative validation studies.

[5]In their study, they compared the estimates from the RR design to a direct questioning format and a list and endorsement experiments.

[6]The authors validated their results at the individual level using official records, which is state-of-the-art in terms of validation approach.

Amongst all the RR variants the crosswise design stands out due to its statistically efficiency and higher procedural adherence. Blair, Imai and Zhou 2015 provides an extensive analysis of different RR designs and their performance by looking at their main statistical properties which finding strong support for the crosswise design. Similarly, Sagoe et al. (2021) conducted a meta-analysis that included 35 studies using the crosswise design. The author found that this variant outperforms the direct questioning technique based on the 'more is better' validation criteria. The author also reported that the crosswise design is particularly effective when measuring the prevalence of highly sensitive issues. Overall, the existing evidence provides a solid basis for using the method to elicit genuine answers.

Randomised Item Response Theory. There have been several advances made in the RR technique's design and analysis. Psychometricians have recently incorporated Item Response Theory (IRT) insights into the RR design (Fox, Veen and Klotzke, 2019; Fox, 2005; Fox and Wyrick, 2008). IRT has been mainly developed in the field of Psychometrics to measure underlying attitudes or abilities as a function of an individual's item responses. In the IRT approach, instead of asking one or a small set of questions about the behaviour of interest, individuals respond to a group/battery of closely related questions to estimate an individual's latent trait of interest. Wu, Wood and Stevenson (2019) explains how the RR design incorporates the IRT within its setup. Rather than asking one sensitive question, I can ask a set of sensitive questions that helps determine each respondent's latent sensitive trait with a high probability.

There are two advantages of utilising RIRT over conventional RR designs and other indirect questioning techniques. Firstly, and most importantly, this technique yields individual-level estimates of the lurking trait of interest (Fox, 2002, 2005). In contrast, all conventional indirect questioning techniques only allow estimating the prevalence of the sensitive behaviour at the population or group level. Secondly, RIRT allows for hierarchical analysis where researchers can investigate group differences of the sensitive traits at sub-levels to the level that the data is gathered. Both features, in theory, would contribute to providing precise estimates of both population, sub-group and individual-level estimates of the behaviour in question. Fox, Avetisyan and van der Palen (2013) validated this technique finding that RIRT effectively identified smokers from non-smokers. The authors compared individual estimates of respondents' smoking trait versus to a post-survey expired-air CO tests results that revealed respondents' smoking status.

We can formalise how the RR design incorporates IRT. In the standard RR design, a respondent will answer affirmatively ($Y_{ijk} = 1$) to an item k is a function of π_{ijk}. This probability is calculated based on each respondent's latent trait. For instance, in this study, the underlying trait of interest is the individual propensity to engage in a particular type of corrupt behaviour. In the IRT literature, the θ_i parameter corresponds to the individual unobservable trait. We can further assume that individuals are nested in j groups, where individuals in some clusters are more likely to answer positively than others (if that is the case, this would be denoted as $\theta_i j$). In the RR set-up, given that we introduce some random noise, the probability of observing a positive answer ($Y_{ijk} = 1$) is the product of π_{ijk} and the probability of also answering affirmatively to the question related to the random device. In this case, the researchers know the probability of the random device, which we can denote as q. Putting all of these parameters

together, we can formalise the probability of answering positively in the RR set up to an item k as follows:

$$P(Y_{ijk} = 1) = p\pi_{ijk} + (1 - p)(1 - \pi_{ijk}) \tag{4.4}$$

In RIRT, the probability of answering positively also depends on each item's attributes. Thus, π_{ijk} is a function of α_k and β_k parameters. α_k represents the discrimination power of an item k. This parameter defines how item k differentiates individuals based on their underlying attribute(s). An item with high discrimination power gives a larger probability of answering positively to respondents with a greater underlying trait. Instead, it provides a lesser probability to those with a lower underlying trait. Conversely, an item with low discrimination power provides roughly the same probability levels to answering positively to questions with different levels of respondent's underlying traits. β_k is the difficulty parameter and determines the percentage of respondents answering favourably to an item. Thus, an item with a higher difficulty level yields a lower chance of positively answering a question. In contrast, an item with a lower difficulty level will provide a higher probability of what of answering affirmatively. If we include these two parameters into the response function, the likelihood of answering affirmatively to an item is as follows.

$$P(\tilde{Y}_{ijk} = 1|\theta_{ij}, \alpha_k, \beta_k) = \Phi(\alpha_k\theta_{ij} - \beta_k) \tag{4.5}$$

In this logistic model, the chance of answering positively to the item k is a function of the cumulative probability distribution function Φ and the parameters mentioned before. We can then integrate this into the RR's set up, where π_{ijk} is equal to $P(\tilde{Y}_{ijk} = 1|\theta_{ij}, \alpha_k, b_k)$:

$$\pi_{ijk} = P(\tilde{Y}_{ijk} = 1|\theta_{ij}, \alpha_k, \beta_k) = \Phi(\alpha_k\theta_{ij} - \beta_k) \tag{4.6}$$

Combining Equations 4.4 and 4.6, we can formally incorporate IRT into the RR design:

$$P(Y_{ijk} = 1) = p\Phi(\alpha_k\theta_{ij} - \beta_k) + (1 - p)(1 - \Phi(\alpha_k\theta_{ij} - \beta_k)) \tag{4.7}$$

As researchers know the probability distribution of the random device, we can then replace the value of p with its chance of answering affirmatively to the unsensitive device question. For example, in the case of a die, the probability of responding positively is equal to $1/6$, thus $p = 1/6$ and $1 - p = 5/6$; therefore, Equation 4.7 is equivalent to:

$$P(Y_{ijk} = 1) = \frac{1}{6}\Phi(\alpha_k\theta_{ij} - \beta_k) + \frac{5}{6}(1 - \Phi(\alpha_k\theta_{ij} - \beta_k)) \tag{4.8}$$

Thus, in addition to estimating the auxiliary parameters α_k, and β_k, we are be able to estimate respondents' latent trait θ_{ij}, which is the parameter of interest for this study.

4.3 Survey Sampling, Instrument, and Design

Sampling For this study, I conducted three online survey waves using the Nuffield Centre for Experimental Social Sciences subjects panel in Chile. The first wave was conducted in January 2020, where 6,044 respondents completed the survey. This survey included the RIRT items plus additional questions about beliefs about corruption, vote intention, partisanship and demographics.

The second wave was conducted five months after the first wave. The same 6,044 respondents who completed the first wave were invited to participate in the follow-up survey. From the 6,044 study participants that answered the first wave, 3,721 completed the second wave, with a 37% overall attrition rate.

I conducted the third wave by randomly sampling 1,163 respondents who answered at least one of the previous waves. In this round, rather than having all eleven items, I included only three items, but in a direct questioning format. The prevalence estimates obtained from this wave served as a benchmark to compare against the estimates obtained from the RIRT waves. Although the results from this survey serves to conduct a comparative validation, this is not the primary goal of this paper. The results of this comparative validation are reported in section 4.7 in the Appendix. I complemented this comparative validation analysis by comparing the population-level estimates of bribes obtained from the RIRT to estimates obtained from a representative survey (n = 2900) conducted annually in Chile [7] that uses a direct questioning format[8].

Within all the RR variants, I chose the crosswise design as it provides a stronger signal of higher levels of confidentiality, higher efficiency, and a moderate level of instruction complexity (Blair, Imai and Zhou, 2015; Hoffmann and Musch, 2016; Höglinger, Jann and Diekmann, 2016; Korndörfer, Krumpal and Schmukle, 2014; Hoffman et al., 2015). Following the design suggestions from Wu, Wood and Stevenson 2019, 11 items were included in the survey, which is a sufficient number of items to achieve lower levels of bias and minimise Root Square Mean Errors.

The selection of questions was based on several surveys, such as the Latin America Public Opinion Project LAPOP (2017), Chile's Consejo para la Transparencia de la Transparencia (2019), and other surveys and studies that have measured clientelism, electoral violence, and vote-buying (Fergusson, Molina and Riano, 2017; Transparente, 2019). Table 4.1 reports the list of items included in the RIRT survey waves.

To ensure respondents understood the RIRT instructions, I included three test rounds before the respondents answered the sensitive questions in the RIRT format. Participants that answered correctly in the first two rounds were allowed to continue with the sensitive questions. Subjects that responded incorrectly in the first two test rounds were asked to answer the third test before continuing with the RIRT section.

[7] This survey is commissioned the Consejo para la Transparencia, which is an independent non-partisan entity in Chile.

[8] In this survey they only asked about bribes, so I were not able to make additional comparison.

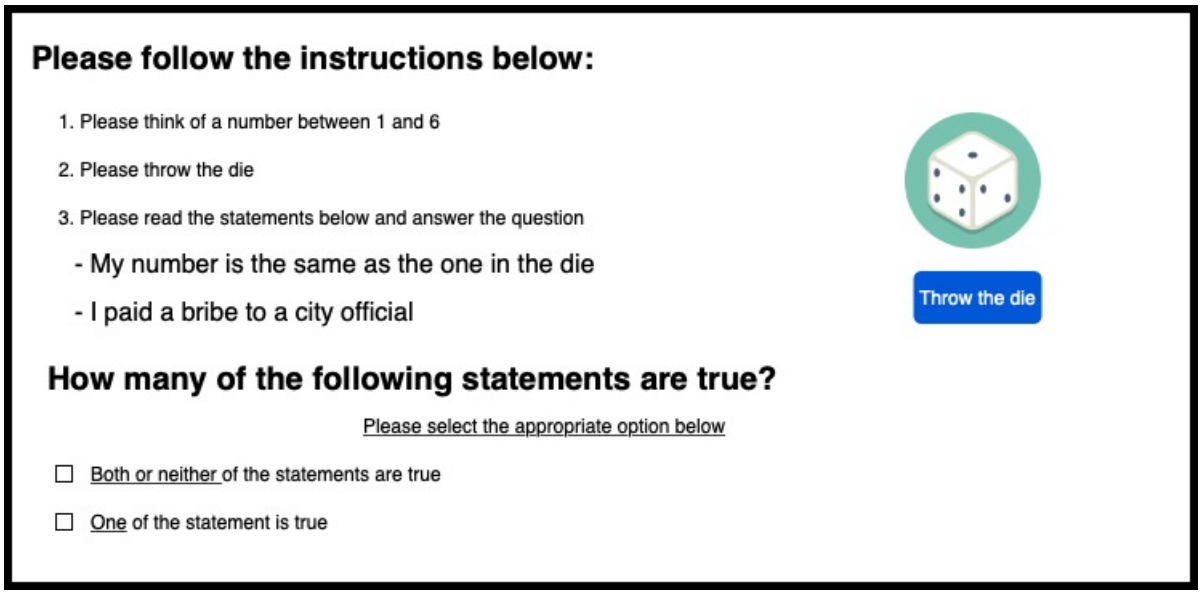
In the second wave of the RIRT, I include four 'nested' items reported at the end of Table 4.1. These questions are designed to flag inattention or poor comprehension, as there are specific behaviours of four core corrupt behaviours. Thus, in theory, nested items should yield lower prevalence estimates relative to their corresponding core item. Finally, I incorporated in the second wave of the RIRT a *Anchor item* that serves as an additional check of attentiveness. The logic of item is that if respondents pay attention, we should expect to yield near-zero prevalence estimates for that item, as the prevalence of that type of behaviour is very unlikely to occur in the population. I provide more details of these strategies to pin down inattentive respondents in Section 4.6 of this manuscript.

Table 4.1 – List of items

	Core Items	**Item name**
1	I know personal cases either of friends or relatives who used their personal or political connections to be hired by the municipality	Patronage
2	I used personal or political connections to get better service than others in my municipality	Political connections
3	I received money or been offered a special favour, gift or benefit by my municipality to vote in a particular way	Vote-buying
4	I had to give money or goods or gifts to a public official of my municipality so that they could give or award me a contract, service, or benefit.	Bribes
5	A public official of the municipality helped me alter administrative documents or information to obtain a service, benefit, or contract	Forgery
6	A public official allowed me to use public resources of the municipality for personal purposes	Personal
7	A public official of my municipality asked me for sexual favours, either suggestively or openly, in exchange for a benefit or service	Sexual Solicitation
8	I pressured municipal officials to skip administrative processes or laws	Lobbying
9	I used personal or political connections or paid public officials of my municipality so they will not oversee an administrative process or project	Compliance
10	A City official threatened me to vote or not vote in a certain way	Electoral violence
11	A City official shared with me privileged information about the municipality	Privilege
	Nested Item	
12	I know cases where close relatives such as *siblings/partners* used their personal or political connections to be hired by the municipality	Patronage
13	I had to give money, gifts or goods to a public official of my municipality so I could obtain a *social protection card*	Bribes
14	A public official of the municipality has helped me alter administrative documents *to obtain my driver's license or a certificate*	Forgery
15	Officials of my municipality shared with me privileged information about *possible social benefits* that the municipality provides	Privilege
	Anchor Item	
16	An official from my municipality asked me to pay a bribe to their bank account abroad	

In the crosswise method implemented in this study, respondents were asked to think of a number between one and six and roll a virtual die. Then, they read two statements: a sensitive statement and an unrelated non-sensitive statement. The non-sensitive statement was, "My number is the same as the one on the die". The sensitive statement was one of the eleven corrupt behaviours reported in Table 4.1. Finally, study participants answered how many of the previous statements were true. Subjects had to chose between option A: *both or neither of the statements are true* or option B: *just one of the statements is true*. Figure 4.1 shows a screenshot of the instructions that respondents saw when they answered the survey.

Figure 4.1 – An example of a randomised response survey item



Note: This figure shows what respondents observed answered the RIRT survey. Study participants were asked to think of a number between 1 and 6. Then, they were asked to throw a virtual die. They were informed that the researchers did not know what was the outcome of the die. Finally, they had to read the two statements, one insensitive and unrelated statement that asked whether the number they thought was the same obtained from rolling the die. The second statement was about sensitive behaviour. The order of the statements was randomised; thus, for some participants, the sensitive statement came up first, and for some respondents, the sensitive statement came up second. Finally, respondents were asked to answer whether the 'Both or neither' statement are true or 'One' of the statement is true.

4.4 Estimation and Inference

Using RIRT allows obtaining population-level estimates of the sensitive behaviours and individual-level estimates of the underlying sensitive trait. We can use either a modified version of a logistic or ordinary least squares models to estimate group-level estimates. Whereas for individual-level estimates, we can use a fully Bayesian estimation method with MCMC sampling for the posterior distribution of the parameters.

Population-level estimates

Modified logistic regression and linear probability models When using binary responses to measure the prevalence of sensitive issues, we can use an adapted version of a logistic regression that incorporates the randomised response design into the logistic function (Bourke and Moran, 1988; Liu and Zhang, 2017*a*,*b*). Following Yu, Tian and Tang (2008)'s formalisation of the RR estimation approach, the underlying trait π is a probabilistic measure of the $P(X = 1|Z)$. This probability of answering yes to the sensitive questions is conditional on set of covariates Z that predict engaging to the sensitive trait. In the logistic regression set up, we can derive β, which is unknown parameter of interest (proportion), which we can estimate using maximum likelihood estimation.

$$\pi = \frac{Z'\beta}{1+Z'\beta} \tag{4.9}$$

As we do not observe X, which is whether respondent has engaged on the sensitive behaviour $(X = 1)$, the maximum likelihood for β is as follows:

$$\ln L(\beta|X, Z) = \sum_{i=1}^{n} [X_i \ln(\pi_i) + (1 - X_i) \ln(1 - \pi_i)] \tag{4.10}$$

In the crosswise design, we have a response variable R that takes the following values, that is conditional on the respondents' exposure to the sensitive trait and the result of the unrelated question:

$$R = \begin{cases} 1, \text{ if response is both } (X = 1 \text{ and } Y = 1 \text{ or } X = 0 \text{ and } Y = 0) \\ 0, \text{ if response only one } (X = 1 \text{ and } Y = 0 \text{ or } X = 0 \text{ and } Y = 1) \end{cases} \tag{4.11}$$

We know the probability of $Pr(Y = 1|Z)$ of answering yes to the unrelated question, which is equal to p. Therefore the probability of $R = 1$ is equal to $\pi p + (1 - \pi)(1 - p)$. Whereas the probability of $R = 0$ is equal to $\pi(1 - p) + (1 - \pi)p$. We then can rewrite the log-likelihood function as follows:

$$\ln L(\beta|R, Z) = \sum_{i=1}^{n} R_i \cdot \ln[\pi_i p_i + (1 - \pi_i)(1 - p_i)] + (1 - R_i) \cdot \ln[\pi_i(1 - p_i) + (1 - \pi_i)p_i] \tag{4.12}$$

$$\ln L(\beta|R, Z) = \sum_{i=1}^{n} R_i \cdot \ln[\pi_i e^{Z_i'\beta} + (1 - \exp^{Z_i'\beta})(1 - p_i)] + (1 - R_i) \cdot \ln[e^{Z_i'\beta}(1 - p_i) + (1 - e^{Z_i'\beta})p_i] \tag{4.13}$$

We can also estimate the proportion of respondents that have engaged in the sensitive behaviour using a modified version of a linear probability model. In this set up, the probability of

engaging of the sensitive behaviour, conditional on a vector of covariates is equal $Pr(X = 1|Z)$. This probability is equal to $Z'\beta$. We can derive β from the following transformation:

$$E(R = 1|Z) = Pr(R = 1|Z) = (Z'\beta)p + (1 - Z'\beta)(1 - p) \tag{4.14}$$

$$E(R = 1|Z) = (Z'\beta)(2p - 1) + 1 - p \tag{4.15}$$

The response variable is equal to $\hat{R} = (R + p - 1)/(2p - 1)$, we can replace this function into equation 4.15:

$$E[(R + p - 1)/(2p - 1)|Z] = \frac{E(R = 1|Z) + p - 1}{2p - 1} = Z'\beta \tag{4.16}$$

I report the linear probability estimates in Figure 4.13 in the Appendix.

Individual Response Probabilities

A key contribution of conducting a RIRT is that it can yield individual-level estimates of the latent trait of interest and probabilities of engaging in the behaviour in question. As I pointed out earlier, there is a debate as to whether these latent estimated traits reflect actual behavioural proclivities.

We can formalise the relationship between the observed randomised responses and the latent and true responses, assuming that respondents are clustered in $kj = 1...J$ groups. Thus there are $i = 1, ..., n_j$ individuals in each group. As before, we can state Y_{ijk} as the individual's latent trait.

$$P(Y_{ijk} = 1) = \phi\Phi(\alpha_k\theta_{ij} - b_k) + \phi(1 - \Phi(\alpha_k\theta_{ij} - b_k)), \text{where } \phi = \frac{1}{6} \tag{4.17}$$

I use a Bayesian approach to estimate individuals' underlying trait by conducting a two-parameter RIRT model (Fox, Veen and Klotzke, 2019; Fox, 2005; Fox and Wyrick, 2008). I use two model identification approaches: First, we identify the model by fixing the mean and variance of the ability parameters (θ_{ij})' to zero and one. This minimum identification strategy yields reliable ability parameter estimates. At the same time, the discrimination parameters do not converge well with this model. I believe it is because most sensitive questions show very low population-level prevalence rates ($< 10\%$).

As an additional identification strategy, I anchored two difficulty item parameters that yield the highest and the lowest group-level prevalence rates. While the individual-level ability parameters are consistent across the two models, I obtain more reliable discrimination parameters using the second model. I use a dispersed normal prior $(N(0, 100))$ for the difficulty parameters and a uniform prior $(U(0, 5))$ for the discrimination parameters. The equations below summarise the estimation strategy

$$Y_{ik}|\alpha_k, \beta_k, \theta_i \sim \text{Bernoulli}(\pi_{ik}) \tag{4.18}$$

$$\pi_{ik} = \phi\Phi(\alpha_k\theta_i - \beta_k) + \phi(1 - \Phi(\alpha_k\theta_i - \beta_k)) \tag{4.19}$$

$$\alpha_k \sim Uniform(0, 5) \tag{4.20}$$

$$\beta_{(-k_{highest}, -k_{lowest})} \sim N(0, 100) \tag{4.21}$$

$$\beta_{k_{highest}} = \Phi^{-1}(\pi_{k_{highest}}) \text{ and } \beta_{k_{lowest}} = \Phi^{-1}(\pi_{k_{lowest}}) \tag{4.22}$$

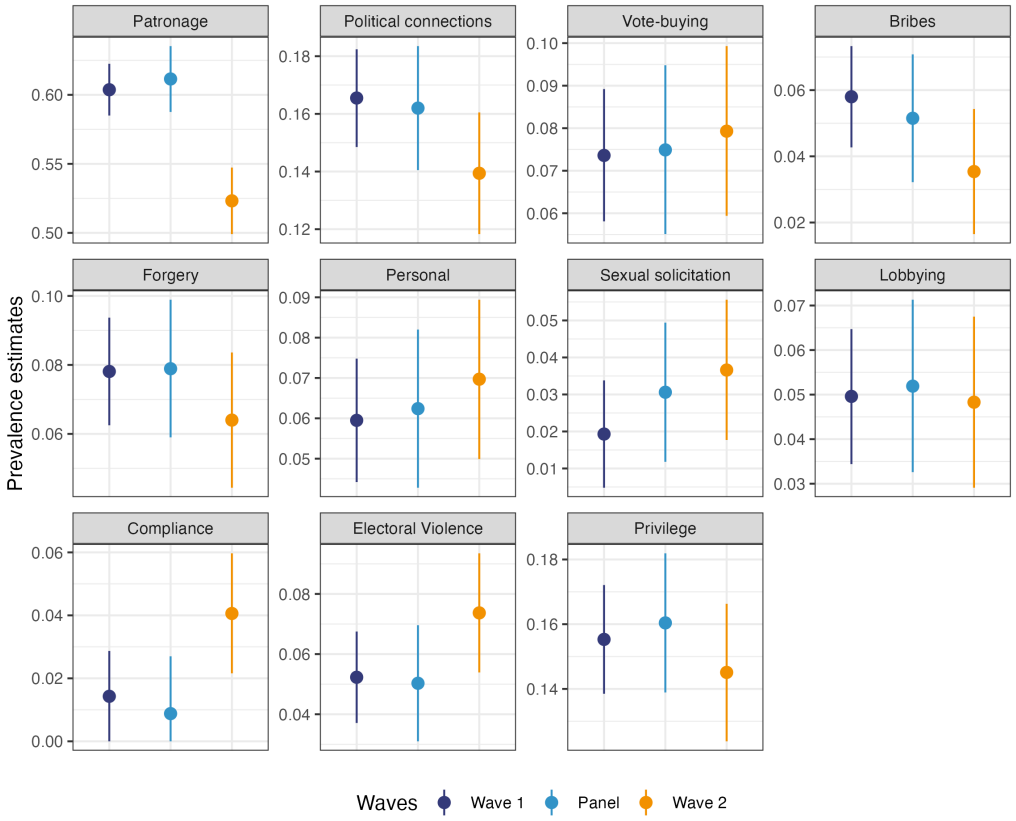
$$\theta_i \sim N(0, 1) \tag{4.23}$$

4.5 Results

Population-level prevalence

I report the prevalence of each sensitive behaviour using equation 4.2. As presented in Figure 4.2, there is variation across the different corrupt behaviours, going from 52-60% of cases related to patronage to around 1-4% to paying bribes to reduce administrative compliance. I can also observe that the estimated prevalence of the different corrupt behaviours tends to be relatively similar across both waves. However, there is a significant drop of around 8% in the case of patronage. Table 4.6 in the appendix reports the *Naive* estimates for both waves, as well as first wave respondents that also took part on the second wave, I refer to this sub-sample in this manuscript as *Panel*. I observe similar estimates between *Panel* and *Wave 2*, where in some items the prevalence estimates are the same to differences of a 9%. I complemented this analysis providing estimates from a linear probability model regression estimates for both waves of the RIRT in Figure 4.13 in the Appendix. Again, estimates tend to be similar across time, except for patronage, which drops substantially on the second wave.

Figure 4.2 – Prevalence estimates of each item - RIRT Waves



Note: This figure report prevalence estimates for all 11 items included both in *Wave 1*, *Panel*, and *Wave 2*. These are the 'Naive' and unweighted estimates. In dark blue, I show estimates for the first wave, in light blue for the *Panel*, and in yellow, the results from *Wave 2*. We can derive from this plot that there is only a statically significant difference for the 'Patronage' item.

I also find considerable differences between *Wave 1*, and *Panel* and *Wave 2* for the 'Compliance' and 'Electoral Violence' items. For the 'Compliance' item, prevalence rates were around 0.015, whereas for the *Wave 2* estimates around 0.04. Similarly, for the 'Electoral Violence' item, estimates both in the first wave and the *Panel* respondents was around 0.05, but then increased up two per cent in the follow-up wave. The differences between *Wave 1* and *Wave 2* are partly due to temporal changes (assuming there is full instruction compliance and full attentiveness) in prevalence rates of corrupt behaviours. We can rule out that these changes are due to changes in the sampling composition, given that the prevalence estimates obtained from *Panel* are closely pegged to the estimates from the first wave. Differences on the prevalence rates and their respective p-values are summarised in Table 4.3.

Given this study was administered using a sample from an opt-in unrepresented respondents' pool, I provide estimates in Figure 4.4 that include sample weights of the naive estimates. I applied inverse probability weighting (Horvitz and Thompson, 1952) to account for population differences across communes. This weighting strategy shows that the results of the estimates are somewhat consistent with this weighting scheme, as roughly eight of the items yield similar estimates. I extended this analysis using different weights. The results of this additional analysis

are documented in Figure 4.14 in the Appendix.

Table 4.2 – Prevalence estimates - Naive Unweighted and Weighted Prevalence Estimates

	Wave 1		Panel		Wave 2	
Core Items	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Patronage	0.60	0.59	0.61	0.60	0.52	0.51
	(0.009)	(0.009)	(0.012)	(0.012)	(0.012)	(0.012)
Political connections	0.17	0.18	0.16	0.18	0.14	0.15
	(0.008)	(0.008)	(0.011)	(0.011)	(0.010)	(0.010)
Vote-buying	0.07	0.07	0.07	0.07	0.08	0.08
	(0.007)	(0.007)	(0.010)	(0.010)	(0.010)	(0.010)
Bribes	0.06	0.07	0.05	0.06	0.04	0.05
	(0.007)	(0.007)	(0.009)	(0.010)	(0.009)	(0.009)
Forgery	0.08	0.09	0.08	0.08	0.06	0.07
	(0.008)	(0.008)	(0.010)	(0.010)	(0.010)	(0.010)
Personal	0.06	0.05	0.06	0.06	0.07	0.08
	(0.007)	(0.007)	(0.010)	(0.009)	(0.010)	(0.010)
Sexual solicitation	0.02	0.02	0.03	0.03	0.04	0.04
	(0.007)	(0.007)	(0.009)	(0.009)	(0.009)	(0.009)
Lobbying	0.05	0.05	0.05	0.06	0.05	0.05
	(0.007)	(0.007)	(0.009)	(0.010)	(0.009)	(0.009)
Compliance	0.01	0.03	0.01	0.03	0.04	0.04
	(0.007)	(0.007)	(0.009)	(0.009)	(0.009)	(0.009)
Electoral Violence	0.05	0.06	0.05	0.07	0.07	0.09
	(0.007)	(0.007)	(0.009)	(0.010)	(0.010)	(0.010)
Privilege	0.16	0.15	0.16	0.15	0.15	0.15
	(0.008)	(0.008)	(0.011)	(0.010)	(0.010)	(0.011)
Nested - Patronage					0.17	0.17
					(0.011)	(0.010)
Nested - Bribes					0.05	0.04
					(0.009)	(0.090)
Nested - Forgery					0.04	0.04
					(0.009)	(0.009)
Nested - Privilege					0.11	0.11
					(0.010)	(0.010)
Num.Obs.	6044	6044	3721	3721	3721	3721

Note: This table reports both unweighted and weighted estimates for *Wave1*, *Wave 2* and *Panel* for all 11 items and nested items. Estimates are reported as proportions of the whole population. Weighted estimates are obtained using Inverse Probability Weighting based on population size.

Table 4.3 – Difference in Prevalence Weighed Naive Estimates and their Respective p-values

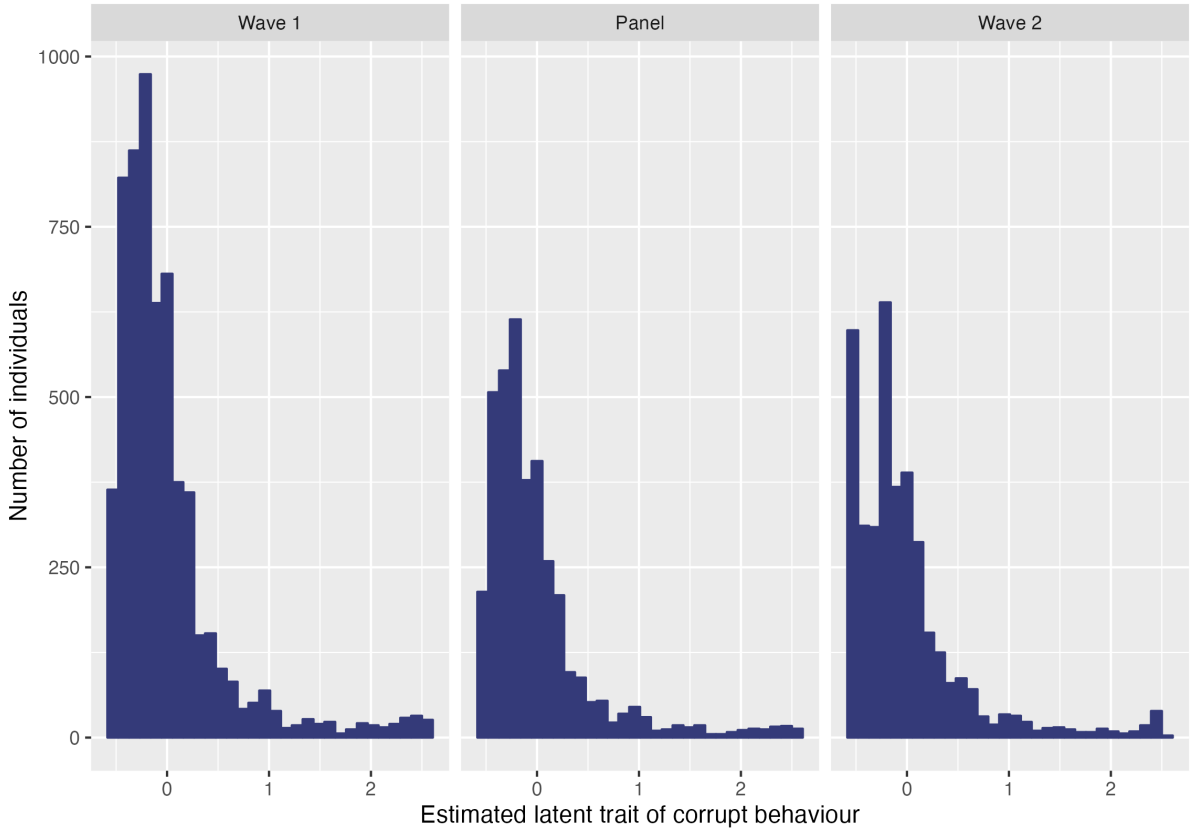
	Wave 1 - Panel		Panel - Wave 2		Wave 1 - Wave 2	
Items	Diff	p-value	Diff	p-value	Diff	p-value
Patronage	-0.011	1.730	0.087	0.000	0.075	0.000
Political connections	-0.002	1.207	0.026	0.002	0.024	0.002
Vote-buying	-0.002	1.253	-0.008	1.818	-0.010	1.931
Bribes	0.007	0.162	0.013	0.011	0.021	0.000
Forgery	0.002	0.783	0.011	0.078	0.013	0.027
Personal	-0.003	1.460	-0.024	2.000	-0.027	2.000
Sexual solicitation	-0.010	1.997	-0.004	1.630	-0.014	2.000
Lobbying	-0.009	1.943	0.014	0.009	0.005	0.308
Compliance	0.000	0.928	-0.014	1.999	-0.014	2.000
Electoral Violence	-0.003	1.403	-0.020	1.999	-0.022	2.000
Privilege	-0.002	1.210	0.008	0.359	0.006	0.452

Note: This table reports the difference in proportion for each of the 11 core items. I only register on this table the differences of prevalence rates using only weighted naive estimates. I documented three differences: 1) Prevalence estimates between *Wave 1* and *Panel*, 2) between *Panel* and *Wave 2*; 3) between *Wave 1* and *Wave 2*.

Individual-level estimates

I estimated individual-level latent traits of corrupt behaviour for the 6,044 respondents in the first wave. I also repeated the estimation for 3,721 respondents in the second wave and *Panel*. I run the MCMC algorithm with 5,000 burn-ins and 5,000 iterations after the burn-ins. Figure 4.3 shows the distribution of individual-level RIRT estimates of the latent trait of corrupt behaviour (θ_i) in each survey wave. As shown in Figure 4.3, many corrupt behaviour items show low group-level prevalence rates. Therefore, the distribution of individual-level latent trait estimates is right-skewed. In other words, because most corrupt behaviours asked did not occur (or were not admitted by respondents), the majority of respondents' latent corruption scores are low. On the other hand, a small group of respondents who have admitted committing these corrupt behaviours receive extremely high estimates.

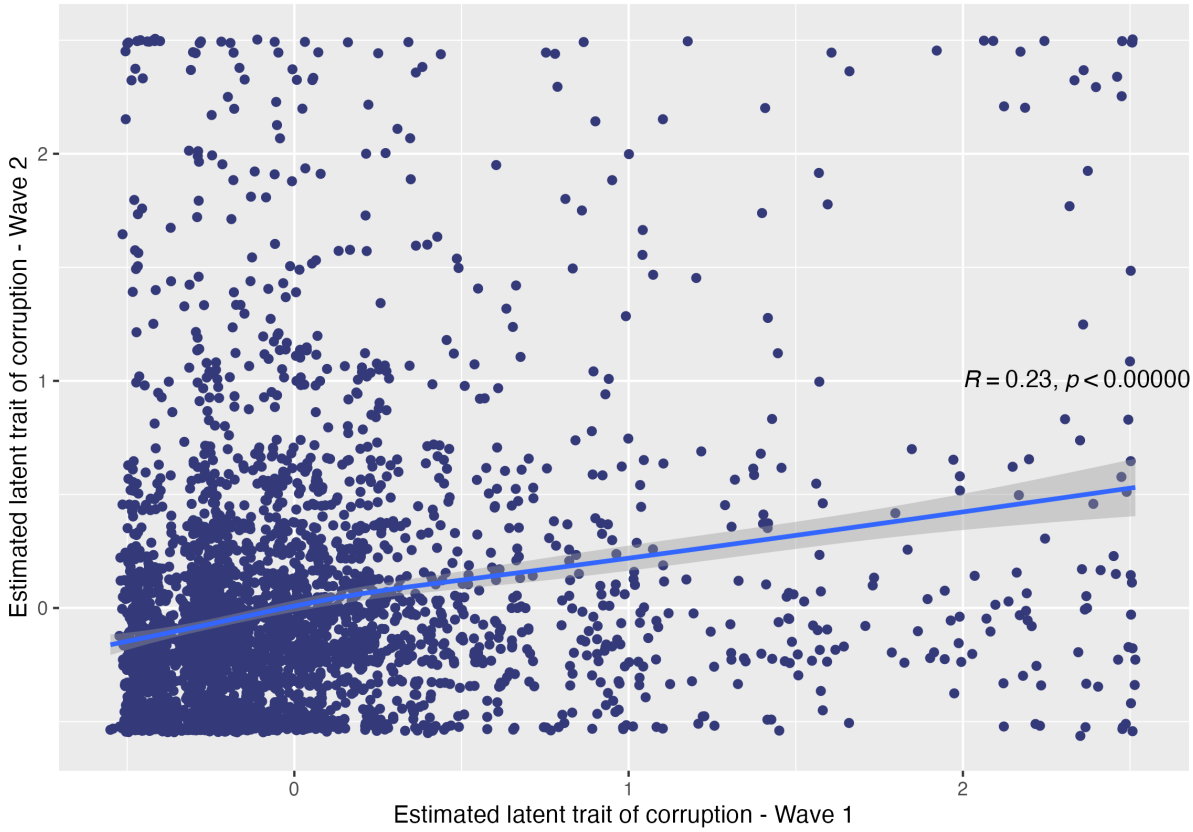
Figure 4.3 – Estimated Individual-level Latent Trait of Corrupt Behaviour



Note: This figure shows the distribution of respondents' latent trait of corrupt behaviour for both *Wave 1* and *Wave 2*. This figure illustrates that subjects' corrupt trait distribution is right-skewed. Most observations are clustered around zero in both waves. We can also see outliers that generate large tails for the highest values of these estimates.

Figure 4.4 shows the relationship between the RIRT estimates in the first and second waves. Although the correlation is not very strong ($R = 0.23$), it is positive. This result would suggest different levels of compliance and attentiveness across respondents. For example, we can explain a subject scoring a lower value on her underlying trait on the second wave for several reasons: 1) The respondent forgot the instances where she was engaged in the corrupt behaviours asked in the surveys. 2) The subject preferred not to disclose one or several corrupt behaviours asked in the second wave. 3) The subject did not comply or understand the instructions in at least one of the waves. In contrast, subjects with higher or relatively similar scores on their underlying trait are perfectly reasonable results. For instance, respondents may or may not have engaged in at least one of the sensitive behaviours asked in the surveys between each wave.

Figure 4.4 – Estimates of Individual-level Latent Trait of Corrupt Behaviour in Both Waves



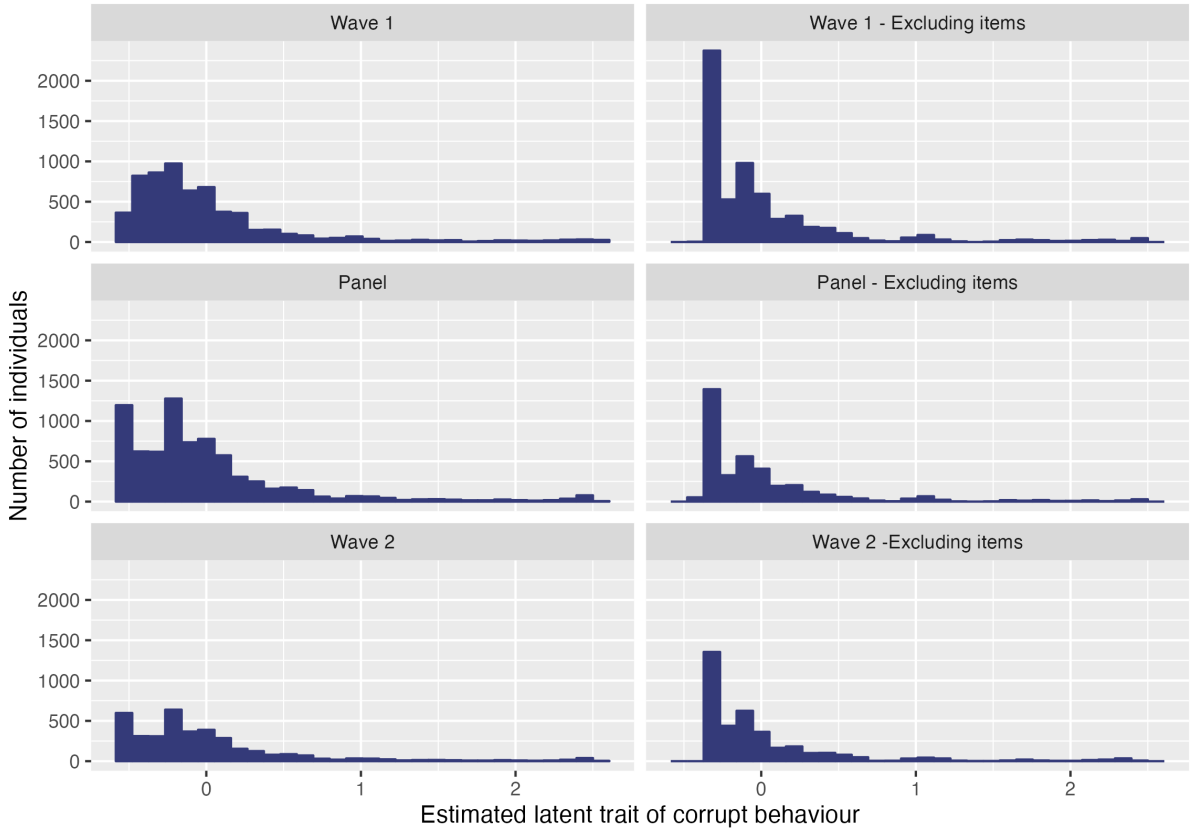
Note: This figure is a scatterplot that reports estimates of respondents' corrupt behaviour trait from *Panel* and *Wave 2*. On the x-axis are the estimates from the *Panel*, and on the y-axis are the estimates from *Wave 2*. I also report the coefficient of correlation and add a regression line that shows a positive relationship between these two estimates.

Item Analysis One concern in estimating respondents' underlying trait is selecting items that accurately measure the dimension of interest. Usually, researchers select questions from a pool of validated items with well-known and precisely estimated parameters such its discrimination and difficulty (Baker and Kim, 2017). However, researchers may want to drop some items in exchange of obtaining more accurate estimates. For example, researchers may want to remove items ex-post because these questions perform poorly in capturing respondents' unobservable trait (Hambleton and Swaminathan, 1985) [9]. In this study, I estimated individuals' underlying trait, dropping the 'Patronage' and 'Sexual solicitation' items. I excluded these items for the following reasons: 1) Both questions may not be capturing the respondent's underlying propensity to engage in corrupt behaviour. In the case of 'Patronage', although this question has been used in surveys that measure patronage (Fornasari et al., 2022), the question opens up the possibility for respondents to claim that they are aware of cases of patronage from third parties. For the case of 'Sexual solicitation', the question's format asks whether respondents have been victims of sexual

[9]For example, Yang et al. (2013) tested the Confusion Assessment Method to identify Delirium (acute confusion), and their suggested dropping items from this instrument to maximise the probability of identifying this illness

coercion, but not whether the corruption behaviour is not initiated by the respondents. The results of this analysis are in Figure 4.5 where we see a considerable change in the distribution of subjects' underlying trait in all waves.

Figure 4.5 – Estimates of Individual-level Latent Trait of Corrupt Behaviour in Both Waves - Excluding Items



Note: This figure shows the distribution of respondents' underlying trait of corrupt behaviour for *Wave 1*, *Wave 2* and *Panel*. I report the original estimates using all 11 core items in the left-hand column. The distributions in the right column correspond to the distribution of respondents' corrupt trait, excluding the 'Patronage' and 'Sexual solicitation' items.

In both cases, estimates are clustered around zero; I find that the distribution of these estimates is susceptible to change once I drop these two items. I performed a Wilcox sign rank test to compare the means of two dependent non-normal distributions. I find statistically significant differences between the two distributions for waves 1 and 2. I do not find statistically significant differences for *Panel* ($p = 0.037$). More details of the magnitude of individual-level changes are in the Appendix in Table 4.7.

4.6 Attention Checks and Estimation Adjustments

One of the shortcomings of the RR designs is that they put a substantial burden on the respondents (Jerke et al., 2021). This additional burden can lead to inattention, lack of coop-

eration and compliance with the RR instructions. In RIRT, this is particularly important as respondents are required to answer a long list of sensitive items.

A common approach to mitigate the inattentive bias is adding a compliance/attention check RR item whose population-level prevalence estimate is approximately zero (Höglinger and Diekmann, 2017; Atsusaka and Stevenson, 2021). Another common approach is to include a test RR item whose randomisation device (e.g., secret number) is known to the researcher. Researchers can use these attention check techniques to screen out non-compliant respondents from the sample.

Atsusaka and Stevenson (2021) proposed another approach where the randomisation device is unknown to researchers and the sensitive item is related to the primary research subject. This approach is better at testing compliance because the item is in the same format as the actual RR items. In their design-based attention check question, their include an RR item known to have zero population-level prevalence rate. Equations 4.25 and 4.27 provides the bias-corrected estimator and its variance proposed by these authors:

$$P(Y=1) = \lambda = \{p\pi_{CM} + (1-p)(1-\pi_{CM})\}\gamma + \kappa(1-\gamma) \tag{4.24}$$

$$B_{CM} \equiv E[\hat{\pi}_{CM}] - \pi = \left(\frac{1}{2} - \frac{1}{2\gamma}\right)\left(\frac{\lambda - \kappa}{p - \frac{1}{2}}\right), \text{ where } \kappa = \frac{1}{2} \tag{4.25}$$

$$\hat{\pi}_{BC} = \hat{\pi}_{CM} - \hat{B}_{CM} \tag{4.26}$$

$$Var(\hat{\pi}_{BC}) = V\left[\frac{\hat{\lambda}}{\hat{\lambda}'}\left(\frac{\frac{1}{2} - p'}{2p-1}\right)\right] \tag{4.27}$$

γ is the proportion of attentive respondents, κ is the probability with which inattentive respondents pick "both or neither is true." B_{CM} is the bias with respect to the quantity of interest caused by inattentive respondents.

Bias-corrected population-level prevalence

To address this issue, I implemented Atsusaka and Stevenson (2021) attention check technique that helps identifying inattentive respondents. I used the 'Anchor item' reported at the end of Table 4.1 to correct the bias due to inattentiveness. Respondents were asked the following question: "An official from my municipality asked me to pay a bribe to their bank account abroad"[10]. This item should yield near zero prevalence estimates; as in this context, it is improbable that local government officials would ask citizens to deposit bribes into to foreign bank accounts.

[10]I use their *cWise* package to produce the bias-corrected population-level prevalence rates.

Table 4.4 – Prevalence estimates - Naive, Weighted and Bias-Corrected

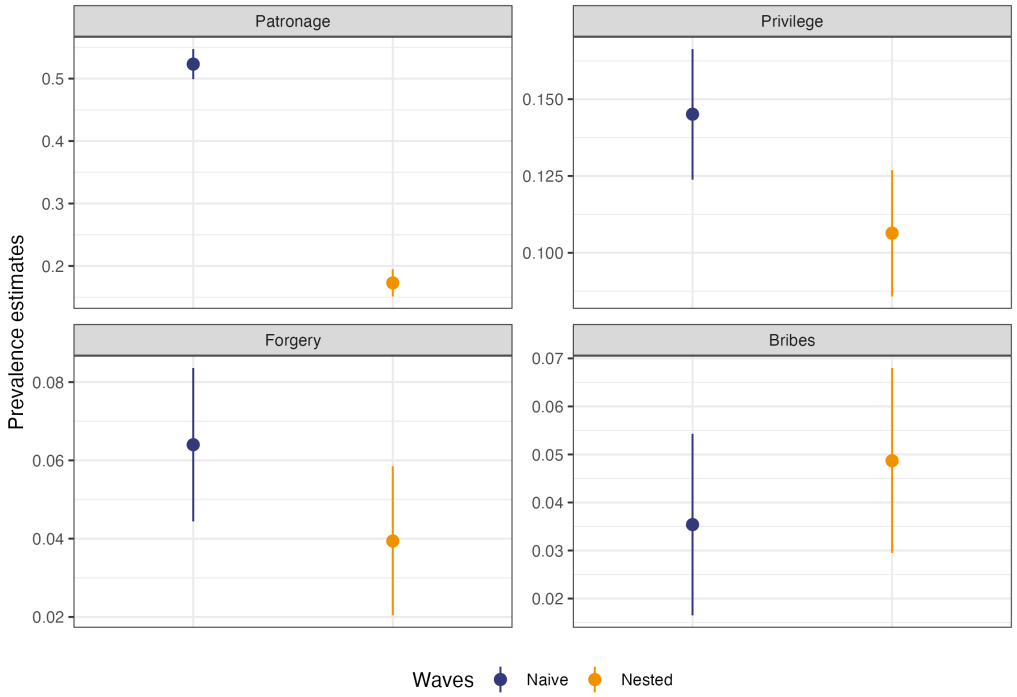
	Wave 1				Wave 2			
	Whole sample		Panel					
Core Items	Unweighted	Weighted	Unweighted	Weighted	Unweighted-Naive	Unweighted-Corrected	Weighted-Naive	Weighted-Corrected
Patronage	0.60	0.59	0.61	0.60	0.52	0.52	0.51	0.51
	(0.009)	(0.009)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.016)
Political connections	0.17	0.18	0.16	0.18	0.14	0.13	0.15	0.15
	(0.008)	(0.008)	(0.011)	(0.011)	(0.010)	(0.013)	(0.010)	(0.016)
Vote-buying	0.07	0.07	0.07	0.07	0.08	0.07	0.08	0.07
	(0.007)	(0.007)	(0.010)	(0.010)	(0.010)	(0.013)	(0.010)	(0.016)
Bribes	0.06	0.07	0.05	0.06	0.04	0.03	0.05	0.04
	(0.007)	(0.007)	(0.009)	(0.010)	(0.009)	(0.012)	(0.009)	(0.016)
Forgery	0.08	0.09	0.08	0.08	0.06	0.05	0.07	0.07
	(0.008)	(0.008)	(0.010)	(0.010)	(0.010)	(0.013)	(0.010)	(0.015)
Personal	0.06	0.05	0.06	0.06	0.07	0.06	0.08	0.08
	(0.007)	(0.007)	(0.010)	(0.009)	(0.010)	(0.012)	(0.010)	(0.015)
Sexual solicitation	0.02	0.02	0.03	0.03	0.04	0.03	0.04	0.03
	(0.007)	(0.007)	(0.009)	(0.009)	(0.009)	(0.012)	(0.009)	(0.014)
Lobbying	0.05	0.05	0.05	0.06	0.05	0.04	0.05	0.05
	(0.007)	(0.007)	(0.009)	(0.010)	(0.009)	(0.012)	(0.009)	(0.015)
Compliance	0.01	0.03	0.01	0.03	0.04	0.03	0.04	0.04
	(0.007)	(0.007)	(0.009)	(0.009)	(0.009)	(0.012)	(0.009)	(0.014)
Electoral Violence	0.05	0.06	0.05	0.07	0.07	0.06	0.09	0.08
	(0.007)	(0.007)	(0.009)	(0.010)	(0.010)	(0.012)	(0.010)	(0.016)
Privilege	0.16	0.15	0.16	0.15	0.15	0.14	0.15	0.14
	(0.008)	(0.008)	(0.011)	(0.010)	(0.010)	(0.013)	(0.011)	(0.015)
Nested - Patronage					0.17	0.17	0.17	0.17
					(0.011)	(0.012)	(0.010)	(0.015)
Nested - Bribes					0.05	0.04	0.04	0.04
					(0.009)	(0.013)	(0.090)	(0.015)
Nested - Forgery					0.04	0.03	0.04	0.04
					(0.009)	(0.012)	(0.009)	(0.016)
Nested - Privilege					0.11	0.10	0.11	0.10
					(0.010)	(0.012)	(0.010)	(0.017)
Num.Obs.	6044	6044	3721	3721	3721	3721	3721	3721

Note: This table reports group-level both Naive and Bias corrected estimates of the corrupt behaviours. These estimates are reported as proportions of the population engaged in the sensitive behaviour. I report weighted and unweighted estimates. I use post-stratification to adjusts the sampling and replicate weights so that the joint distribution of a set of post-stratifying variables matches the known population joint distribution. For post-stratification weights I used demographic variables such as age, education attainment, occupation, and gender.

The bias-corrected estimates using the check item are reported in Table 4.4. Using only the sample of the second wave, the estimated attentive rate is 97.8%. Because of the high attentive rate, there is little difference between the 'naive' RIRT estimates and the bias-corrected RIRT estimates. Overall, The bias-corrected estimates are lower than the naive estimates by $\sim 1\%$.

Nested Items As I stated before, I included four nested items to evaluate respondents' attentiveness and compliance with the RIRT design. These nested items are a sub-category of the eleven core items included in the survey; therefore, I should expect lower prevalence estimates once we compare them to their respective core items. Figure 4.6 reports the proportions of both core and nested items of the second wave. I find lower prevalence estimates for three out of the fourth nested items, except for the 'Bribes' item. This evidence implies a reasonable level of diligence, comprehension, and compliance from respondents. Table 4.8 in the Appendix reports proportions and their respective confidence intervals for unweighted, weighted, naive and bias-corrected estimates.

Figure 4.6 – Comparison of Naive RR Estimates Versus Nested Items



Note: This figure shows the point estimates and confidence intervals for the 'Patronage', 'Privilege', 'Forgery' and 'Bribes' items. It provides estimates both from the core and nested items. We can see that the prevalence rate of the core items is higher for three out of the fourth items. Both estimates are from the second wave, given that we only implemented this attention check technique for this wave.

Sub-sample individual-level estimates

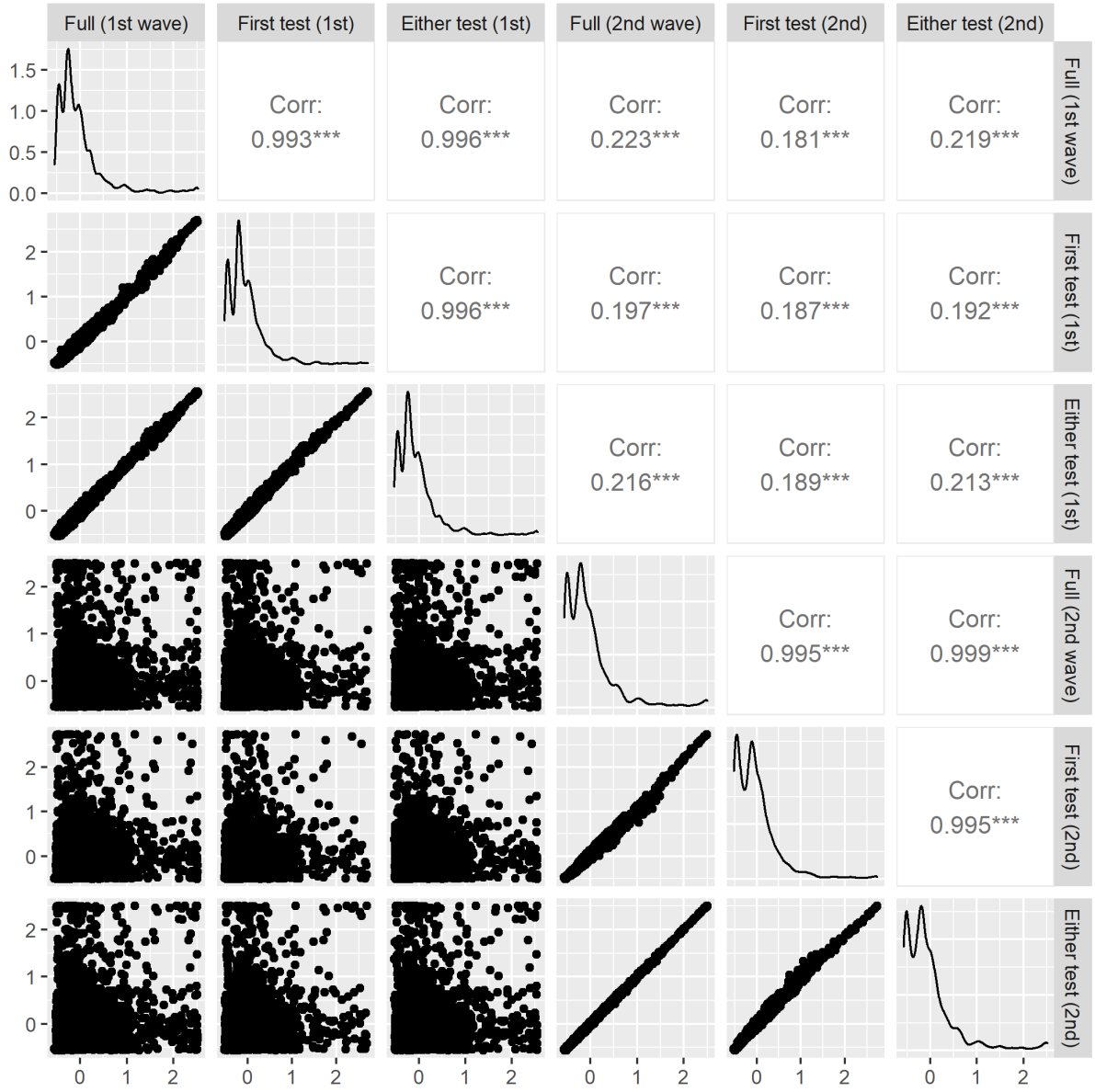
The above strategy to correct inattentiveness bias is only available for population-level estimates. Therefore, I take a different approach to examining the effect of inattentiveness on individual-level estimates. At the beginning of the survey, respondents answered two RIRT compliance test questions, which help me to examine respondents' comprehension of the RR design. If a respondent answers the first and second test questions correctly, the person does not get the third question. If a respondent failed to answer the first or second test question correctly, they had to answer the third test question. I use this design to ensure they understand the RR design well.

Utilising this information, I conduct two additional RIRT estimations for each wave. First, I re-estimate the individual-level estimates for those who correctly answered the first RIRT compliance test question. In the first wave, 5,217 respondents out of 6,044 answered this question correctly. In the second wave, 3,263 out of 3,721 answered the first test question correctly.

Second, I also re-estimate the RIRT model for those who either answered the first question correctly or the follow-up question correctly. In other words, it includes the respondents who answered the initial question incorrectly but answered the follow-up question correctly. 5,753 respondents are in this group for the first wave, and 3,560 respondents are in this group for the

second wave.

Figure 4.7 – Comparison of RIRT Estimates across Subsamples



Note: This figure shows scatterplots and pairwise correlations between individual-level estimates of respondents' sensitive trait from *Wave 1* and *Wave 2*. It breaks down this analysis based on whether respondents correctly answered the test questions conducted in one or both waves. I find a strong positive correlation between the two groups of estimates.

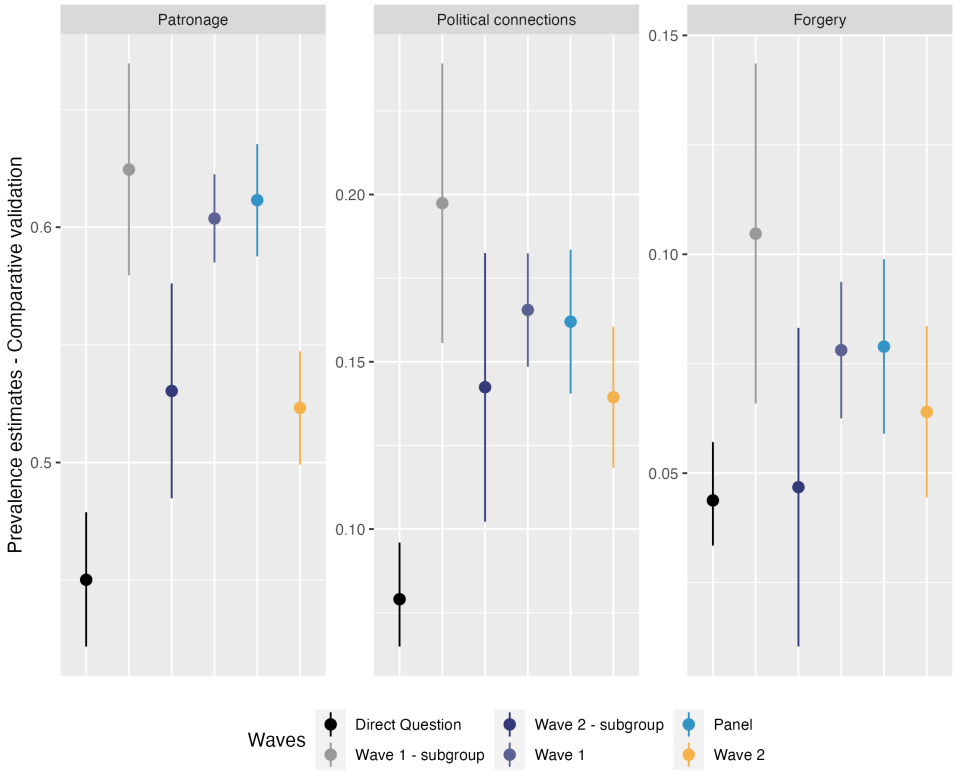
Figure 4.7 shows the density of each set of individual-level estimates, pairwise scatter plots and pairwise correlations among the estimates. There are several notable findings. First, there is hardly any difference between the individual-level estimates obtained from the total sample and the estimates from either sub-samples. The scatter plots show that almost all of the estimates align on a straight line, $y = x$, and the pairwise correlations are ~ 0.99. It indicates that the individual-level estimates are reliable and are not very sensitive to respondents' compliance.

The pairwise comparisons between the two waves are also consistent. The correlations are ~ 0.2, and the scatter plots also show consistent shape across all pairs. Furthermore, the shape of density plots is also very similar across different estimates. The estimates for the majority are pretty low, indicating that the overall prevalence rate of corrupt behaviours is low. Still, the distribution is skewed to the right, revealing a small number of cases where the estimated latent trait of corruption is high.

4.7 Comparison with direct question format from opt-in and representative samples

As pointed out, I conducted a third survey wave where 1,163 randomly selected respondents from the first wave. In this survey, respondents were asked to answer 3 out of 11 sensitive items in a direct questioning format. Figure 4.8 compares the results for all three sensitive items, and we can see that the RIRT yields higher prevalence levels. However, in the case of 'Forgery', I did not find a statistically significant difference between the two questioning formats only once we compared these estimates to the second wave. Still, there were substantial differences for the 'Patronage' and 'Political connections. Given that there are changes in the sample composition between the RIRT waves and the third direct question wave, I provide RIRT estimates from only those respondents who took part in the third wave. We can see that the results still show significant differences in prevalence estimates, with the exception of 'Forgery' item.

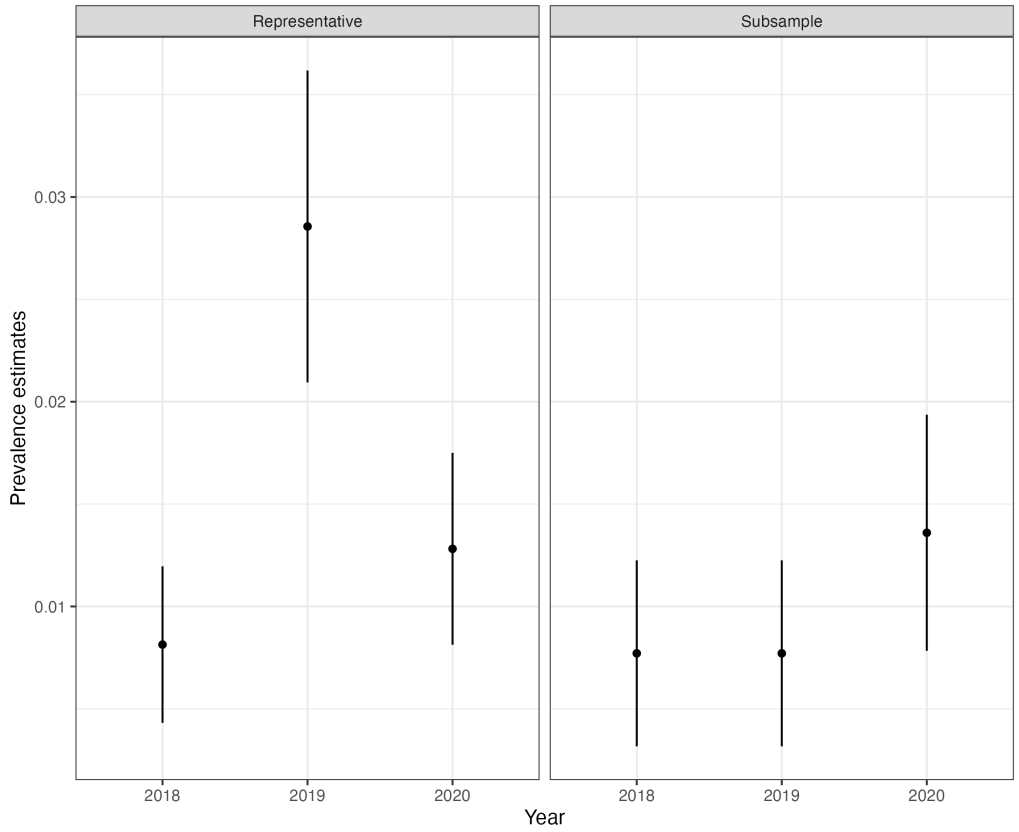
Figure 4.8 – Prevalence estimates RIRT versus Direct Questioning



Note: This figure reports estimated prevalence rates from *Wave 1*, *Panel*, *Wave 2*, and the last wave *Wave 3*. *Wave 3* included all items asked in a direct question format. I report the direct question wave in this plot as *Direct Question*. *Wave* included only three items: 'Forgery, 'Patronage', and 'Political connections', which corresponded to the three items with the highest prevalence rates on *Wave 1*. All estimates reported are naive and unweighted estimates.

Prevalence estimates from representative surveys I complemented this analysis by looking at the prevalence rates of 'Bribes' from an annual survey commissioned by the *Chilean Consejo para la Transparencia* (para la Transparencia, 2018, 2019, 2020). The Consejo para la Transparencia is an independent, non-partisan governmental organisation that seeks to promote transparency in Chile. This survey is a representative survey containing the same 'Bribe' item question in the RIRT but in a direct question format. The estimates displayed in Figure 4.9 are significantly lower than the bribe estimates obtained using RIRT. I provide two estimates from this sample: 1) Weighted estimates from the whole sample; 2) Weighted estimates for a subset that only contains respondents from the same communes included in the RIRT. I find that the estimated proportion of bribes since 2018 has been around 0.008 and 0.01.

Figure 4.9 – Prevalence Estimates - Bribes Item



Note: This figure shows the prevalence estimates of the 'Bribe' item from a representative survey conducted annually in Chile. I find that under 1% of respondents claimed to pay a bribe to a local official. Estimates are weighted to be nationally representative. The sample size by year: 2018 (n = 2,860), 2018 (n = 2,850) and 2020 (2,900). For the sub-sample, the sample sizes as follow 2018 (n = 2,100), 2019 (n = 1980) and 2020 (n = 2,170).

From this analysis, I find that prevalence rates from the representative survey are considerably lower than those obtained from RIRT. Figure 4.10 summarises the weighted estimates from the representative sample (and subsample) and all waves of the RIRT.

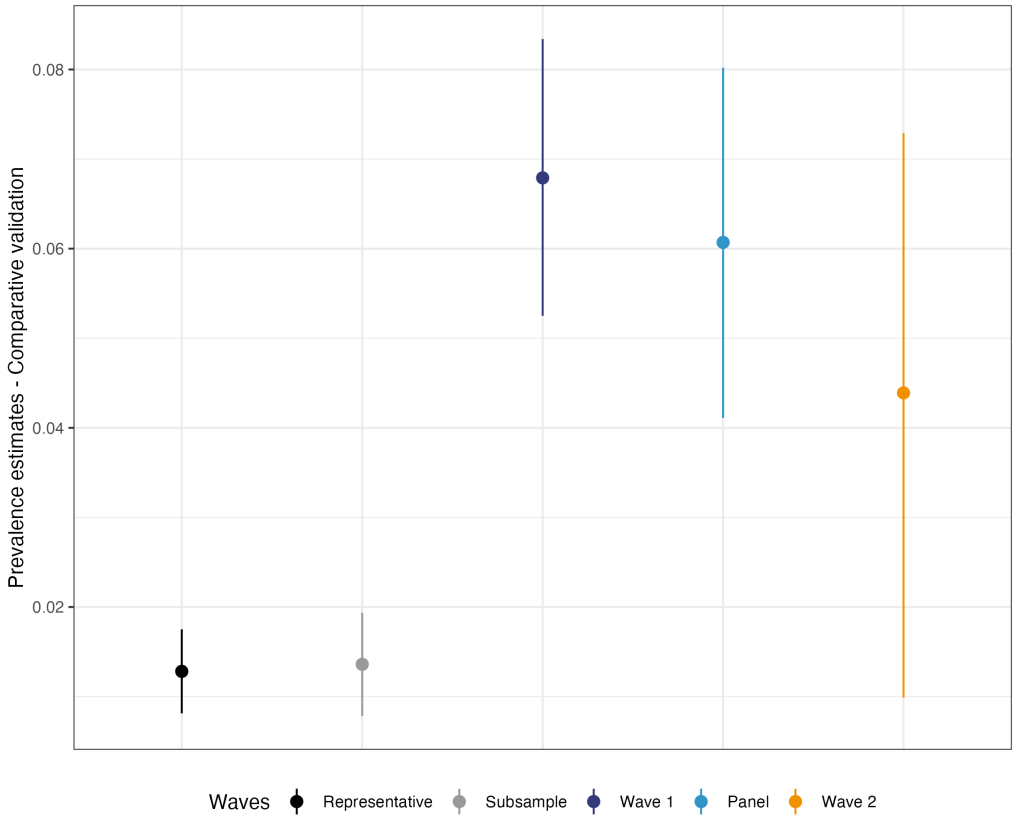


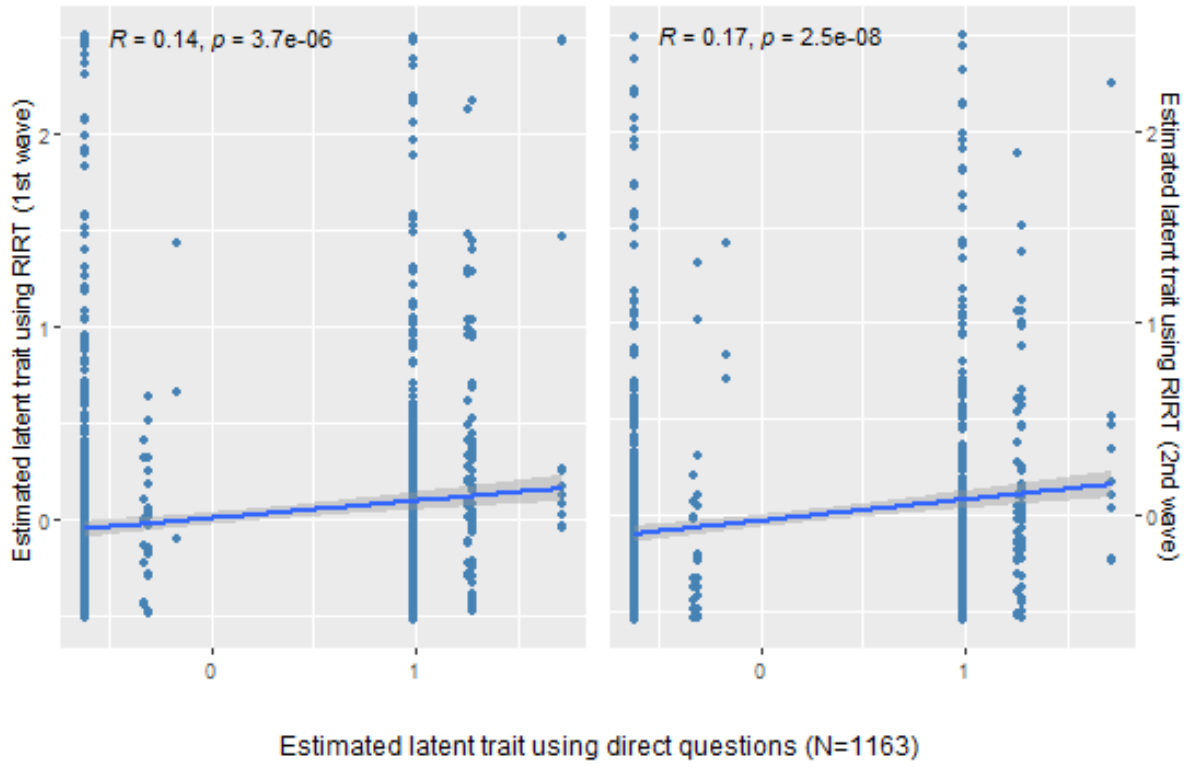
Figure 4.10 – Prevalence estimates - Bribes item RIRT and representative survey

Note: This figure shows the prevalence estimates of the 'Bribe' item from a representative survey conducted annually in Chile. From this analysis, I find that under 1% of respondents claimed to pay a bribe to a local official. Estimates are weighted to be nationally representative. The sample size by year: 2018 (n = 2,860), 2018 (n = 2,850) and 2020 (2,900). For the sub-sample, the sample sizes are as follows: 2018 (n = 2,100), 2019 (n = 1980) and 2020 (n = 2,170). The estimates from the RIRT for *Wave 1* and *Panel* are weighted estimates. *Wave 2* are weighted and bias-corrected estimates.

Individual-level estimates comparison Figure 5 illustrates a comparison of individual-level RIRT estimates to IRT estimates. The IRT estimates are obtained from the three items using direct questioning. I use R's *ltm* package to run the IRT model. The x-axis of Figure 5 is the individual-level estimates of corruption behaviour using direct questioning. The y-axes are the individual-level estimates of corruption behaviour using Crosswise RR questioning. The fitted line and Pearson's correlation coefficient on each plot describe the statistical association between each pair of estimates.

Due to the small number of questions, the IRT estimates' variability is limited. However, I could still find a positive correlation between the IRT and RIRT estimates, validating the overall consistency of RIRT estimates.

Figure 4.11 – Comparison of Individual-level Estimates between IRT and RIRT



Note: This figure depicts the relationship of respondents' underlying trait for *Wave 1* and *Wave 2* and subjects' underlying trait computed from the three direct questions used in the *Wave 3*. I observe a positive correlation between these two distributions. One corresponds to respondents that answered 'Yes' to at least one of the sensitive questions and '0' to those who responded 'No' to all questions.

I extended the previous analysis by looking at the relationship between respondents' probability of engaging in the corrupt behaviour (based on their underlying trait obtained in *Wave 1* and their answers to the direct questioning format in *Wave 3*. I depict this relationship in Figure 4.12. On the x-axis is the respondent's answer to the sensitive question in a direct question format, and the y-axis is the probability of individuals engaging in each sensitive behaviour computed based on the respondents' underlying trait. We can observe that for the 'Patronage' item for respondents that answered yes (n = 523), most also yielded a higher probability of engaging in that action. However, we also can see that respondents who answered 'No' also obtained a high likelihood. This result may occur due to the low discrimination power of this item. In the case of 'Political connections', most of the 92 respondents claimed to engage in this behaviour yielded a higher probability of committing this behaviour. Nevertheless, we also observe that most respondents that answered 'No' also obtained a higher probability. Finally, for the case of forgery, RIRT accurately predicted a higher probability for most of the 51 respondents who reported engaging in this behaviour. Again, we observe that RIRT clusters respondents into two groups for those that answered no: 1) A group with a high probability near o equal to 1. 2) A group clustered around a 0.6 probability.

Figure 4.12 – Comparison of Direct Question answer and Probability of Engaging on the Sensitive Behaviour



Note: This figure shows the distribution of the probabilities of engaging in sensitive behaviour for the 'Patronage', 'Political connections', and 'Forgery' items. On the x-axis is whether respondents answered 'Yes', 'No' or 'Prefer not to say to the same question, but in a direct question format in *Wave 3*. The probability of engaging in the sensitive behaviour was computed using respondents' underlying trait, discrimination and difficulty parameter.

4.8 Discussion

The results shows that we can leverage RIRT to estimate individual-level prevalence rates of the propensity of each respondent to engage in the sensitive behaviour in question. These estimations can be crucial not only to obtain prevalence rates in a particular population of interest, but they could also play an important role as an individual-level predictor whenever researchers are attention checks and investigating topics related to corruption.

Randomised Response technique and its extensions, such as RIRT, may be a promising method to minimise social desirability bias. However, it is still unclear how RIRT's additional burden on respondents may prone them to comply less with the instructions or to answer incorrectly. To address this, I provided bias-corrected estimates at the population-level and subsample analyses at the individual-level. From the evidence gathered of the different strategies to measure attentiveness I find encouraging, but somewhat puzzling findings. In on end, There

are around 14% of respondents failed to understand the instructions in the first round of test included before the surveys[11]. Meantime, I found strong evidence of high attentiveness and comprehension obtained from the bias-corrected estimates by incorporating an 'Anchor item'. In the case, of nested items, the results also signal high level of attentiveness, there are 3 out the 4 nested item which yielded lower level of prevalence rates to their respective core items.

As indicated before, it is important to include sample weights to calibrate the estimations based on the proportions of different groups or populations. I tested various weighting schemes to make accurate inferences about prevalence rates in the Chile population, and I shown that the results are robust different weighting schemes. One particular that I did not exploit in this essay is that we can use individual-level estimates of corrupt behaviour to produce predictions at geographical smaller areas of interest using techniques such as Multilevel Regression and Post-Stratification.

One of most prevalent criticisms of all indirect methods techniques, is the over-reliance on what some call the "more-is-better assumption" approach. Such an approach means that whichever technique yields higher prevalence rates is closer to the actual prevalence. According to Höglinger and Jann (2018), most RR studies use the "the more is better" validation strategy to evaluate how the different RR variants perform. They also found that different RR variants can capture around 44% to 71% of the actual levels of prevalence of the sensitive behaviour. At the same time, Wolter and Preisendörfer (2013) found slightly more conservative results, with barely any difference in the proportion of respondents that answered truthfully to a RR question versus a direct question format[12]. Hence, the results of studies that use the "more is better" testing approach must be critically evaluated and taken with a grain of salt.

The results show that RIRT yields higher prevalence rates at the aggregate level than the direct questioning format, at least for the three items evaluated in the direct questioning wave (third wave). This is in line with previous findings where the RR obtains higher prevalence rates once compared to the direct questioning format. Although these results appear promising, I do not make strong claims on whether the RIRT yields estimates closer to the unobserved true prevalence rate.

Once I examine prevalence rates from the RIRT to rates obtained from a representative sample survey, I also find significant differences between these questioning formats. It is important to stress that several factors may explain this difference—for example, different sample composition to weighting schemes. Furthermore, I compare one item, which limits the possibility of making claims on whether they are systematic differences between these survey formats.

Finally, once we compare the propensity to engage in sensitive behaviour to respondents' answers from a direct questioning format survey, I find some evidence that RIRT accurately predicts their self-reported behaviour. We observe that RIRT yields a higher probability to those respondents that answered 'Yes' to the items. However, the distribution of the probability values for respondents who answered 'No' is puzzling. RIRT tends to cluster the probabilities

[11] In both waves.

[12] It is important to state that both studies relied on official records at the individual level to estimate the true prevalence, making their validation approach very robust.

in very high or very low values of the probability distribution for the 'Political connections' and 'Forgery' items. Furthermore, it yields high probabilities values near one for a considerable number of respondents, which raises questions about the accuracy of this technique.

In summary, the findings of this study support the claim that attentiveness and compliance are high for this crosswise design. Furthermore, it shows that prevalence rates using indirect questioning techniques tend to be higher than estimates obtained from direct survey surveys. However, more research is needed to examine the accuracy and reliability of individual-level estimates, specifically in validation studies where prevalence estimates are compared to observed behaviour or official records.

Appendix

Descriptives

Table 4.5 reports the means of relevant variables for all three waves. I observe that, across all waves, most point estimates remain similar, except income, which drops in the direct questioning wave (DC).

Table 4.5 – Means - all waves

	Survey Wave			
	Wave 1	*Panel*	*Wave 2*	*Wave 3*
	RIRT	*RIRT*	*RIRT*	*DC*
Covariates				
Female	65%	66%	66%	67%
Education (Yrs)	15	15	15	15.2
Age	35.9	35.7	35.7	35.9
Partisanship	4.48	4.46	4.46	4.37
Income (USD)	529	506	506	470
Sample	6050	3723	3723	1163

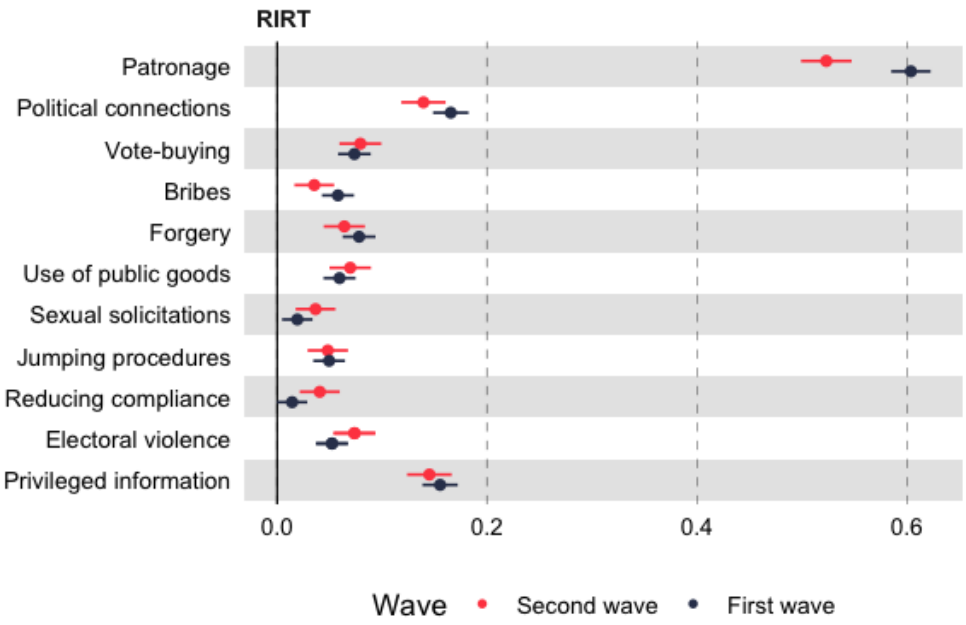
Prevalence estimates

Table 4.6 – Prevalence Estimates Waves 1 and 2 and Panel

Items	Wave 1	Panel	Wave 2
Patronage	0.59	0.61	0.52
	(0.010)	(0.012)	(0.012)
Political connections	0.18	0.16	0.14
	(0.009)	(0.011)	(0.011)
Vote-buying	0.07	0.07	0.08
	(0.008)	(0.010)	(0.010)
Bribes	0.07	0.05	0.04
	(0.008)	(0.010)	(0.010)
Forgery	0.09	0.08	0.06
	(0.008)	(0.010)	(0.010)
Personal	0.05	0.06	0.07
	(0.008)	(0.010)	(0.010)
Sexual solicitation	0.02	0.03	0.04
	(0.007)	(0.010)	(0.010)
Lobbying	0.05	0.05	0.05
	(0.008)	(0.010)	(0.010)
Compliance	0.03	0.01	0.04
	(0.007)	(0.009)	(0.010)
Electoral Violence	0.06	0.05	0.07
	(0.008)	(0.010)	(0.010)
Privilege	0.15	0.16	0.15
	(0.009)	(0.011)	(0.011)

Note: This table reports naive prevalence estimates using post-stratification survey weights.

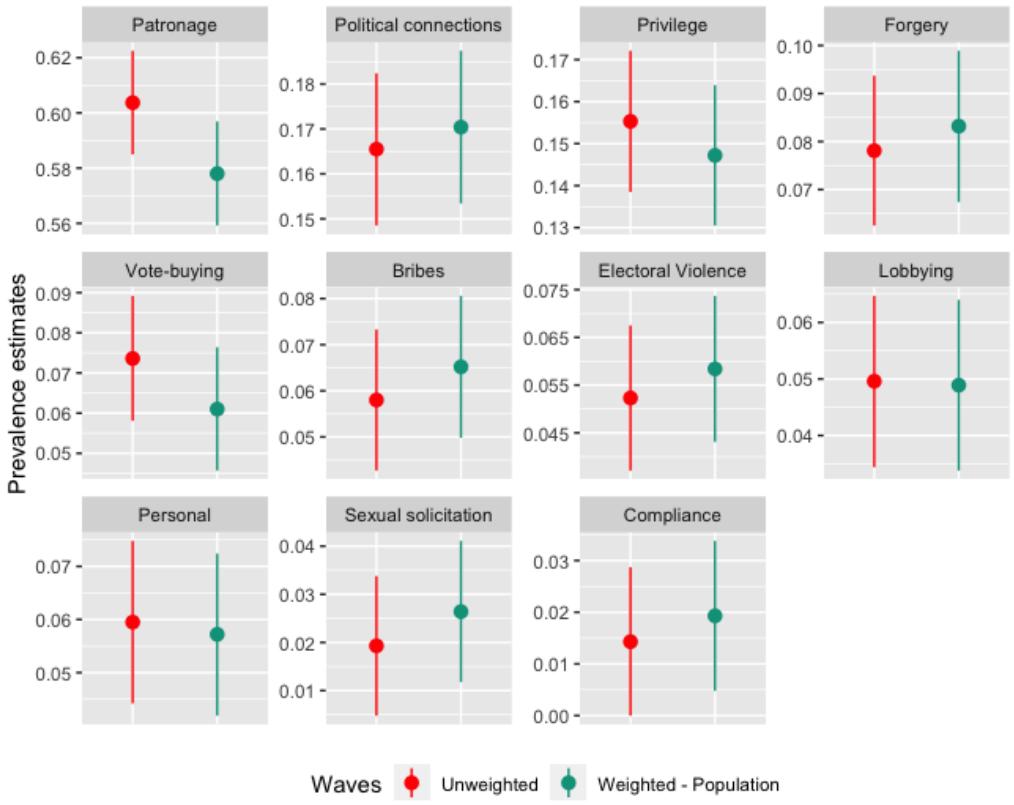
Figure 4.13 – Naive Prevalence Estimates using Univariate Estimation



Note: This figure reports the prevalence rates using a a modified univariate regression reported in the estimation section of this essay. This estimates are naive and unweighted estimates for all 11 items. Overall, I find no significant differences from the estimates obtained from the raw and logistic model.

Weighting Given that respondents were randomly selected from an unrepresentative opt-in sample, we included sample weights in our estimations using different sampling weights and weighting techniques. Figure 4.14 provides weighted prevalence estimates once you incorporate population weights, in addition to other demographic parameters such as age, gender, and education. Surprisingly, we observed six items yield higher prevalence estimates once compared to the naive unweighted estimates.

Figure 4.14 – Naive Weighted IPW Prevalence Estimates



Note: This figure summarises Naive prevalence rates from unweighted and weighted estimates. Weighted estimates are weighted using Inverse Probability Weights based on population. We report these estimates for all 11 items. We can identify that prevalence rates somewhat changed once we introduced weights. We observe that most prevalence estimates increased. For example, 'Political connections', 'Bribes', 'Sexual solicitation', 'Forgery' and 'Compliance'.

Table 4.7 – Proportion of Underlying Trait Dropping two Selected Items

Wave	$< 0.5\sigma$	$>= 0.5\sigma$
Wave 1	0.962	0.038
Wave 2	0.923	0.077

Table 4.8 – Difference between Core and Nested Items

Items	Core			Nested			
Unweighted - Naive	Prevalence	Lower C.I	Upper C.I	Prevalence	Lower C.I	Upper C.I	Difference
Patronage	0.523	0.499	0.547	0.173	0.151	0.195	0.350***
Bribes	0.035	0.016	0.054	0.049	0.029	0.068	-0.013
Forgery	0.064	0.044	0.084	0.039	0.020	0.059	0.025
Privilege	0.145	0.124	0.166	0.106	0.086	0.127	0.039
Weighted - Bias-corrected							
Patronage	0.512	0.478	0.545	0.167	0.137	0.196	0.345
Forgery	0.070	0.040	0.103	0.035	0.004	0.062	0.034
Privilege	0.144	0.114	0.173	0.103	0.070	0.134	0.041
Bribes	0.044	0.010	0.073	0.039	0.008	0.064	0.005

Note: This table summarises the prevalence rates for all four core and nested items included in Wave 2. I report the difference between the prevalence rates from core minus nested items. In this analysis, we conducted the differences between two group of population estimates. The difference between unweighted and naive. The second set of differences is between weighted and bias-corrected estimates. I only find differences for the 'Patronage' item $^{***}p < 0.001$.

Extension on population estimation strategies

In the following section, I provide more details on population-level estimates. I largely borrowed from (Heck and Moshagen, 2018).

For the logistic regression, the expected prevalence of the sensitive behaviour is given by the following expression :

$$\pi_i = \text{logit}^{-1}(\mathbf{X}_i\beta) = \frac{1}{1+\exp(-\mathbf{X}_i\beta)} \tag{4.28}$$

In this expression, β is the vector of regression coefficients and $\mathbf{X}_i$ is the i-th row of the design matrix $\mathbf{X}$. In the case of RR designs, the responses are binomially distributed with means given by the expected probabilities λ. λ is equal to the product of a misclassification matrix $\mathbf{P}$ and the $\boldsymbol{\pi}$ matrix that contains the true probabilities of the sensitive behaviour. The entries of P_{ij} of the $\mathbf{P}$ matrix are set by the conditional probabilities to response i (i.e. *Yes* or *no*) to the sensitive item given the true state of a respondent is j (i.e. *Engaged in corrupt behaviour* or *Did not engage in corrupt behaviour*). In the general case, the likelihood function is the following:

$$f(\boldsymbol{y}|\beta) = \prod_{i=1}^{N} \binom{n_i}{y_i} (P_{10}(1-\pi_i) + P_{11}\pi_i)^{y_i} (P_{00}(1-\pi_i) + P_{01}\pi_i)^{n_i - y_i} \tag{4.29}$$

In equation 4.29, y_i is the number of affirmative responses to the sensitive question, and n_i is the total responses. $P_{01}...P_{11}$ are the entries of the misclassification matrix where P_{00} is the conditional probability of answering *'no'* to the sensitive question, dependent on that the respondent has not engaged in corrupt behaviour. Conversely, P_{11} is the conditional probability of answering affirmatively, conditional on whether the respondent has engaged in corrupt conduct.

Using an expectation-maximisation (EM) algorithm, we can retrieve the maximum likelihood estimate of the parameters of interest. This algorithm serves cases with an incomplete-data problem. The EM algorithm works as follows: 1) define the distribution of the Z unobservable variable. 2) estimate the maximum likelihood with the complete data $L_c(\theta|Y,Z)$ as both X and Y variables were observed. 3) The third step is to set a starting or initial value for the parameter to be estimated θ_0 4) estimate the expectation of the likelihood function.

$$Q(\theta,\theta_0) = E[\ln L_c(\theta|Y,Z,\theta_0)] \tag{4.30}$$

This algorithm proceeds with the maximisation step after obtaining the expectation of the starting value and calculating the current conditional value of the parameter of interest. This is stated formally in equation 4.31:

$$\max_{\theta} Q(\theta, \theta_1) \tag{4.31}$$

This algorithm continues in this iterative process of calculating the expected value for a given value of θ and find the value that maximises the value of this parameter until it converges.

We can extend the modified univariate logistic regression to include predictors of experiencing the sensitive behaviour. Ben, Jerke and Krumpal (2012) and van den Hout, van der Heijden and Gilchrist (2007), Hsieh and Perri (2020) formalise this modified version, where $\pi = P(X = 1|Z)$ is the unknown probability of answering 'yes' to the sensitive item, conditional on covariates Z.

Linear estimation. The linear estimation approach for RR can be described as a misclassification scheme that adds random noise to the true responses Heck and Moshagen (2018). The likelihood function of this linear RR regression model contains two elements. 1) The likelihood of the unobservable true states is a function of the mixture distribution of the observed responses and the missclasification matrix $\boldsymbol{P}$. 2) n_j^* are the observed frequencies of the $j = 1....J$ possible RR responses. We can express this likelihood function as follows:

$$log f_1(\boldsymbol{y}|\boldsymbol{\pi}) = \sum_{k=1}^{J} n_j^* \sum_{k=1}^{J} P_{jk}\pi_k + C \tag{4.32}$$

The second element is the dependent variable $\boldsymbol{y}$ (the number of affirmative responses to the sensitive question), which is conditionally normally distributed with standard deviation σ. Then, the expected value is a function of the *true*, *latent* states on the predictors and the observed value for the observed variables. As we do not observe the *true* states, the marginal likelihood of the regression coefficient β has to be computed to reduce the uncertainty induced by the misclassification matrix.

$$log f_2(\boldsymbol{y}|\beta, \sigma, \boldsymbol{\pi}) = \sum_{i=1}^{N} log \left[\sum_{j=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{y_i - \boldsymbol{X}_i^j}{\sigma})^2) \frac{P_{(i)j}\pi_j}{\sum_{k=1}^{J} P_{(i)k}\pi_k} \right] \tag{4.33}$$

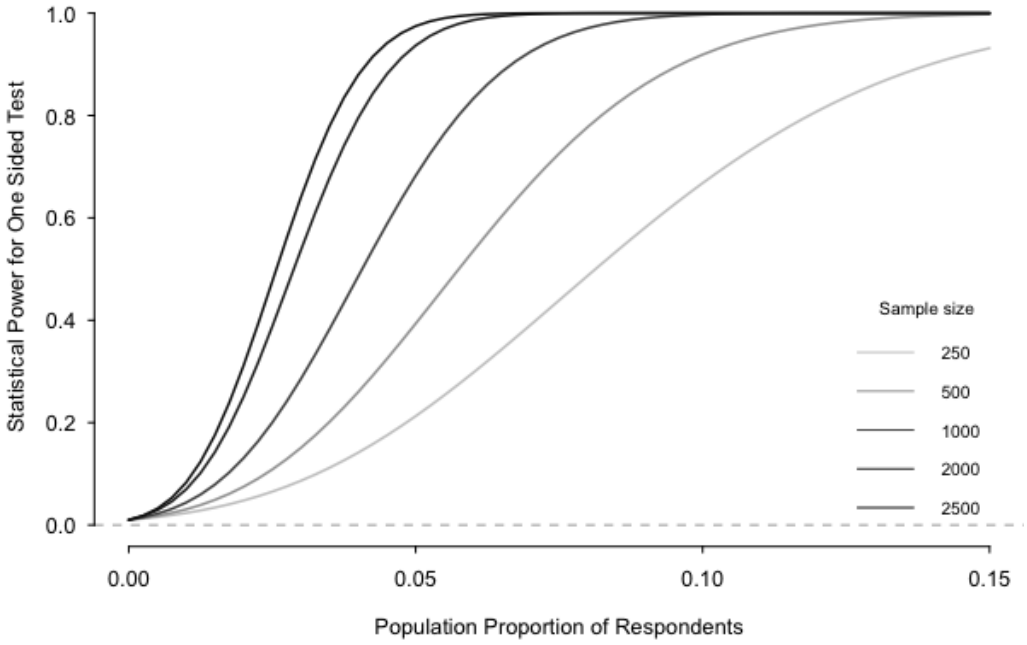
Combining equations 4.32 and 4.33, we can obtain the joint loglikelihood of the whole model.

By maximising this function, we obtain the maximum likelihood of the parameter of interest.

Power Calculations

Power calculations were conducted previous to this study. For population prevalence estimates above 2-3%, all the RIRT waves (first and second waves) are well-powered above 80%. For low-prevalence items, such as sexual solicitations, both the first and second waves would be under-powered.

Figure 4.15 – Power calculations Randomised Response



Research questions and hypotheses

Research questions As I mentioned, researchers should seek to minimise the effects of social desirability bias. Tackling this issue is particularly important for studies that rely on survey-based measurements of a sensitive issue or behaviour. Thus, one of the goals of this study is to measure the prevalence of corrupt behaviour at the population level using RIRT.

- *RQ 1a: What the prevalence estimates of corrupt behaviour at the population level?*

Most of the literature on RR design has focused on validating this method against other questioning formats. However, no study has investigated whether this technique provides robust estimates. Thus, this question seeks to explore further on this area:

- *RQ 2: Does the RR method provide robust estimates of the prevalence of the sensitive behaviour and subjects' underlying trait?*

One major drawback of RR designs is the additional cognitive burden and response times on respondents, which may reduce compliance and attentiveness (Atsusaka and Stevenson, 2021). Thus, the contribution of this paper is to provide new empirical evidence on the magnitude of this bias and to provide bias-corrected estimates of sensitive behaviour.

- *RQ 3: Does bias due to inattention lead to significantly larger estimates of the sensitive behaviour?*

Hypotheses Based on recent survey results (LAPOP, 2021, 2019, 2017; para la Transparencia, 2020, 2019, 2018), we should expect to obtain positive prevalence estimates of corrupt behaviours in Chile. I do not have a strong prior about the magnitude of this prevalence estimates, as these estimates vary from one survey to another. Furthermore, there are corrupt behaviours included in this study that have not been measured before.

- *H1: RIRT will provide positive prevalence estimates of corrupt behaviours.*

It is reasonable to expect that the lack of comprehension of the RR questioning procedures would yield biased estimates of the prevalence of the behaviour of interest (Hoffmann et al., 2017). According to Atsusaka and Stevenson (2021), naive group-level prevalence estimates

that ignore respondents' inattentiveness always overestimate the population prevalence under certain regularity conditions. The conditions are that the population prevalence (π) and the randomisation probability (p) are less than 0.5. Because my analysis shows these conditions met, I explore if naive RR estimates are significantly larger than the bias-corrected estimates at the population level—the inattention bias increases as the per cent of inattentive survey respondents increases.

- *H2: Naive estimates of the sensitive behaviours would yield higher prevalence estimates than bias-corrected estimates with inattentive respondents.*

As pointed out before, no studies focus on whether this technique yields robust estimates of the prevalence of the sensitive behaviours in question. However, there is also no systematic evidence or theoretical grounds that would suggest that this method would yield noisy estimates of the behaviours of interest. Thus, my starting point is that this technique should produce robust estimates of the corrupt behaviours at the population and individual levels.

- *H3: RIRT will provide robust estimates of the prevalence of corrupt behaviours.*

Bibliography

Aarslew, Laurits F. 2022. "Why Don't Partisans Sanction Electoral Malpractice?" *British Journal of Political Science* pp. 1–17.

Adida, Claire, Jessica Gottlieb, Eric Kramon and Gwyneth McClendon. 2022. "Breaking the Clientelistic Voting Equilibrium: The Joint Importance of Salience and Coordination." *Quarterly Journal of Political Science* .

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

Ahlquist, John S. 2018. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis* 26(1):34–53.

Ajzenman, Nicolás. 2021. "The Power of Example: Corruption Spurs Corruption." *American Economic Journal: Applied Economics* 13(2):230–57.

Alesina, Alberto, Armando Miano and Stefanie Stantcheva. 2018. Immigration and Redistribution. Working Paper 24733 National Bureau of Economic Research.

Alesina, Alberto, Stefanie Stantcheva and Edoardo Teso. 2018. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108(2):521–54.
URL: *http://www.aeaweb.org/articles?id=10.1257/aer.20162015*

Anderson, Kaitlin, Gema Zamarro, Jennifer Steele and Trey Miller. 2021. "Comparing Performance of Methods to Deal With Differential Attrition in Randomized Experimental Evaluations." *Evaluation Review* 45(1-2):70–104.

Anduiza, Eva, Aina Gallego and Jordi Muñoz. 2013. "Turning a Blind Eye: Experimental Evidence of Partisan Bias in Attitudes Toward Corruption." *Comparative Political Studies* 46(12):1664–1692.

Angrist, Joshua D. and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90(430):431–442.

Aoyagi, Masaki, Rm Guillaume Fréchette and Sevgi Yuksel. 2021. "Beliefs in Repeated Games." No notes.

Arel-Bundock, Vincent, André Blais and Ruth Dassonneville. 2021. "Do Voters Benchmark Economic Performance?" *British Journal of Political Science* 51(1):437–449.

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*

Armand, Alex, Alexander Coutts, Pedro C. Vicente and Inês Vilela. 2020. "Does Information Break the Political Resource Curse? Experimental Evidence from Mozambique." *American Economic Review* 110(11):3431–53.

Atsusaka, Yuki and Randolph T. Stevenson. 2021. "A Bias-Corrected Estimator for the Crosswise Model with Inattentive Respondents." *Political Analysis* pp. 1–15.

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*

Baker, Frank B. and Seock-Ho Kim. 2017. *Test Calibration*. Cham: Springer International Publishing pp. 105–125.
URL: *https://doi.org/10.1007/978-3-319-54205-8*$_7$

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

Ben, Jann, Julia Jerke and Ivan Krumpal. 2012. "Asking Sensitive Questions using the Crosswise Model: An Experimental Survey Measuring Plagiarism." *American Association for Public Opinion Research* 76(1):32–49.

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

Berliner, Daniel and Joachim Wehner. 2022. "Audits for Accountability: Evidence from Municipal By-Elections in South Africa." *The Journal of Politics* 0(ja):null.

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*

Blair, Graeme, Kosuke Imai and Jason Lyall. 2014. "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan." *American Journal of Political Science* 58(4):1043–1063.
URL: *http://www.jstor.org/stable/24363542*

Blair, Graeme, Kosuke Imai and Yang-Yang Zhou. 2015. "Design and Analysis of the Randomized Response Technique." *Journal of the American Statistical Association* 110(511):1304–1319.
URL: *https://doi.org/10.1080/01621459.2015.1050028*

Blair, Graeme, Winston Chou and Kosuke Imai. 2019. "List Experiments with Measurement Error." *Political Analysis* pp. 1–26.

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.

Bonica, Adam. 2018. "Are Donation-Based Measures of Ideology Valid Predictors of Individual-Level Policy Preferences?" *The Journal of Politics* 81:000–000.

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.

Bourke, Patrick D. and Michael A. Moran. 1988. "Estimating Proportions from Randomized Response Data Using the EM Algorithm." *Journal of the American Statistical Association* 83(404):964–968.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478685*

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

Broockman, David E., Joshua L. Kalla and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." *Political Analysis* 25(4):435?464.

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*

Campello, Daniela and Cesar Zucco. 2016. "Presidential Success and the World Economy." *The Journal of Politics* 78(2):589–602.

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.

Cavallo, Alberto, Guillermo Cruces and Ricardo Perez-Truglia. 2017. "Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments." *American Economic Journal: Macroeconomics* 9(3):1–35.

Chang, Eric C. C., Miriam A. Golden and Seth J. Hill. 2010. "Legislative Malfeasance and Political Accountability." *World Politics* 62(2):177–220.

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.

Clifford, Scott, Geoffrey Sheagley and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* pp. 1–18.

Coibion, Olivier, Dimitris Georgarakos, Yuriy Gorodnichenko and Maarten van Rooij. 2019. How Does Consumption Respond to News about Inflation? Field Evidence from a Randomized Control Trial. Working Paper 26106 National Bureau of Economic Research.

Coppock, Alexander, Seth J. Hill and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6(36).

Corbacho, Ana, Daniel W. Gingerich, Virginia Oliveros and Mauricio Ruiz-Vega. 2016. "Corruption as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica." *American Journal of Political Science* 60(4):1077–1092.

Costas-Pérez, Elena, Albert Solé-Ollé and Pilar Sorribas-Navarro. 2012. "Corruption scandals, voter information, and accountability." *European Journal of Political Economy* 28(4):469–484.

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.

Cruces, G., R. Perez-Truglia and M. Tetaz. 2013. "Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment." *Journal of Public Economics* 98:100–112.

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

de la Transparencia, Consejo. 2019. "Estudio Nacional de Funcionarios Públicos.".

DellaVigna, Stefano and Matthew Gentzkow. 2010. "Persuasion: Empirical Evidence." *Annual Review of Economics* 2(1):643–669.

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

Di Tella, Rafael and Ignacio Franceschelli. 2011. "Government Advertising and Media Coverage of Corruption Scandals." *American Economic Journal: Applied Economics* 3(4):119–51.
URL: *https://www.aeaweb.org/articles?id=10.1257/app.3.4.119*

Druckman, James N. 2022. "A Framework for the Study of Persuasion." *Annual Review of Political Science* 25(1):65–88.
URL: *https://doi.org/10.1146/annurev-polisci-051120-110428*

Duch, Raymond, Denise Laroze, Thomas Robinson and Pablo Beramendi. 2020. "Multi-modes for Detecting Experimental Measurement Error." *Political Analysis* pp. 1–21.

Duch, Raymond M, Denise Laroze, Constantin Reinprecht and Thomas S Robinson. 2020. "Nativist policy: the comparative effects of Trumpian politics on migration decisions." *Political Science Research and Methods* pp. 1–17.

Duch, Raymond M. and Randy Stevenson. 2008. *The Economic Vote: How Political and Economic Institutions Condition Election Results*. Cambridge: Cambridge University Press.

Duch, Raymond, Paulina Granados, Denise Laroze, Mauricio Lopez, Marian Ormeño and Ximena Quintanilla. 2020. What works? An experiment on highlighting the losses in provider selection. Working paper Nuffield College Centre for Experimental Social Sciences.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, Gareth Nellis, Claire L. Adida, Eric Arias, Clara Bicalho, Taylor C. Boas, Mark T. Buntaine, Simon Chauchard, Anirvan Chowdhury, Jessica Gottlieb, F. Daniel Hidalgo, Marcus Holmlund, Ryan Jablonski, Eric Kramon, Horacio Larreguy, Malte Lierl, John Marshall, Gwyneth McClendon, Marcus A. Melo, Daniel L. Nielson, Paula M. Pickering, Melina R. Platas, Pablo Querubín, Pia Raffler and Neelanjan Sircar. 2019. "Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials." *Science Advances* 5(7).

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. 2014. "Partisanship and Electoral Accountability: Evidence from the UK Expenses Scandal." *Quarterly Journal of Political Science* 9(4):441–472.

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.

Engel, Christoph and Oliver Kirchkamp. 2018. "Measurement Errors of Risk Aversion and How to Correct Them." Working Paper.

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.

Fergusson, Leopoldo, Carlos Molina and Juan Riano. 2017. "I Sell My Vote and So What? A New Database and Evidence from Colombia.".

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.

Finan, Frederico and Maurizio Mazzocco. 2021. Combating Political Corruption with Policy Bundles. Working Paper 28683 National Bureau of Economic Research.

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

Fisher, Robert J. 1993. "Social Desirability Bias and the Validity of Indirect Questioning." *Journal of Consumer Research* 20(2):303–315.
URL: *https://doi.org/10.1086/209351*

Fishman, Ray and Miriam A. Golden. 2017. *Corruption: What Everybody Needs to Know.* Oxford: Oxford University Press.

Fong, Christian and Justin Grimmer. 2022. "Causal Inference with Latent Treatments." *American Journal of Political Science* n/a(n/a).

Fornasari, Margherita, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, Kay Mukhtarova, Roger Daniel and Zahid Hasnain. 2022. "Global Survey of Public Servants Remote Work Surveys Data Set.".

Fox, J.-P. and Cheryl Wyrick. 2008. "A Mixed Effects Randomized Item Response Model." *Journal of Educational and Behavioral Statistics* 33(4):389–415.
URL: *https://doi.org/10.3102/1076998607306451*

Fox, J.-P., M. Avetisyan and J. van der Palen. 2013. "Mixture randomized item-response modeling: a smoking behavior validation study." *Statistics in Medicine* 32(27):4821–4837.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5859*

Fox, Jean-Paul. 2005. "Randomized Item Response Theory Models." *Journal of Educational and Behavioral Statistics* 30(2):189–212.

Fox, Jean-Paul, Duco Veen and Konrad Klotzke. 2019. "Generalized linear mixed models for randomized responses." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 15.

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression.* Thousand Oaks: Sage Publications.

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.

Freeder, Sean. 2020. It's No Longer the Economy, Stupid: Selective Perception and Attribution of Economic Outcomes. Working paper.

Fu, Chao. 2014. "Equilibrium Tuition, Applications, Admissions, and Enrollment in the College Market." *Journal of Political Economy* 122(2):225–281.

Gillen, Ben, Erik Snowberg and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 127(4):1826–1863.

Gingerich, Daniel W. 2010. "Understanding Off-the-Books Politics: Conducting Inference on the Determinants of Sensitive Behavior with Randomized Response Surveys." *Political Analysis* 18(3):349–38.

Gingerich, Daniel W., Virginia Oliveros, Ana Corbacho and Mauricio Ruiz-Vega. 2016. "When to Protect? Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior." *Political Analysis* 24(2):132–156.

Gonzalez-Ocantos, Ezequiel, Chad Kiewiet de Jonge, Carlos Melendez, Javier Osorio and David W. Nickerson. 2012. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56(1):202–217.

Gonzalez-Ocantos, Ezequiel, Chad Kiewiet de Jonge and David W. Nickerson. 2014. "The Conditionality of Vote Buying Norms: Experimental Evidence from Latin America." *American Journal of Political Science* 58(1):197–211.

Gottlieb, Jessica. 2016. "Greater Expectations: A Field Experiment to Improve Accountability in Mali." *American Journal of Political Science* 60(1):143–157.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Grigorieff, Alexix, Christopher Roth and Diego Ubfal. 2020. "Does Information Change Attitudes Toward Immigrants?" *Demography* 57:1117–1143.

Grossman, Guy and Kristin Michelitch. 2018. "Information Dissemination, Competitive Pressure, and Politician Performance between Elections: A Field Experiment in Uganda." *American Political Science Review* 112(2):280–301.

Haaland, Ingar, Christopher Roth and Johannes Wohlfart. 2020. "Designing Information Provision Experiments.". CESifo Working Paper No. 8406.

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).

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27(2):163–192.

Hambleton, Ronald K. and Hariharan Swaminathan. 1985. *Item Banking*. Dordrecht: Springer Netherlands pp. 255–279.
URL: *https://doi.org/10.1007/978-94-017-1988-9*$_1$2

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 Gabriel S. Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58(1):31–47.
URL: *http://dx.doi.org/10.1111/ajps.12053*

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

Heck, Daniel W. and Morten Moshagen. 2018. "RRreg: An R Package for Correlation and Regression Analyses of Randomized Response Data." *Journal of Statistical Software* 85(2):1–29.
URL: *https://www.jstatsoft.org/index.php/jss/article/view/v085i02*

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Hill, Seth J. 2017. "Learning Together Slowly: Bayesian Learning about Political Facts." *The Journal of Politics* 79(4):1403–1418.

Hoffman, Adrian, Birk Diedenhofen, Bruno Verschuere and Jochen Musch. 2015. "A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior." *Experimental Psychology* 62:403–414.

Hoffmann, Adrian, Berenike Waubert de Puiseau, Alexander F. Schmidt and Jochen Musch. 2017. "On the comprehensibility and perceived privacy protection of indirect questioning techniques." *Behavior Research Methods* 49(4):1470–1483.
URL: *https://doi.org/10.3758/s13428-016-0804-3*

Hoffmann, Adrian and Jochen Musch. 2016. "Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model." *Behavior Research Methods* 48(3):1032–1046.
URL: *https://doi.org/10.3758/s13428-015-0628-6*

Höglinger, M., B. Jann and A Diekmann. 2016. "Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model." *Survey Research Methods* 10(3):171–187.

Höglinger, Marc and Andreas Diekmann. 2017. "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT." *Political Analysis* 25(1):131–137.

Höglinger, Marc and Ben Jann. 2018. "More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model." *PLOS ONE* 13(8):1–22.
URL: *https://doi.org/10.1371/journal.pone.0201770*

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.

Hsieh, Shu-Hui and Pier Francesco Perri. 2020. "A Logistic Regression Extension for the Randomized Response Simple and Crossed Models: Theoretical Results and Empirical Evidence." *Sociological Methods & Research* 0(0).
URL: *https://doi.org/10.1177/0049124120914950*

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.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J. Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22(1):129–146.

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

Jerke, Julia, David Johann, Heiko Rauhut, Kathrin Thomas and Antonia Velicu. 2021. "Handle with Care: Implementation of the List Experiment and Crosswise Model in a Large-Scale Survey on Academic Misconduct.".

Jo, Myung-Soo. 2000. "Controlling social-desirability bias via method factors of direct and indirect questioning in structural equation models." *Psychology & Marketing* 17(2):137–148.

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.

Kayser, Mark Andreas and Michael Peress. 2012. "Benchmarking across Borders: Electoral Accountability and the Necessity of Comparison." *American Political Science Review* 106(3):661–684.

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.

Korndörfer, Martin, Ivar Krumpal and Stefan C. Schmukle. 2014. "Measuring and explaining tax evasion: Improving self-reports using the crosswise model." *Journal of Economic Psychology* 45:18–32.
URL: *https://www.sciencedirect.com/science/article/pii/S0167487014000609*

Kunda, Z. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3):480–498.

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

LAPOP. 2017. "Latin American Public Opinion Project.".

LAPOP. 2019. "Latin American Public Opinion Project.".

LAPOP. 2021. "Latin American Public Opinion Project.".

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.

Le Moglie, Marco and Gilberto Turati. 2019. "Electoral cycle bias in the media coverage of corruption news." *Journal of Economic Behavior Organization* 163:140–157.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268119301441*

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76(3):1071–102.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden and Cora J. M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods & Research* 33(3):319–348.
URL: *https://doi.org/10.1177/0049124104268664*

Lenz, Gabriel. 2013. *Follow the Leader: How Voters Respond to Politicians' Policies and Performance.* The University of Chicago Press.

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.

Liu, Haiyan and Zhiyong Zhang. 2017*a*. "Logistic regression with misclassification in binary outcome variables: a method and software." *Behaviormetrika* 44(2):447–476.
URL: *https://doi.org/10.1007/s41237-017-0031-y*

Liu, Haiyan and Zhiyong Zhang. 2017*b*. "Logistic regression with misclassification in binary outcome variables: a method and software." *Behaviormetrika* 44(2):447–476.
URL: *https://doi.org/10.1007/s41237-017-0031-y*

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.

Mian, Atif, Amir Sufi and Nasim Khoshkhou. 2021. "Partisan Bias, Economic Expectations, and Household Spending." *The Review of Economics and Statistics* pp. 1–46.

Muralidharan, Karthik, Mauricio Romero and Kaspar Wüthrich. 2019. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments.".

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*

OCDE. 2016. *Supreme Audit Institutions and Good Governance.*

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*

Oliveros, Virginia and Daniel W Gingerich. 2019. "Lying About Corruption in Surveys: Evidence from a Joint Response Model." *International Journal of Public Opinion Research* 32(2):384–395.
URL: *https://doi.org/10.1093/ijpor/edz019*

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?".

para la Transparencia, Consejo Nacional. 2018. "Estudio Nacional de Transparencia.".

para la Transparencia, Consejo Nacional. 2019. "Estudio Nacional de Transparencia.".

para la Transparencia, Consejo Nacional. 2020. "Estudio Nacional de Transparencia.".

Pérez, Efrén O. 2015. "Xenophobic Rhetoric and Its Political Effects on Immigrants and Their Co-Ethnics." *American Journal of Political Science* 59(3):549–564.

Peterson, Erik, Sean J. Westwood and Shanto Iyengar. 2021. *Beyond Attitudes: Incorporating Measures of Behavior in Survey Experiments.* Cambridge University Press.

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*

Ridgeway, Greg, Dan McCaffrey, Andrew Morral, Matthew Cefalu, Lane Burgette, Joseph Pane and Beth Ann Griffin. 2021. Toolkit for Weighting and Analysis of Nonequivalent Groups: A guide to the twang package. R package RAND.
URL: *https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf*

Rosenfeld, Bryn, Kosuke Imai and Jacob N. Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60(3):783–802.

Roth, Christopher and Johannes Wohlfart. 2020. "How Do Expectations about the Macroeconomy Affect Personal Expectations and Behavior?" *Review of Economics and Statistics* 102(4):731–748.

Roth, Christopher, Sonja Settele and Johannes Wohlfart. 2021. "Beliefs about public debt and the demand for government spending." *Journal of Econometrics* .

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*

Sagoe, Dominic, Maarten Cruyff, Owen Spendiff, Razieh Chegeni, Olivier de Hon, Martial Saugy, Peter G. M. van der Heijden and Andrea Petróczi. 2021. "Functionality of the Crosswise Model for Assessing Sensitive or Transgressive Behavior: A Systematic Review and Meta-Analysis." *Frontiers in Psychology* 12:2264.
URL: *https://www.frontiersin.org/article/10.3389/fpsyg.2021.655592*

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 Jaroslav Tir. 2022. "Threat-Inducing Violent Events Exacerbate Social Desirability Bias in Survey Responses." *American Journal of Political Science* n/a(n/a).
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12615*

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.".

Smallpage, Steven M., Adam M. Enders, Hugo Drochon and Joseph E. Uscinski. 2022. "The impact of social desirability bias on conspiracy belief measurement across cultures." *Political Science Research and Methods* 1(15):1–15.

Solaz, Hector, Catherine E. De Vries and Roosmarijn A. de Geus. 2019. "In-Group Loyalty and the Punishment of Corruption." *Comparative Political Studies* 52(6):896–926.

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.

Sovey, Allison J. and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

Stanig, Piero. 2015. "Regulation of Speech and Media Coverage of Corruption: An Empirical Analysis of the Mexican Press." *American Journal of Political Science* 59(1):175–193.

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*

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 5:859–83.

Transparente, Chile. 2019. Encuesta Nacional de Integridad. Technical report Chile Transparente.

van den Hout, Ardo, Peter G.M. van der Heijden and Robert Gilchrist. 2007. "The logistic regression model with response variables subject to randomized response." *Computational Statistics Data Analysis* 51(12):6060–6069.
URL: *https://www.sciencedirect.com/science/article/pii/S0167947306004774*

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*

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60(309):63–69. PMID: 12261830.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1965.10480775*

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*

Wolter, Felix and Peter Preisendörfer. 2013. "Asking Sensitive Questions: An Evaluation of the Randomized Response Technique Versus Direct Questioning Using Individual Validation Data." *Sociological Methods & Research* 42(3):321–353.
URL: *https://doi.org/10.1177/0049124113500474*

Wu, Ahra, R. Andrew Wood and Randolph T Stevenson. 2019. "A Validation Study of Individual-Level Survey Methodologies for Sensitive Questions.".

Wysmułek, Ilona. 2019. "Using public opinion surveys to evaluate corruption in Europe: trends in the corruption items of 21 international survey projects, 1989–2017." *Quality & Quantity* 53(5):2589–2610.

Yang, Frances M., Richard N. Jones, Sharon K. Inouye, Douglas Tommet, Paul K. Crane, James L. Rudolph, Long H. Ngo and Edward R. Marcantonio. 2013. "Selecting optimal screening items for delirium: an application of item response theory." *BMC Medical Research Methodology* 13(1):8.
URL: *https://doi.org/10.1186/1471-2288-13-8*

Yu, Jun-Wu, Guo-Liang Tian and Man-Lai Tang. 2008. "Two new models for survey sampling with sensitive characteristic: design and analysis." *Metrika* 67(3):251–263.
URL: *https://doi.org/10.1007/s00184-007-0131-x*

Yusaku Horiuchi Horiuchi, Yusaku, Zachary Markovich and Teppei Yamamoto. 2021. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis* 1(15).

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*

Zwane, Alix Peterson, Jonathan Zinman, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean S. Karlan, Richard Hornbeck, Xavier Giné, Esther Duflo, Florencia Devoto, Bruno Crepon and Abhijit Banerjee. 2011. "Being surveyed can change later behavior and related parameter estimates." *Proceedings of the National Academy of Sciences* 108(5):1821–1826.