

UNRAVELING THE PARADOX OF ANTICORRUPTION MESSAGING: EXPERIMENTAL EVIDENCE FROM A TAX ADMINISTRATION REFORM

Nicolas Ajzenman (McGill University)
Martín Ardanaz (Inter-American Development Bank)
Guillermo Cruces (Universidad de San Andrés-CONICET & University of Nottingham)
Germán Feierherd (Universidad de San Andrés)
Ignacio Lunghi (New York University & CEDLAS-IIE-UNLP)

Primera versión: Noviembre 2025
Esta versión: Noviembre 2025

Documento de Trabajo Nº 173
Departamento de Economía
Universidad de San Andrés

*Vito Dumas 284, B1644BID, Victoria,
San Fernando, Buenos Aires, Argentina
Teléfono 7078-0400
Email: economia@udesa.edu.ar*

Unraveling the Paradox of Anticorruption Messaging: Experimental Evidence from a Tax Administration Reform*

Nicolas Ajzenman, *McGill University*

Martín Ardanaz, *Inter-American Development Bank*

Guillermo Cruces, *Universidad de San Andrés-CONICET & University of Nottingham*

Germán Feierherd, *Universidad de San Andrés*

Ignacio Lunghi, *New York University & CEDLAS-IIE-UNLP*

November 2025

Abstract

Corruption—and the widespread perception of it—poses significant obstacles to development by eroding institutional trust and reducing citizens’ willingness to pay taxes. Yet, government efforts to improve public perceptions by combating corruption may prove ineffective—or even backfire—when confronted with entrenched pessimistic beliefs. We propose that providing an external benchmark of corruption to shift the reference point before highlighting government actions can mitigate these negative effects. In a survey experiment exploiting an institutional reform within Honduras’s tax agency, we find that messages focusing solely on reform efforts have limited or negative effects. By contrast, a combined message that first corrects pessimistic beliefs and then highlights anti-corruption efforts significantly reduces perceived corruption and tax evasion intentions. A field experiment with approximately 45,000 taxpayers confirms that this sequencing approach increases actual tax compliance. These findings suggest that belief updating is possible—but only when information is structured to first engage and recalibrate skeptical priors.

Keywords: Corruption, Tax Administration, Tax Evasion, Field Experiment.

*This project was reviewed and approved in advance by the Institutional Review Board at the University of Nottingham. The design for this experiment was preregistered in the AEA RCT Registry (RCT ID: *AEARCTR-0012039*). We also had approval for the use of anonymized taxpayer data from the CEDLAS ethics committee (IRB number 20241223). We would like to thank Nicolás de la Cerda, Aaron Erlich, and Horacio Larreguy for their valuable comments and support. We are also grateful to participants at Tulane’s CIPR workshop, the CEDLAS internal seminar, and the LACEA-RIDGE Political Economy meeting for their helpful feedback and suggestions. The findings and interpretations in this paper are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank or the governments it represents.

1. Introduction

Corruption remains a central obstacle to development, reducing economic growth (Mauro, 1995; Fisman and Svensson, 2007) and distorting the allocation of public resources (Acemoglu and Verdier, 2000; Olken, 2006). Yet the consequences of corruption extend beyond its direct effects. When corruption is perceived as widespread, institutional trust erodes (Daniele et al., 2023), corrupt social norms become more entrenched (Corbacho et al., 2016; Ajzenman, 2021; Gulino and Masera, 2023), and citizens’ willingness to contribute to collective goods—such as paying taxes—declines (Levi et al., 2009; Yamou et al., 2024). These dynamics are particularly acute in low-capacity states, where institutional reforms must overcome both real governance deficits and deeply pessimistic public beliefs (Okunogbe and Tourek, 2024).

In response, governments increasingly invest in public communication campaigns to highlight institutional reforms or demonstrate progress in addressing corruption. Yet the effectiveness of anticorruption messages remains highly uneven. At best, such messages are ignored or dismissed; at worst, they may backfire—reinforcing cynicism or distrust among already skeptical citizens. A growing body of research suggests that messages emphasizing the persistence or prevalence of corruption can produce unintended consequences. For instance, studies show that messages describing corruption as widespread or harmful often reduce civic confidence and, among pessimistic respondents, may even increase tolerance for bribery (Peiffer, 2020; Cheeseman and Peiffer, 2022). Corbacho et al. (2016), while not designed as an anticorruption campaign, demonstrates that information about widespread impunity increases citizens’ willingness to bribe, reinforcing permissive social norms. Similarly, Beesley and Hawkins (2022) find that priming citizens with information about either petty or grand corruption—especially when framed in terms of negative consequences—can lower institutional trust and political efficacy. Likewise, Chong et al. (2015) show that even audit-based information about local corruption can reduce turnout and support for challengers, suggesting that citizens may disengage when corruption is seen as systemic. Even messages framed around government reform efforts or civic engagement may have demobilizing effects if they fail to overcome entrenched beliefs about institutional failure (Peiffer, 2020; Cheeseman and Peiffer, 2022).

Nevertheless, the broader literature on political misperceptions paints a more optimistic picture. A series of high-powered studies in political behavior suggest that factual backfire effects—where counter-attitudinal messages increase belief in the original falsehood—are rare and often overstated. Instead, the main barrier to persuasion is limited receptivity, especially when messages contradict prior beliefs or originate from

distrusted sources (Wood and Porter, 2019; Guess and Coppock, 2020; Nyhan, 2020, 2021). Related research in the anticorruption domain highlights how norm-based messaging strategies can be more effective. Köbis et al. (2015, 2022) use descriptive norm messages—such as “less and less people are paying bribes”—to reduce corrupt behavior in both lab and field contexts. Blair et al. (2019) show that media portrayals of reporting and text-based nudges increase citizen willingness to report corruption. Agerberg (2022) shows that informing citizens about others’ moral rejection of bribery reduces bribery intentions, especially among individuals with pessimistic priors about societal values. Erlich and Gans-Morse (2025) further shows that descriptive messages about declining corruption are effective only when recipients already find such claims credible, particularly among those predisposed to believe improvement is possible.

Taken together, these studies underscore the central role of citizens’ prior beliefs, particularly about the prevalence and persistence of corruption, in shaping how anticorruption messages are received. These priors are not passive baselines but active filters: they guide attention, shape interpretation, and influence whether new information is accepted, discounted, or reinterpreted. Classic models of information processing—including motivated reasoning (Taber and Lodge, 2006), Bayesian updating (e.g., Bullock, 2009; Christensen, 2023; Broockman and Kalla, 2023), the Receive-Accept-Sample model (Zaller, 1992), and priming theory (Iyengar and Kinder, 2010)—converge on this point: prior beliefs structure the cognitive processing of political information.

This filtering process helps explain why anticorruption messages may fall flat—or even backfire—when they fail to align with, or explicitly challenge, deeply held expectations. In low-trust settings, even positively framed institutional messages may be dismissed unless they directly engage with citizens’ informational priors. A reform message that acknowledges corruption may inadvertently reinforce the idea that corruption remains the norm. If recipients initially believed that public officials were mostly honest, this new information may update their beliefs downward—normalizing corruption and weakening civic norms. If recipients already believed that corruption was widespread, the same message may simply confirm their priors and prompt skepticism about the plausibility of reform.

A more promising strategy, then, is to recalibrate perceived norms before introducing reform messages. By first correcting these assumptions—reframing what is typical or expected public sector behavior—subsequent information about reform becomes more credible and more likely to shift beliefs and behavior. This sequential approach—first providing a benchmark, then presenting evidence of reform—does not merely avoid priming pessimism; it changes the normative baseline against which institutional performance

is judged. We test this approach in Honduras, where the government implemented a major reform of its national tax authority (*Servicio de Administración de Rentas*, or SAR). The reform involved replacing the majority of personnel, introducing integrity vetting procedures, and rebranding the agency to signal a break with past corruption.

To evaluate how citizens respond to messages about such reforms, we designed and pre-registered a survey experiment that later informed a government-led field experiment using real taxpayer behavior. Participants were randomly assigned to receive one of three informational messages: a “reform” message highlighting the administrative overhaul; a “debiasing” message presenting data from a neutral, non-governmental source (Latino-barómetro) showing that perceptions of corruption in Honduras’s tax agency were among the lowest in the region; and a “combined” message presenting the benchmark first, followed by the reform. We hypothesize that this two-part strategy will be more persuasive in shaping beliefs and compliance behavior. We examine the effects of these messages on beliefs about institutional corruption, attitudes toward corruption more broadly, and self-reported tax evasion intentions in the survey experiment. We then assess behavioral outcomes in the field, where similar messages were randomly assigned to around 45,000 taxpayers ahead of the income tax filing deadline.

Our findings suggest that skepticism—not classic backfire—is the dominant obstacle to persuasive institutional messaging. Participants in our survey overwhelmingly perceived corruption as widespread, with mean and median scores of 8 and 9 (on a 10-point scale). The “reform” message alone did not improve respondents’ beliefs or attitudes about institutional integrity. On the contrary, it marginally reinforced negative perceptions, though these effects were not statistically significant ($p = 0.12$). In contrast, the “debiasing” message significantly improved beliefs about corruption within the tax authority. The “combined” treatment, which presented the benchmark first and then the reform message, produced the largest and most consistent improvements, surpassing the “reform” treatment arm with effects significant at the 1% level.

Because positive beliefs about tax authorities are theoretically linked to improved tax compliance (Martin, 2023; Baum et al., 2024), we also examined outcomes capturing tax evasion intentions. In the survey experiment, we used a crosswise randomization technique to reduce social desirability bias when measuring tax compliance. This method pairs a sensitive question with an unrelated one (e.g., a parent’s birth month), allowing us to estimate the true prevalence of tax evasion intent without direct disclosure (Jann et al., 2012). Neither the “reform” nor “debiasing” message significantly changed these intentions on their own, but the combined message led to a statistically significant decrease in evasion propensity. Self-reported (non-crosswise) measures showed no treatment

effects, as expected given their susceptibility to social desirability bias.

To assess behavioral consequences beyond responses in a survey, we collaborated with the Honduran tax authority to implement a large-scale field experiment. In the weeks leading up to the income tax filing deadline, taxpayers were randomly assigned to receive either the debiasing message, the combined message, or no message (control).¹ The results closely mirrored those of the survey experiment: the combined message significantly increased declared taxable income—by 6.8% on average, and by 11.5% among high-risk taxpayers—while the debiasing message alone produced a positive but statistically insignificant effect. These findings suggest that even in low-trust environments, informational strategies that begin with a credible external benchmark—re-anchoring expectations before institutional claims are introduced—can meaningfully shape belief updating and promote compliance behavior.

This study makes several contributions. First, it speaks to ongoing debates about the prevalence of backfire effects in response to counter-attitudinal information. While early research raised concerns that corrective messages could entrench misperceptions (Nyhan and Reifler, 2010) or that corruption-related information could provoke cynicism and disengagement (Chong et al., 2015; Corbacho et al., 2016), more recent large-scale studies suggest that backfiring effects are rare and often overstated (Nyhan, 2020; Guess and Coppock, 2020; Nyhan, 2021; Erlich and Gans-Morse, 2025). Our findings are consistent with this more optimistic view. We observe little evidence of genuine backfire, which is at most suggestive. Instead, skepticism and belief rigidity limit the persuasive reach of institutional messages, especially in low-trust contexts.

Second, we contribute to a growing literature on anticorruption messaging, which has yielded mixed results. Several studies suggest that messages emphasizing widespread or persistent corruption can reinforce cynicism, particularly among citizens with more pessimistic priors (Peiffer, 2020; Cheeseman and Peiffer, 2022). At the same time, other interventions that highlight positive social norms, frame institutional improvement as a trend, or draw on neutral and credible sources have been shown to reduce bribery intentions, increase civic cooperation, and strengthen perceptions of institutional integrity (Köbis et al., 2015; Blair et al., 2019; Köbis et al., 2022; Erlich and Gans-Morse, 2025). We extend this work by testing whether reform messages are more effective when preceded by a credible external benchmark that re-anchors beliefs. Our results suggest that such a two-pronged strategy—explicitly accounting for prior beliefs—can reduce resistance and improve compliance outcomes.

¹Due to concerns that the standalone reform message might reduce compliance, it was omitted from this phase.

Third, we contribute to research on belief formation and the psychological mechanisms that explain failures of information updating. A large body of work in political psychology shows that individuals often depart from Bayesian learning due to motivated reasoning (Taber and Lodge, 2006; Taber et al., 2009; Epley and Gilovich, 2016), confirmation bias (Little et al., 2022), or source distrust (Flynn et al., 2017). These belief distortions shape attitudes toward government (Enríquez et al., 2025), responsiveness to elite cues (Bullock, 2011; Nicholson, 2012; Boudreau and MacKenzie, 2014) and reactions to corrective information (Barrera et al., 2020). While many studies document these biases, fewer link them to downstream behavioral consequences. Our study helps fill this gap by showing that belief-correcting interventions can influence concrete behaviors—such as tax compliance.

Finally, we contribute to a growing experimental literature on tax behavior in low-capacity states. Existing field experiments typically focus on public lotteries (Dunning et al., 2025), enforcement cues (Slemrod, 2019) or moral appeals rooted in reciprocity and intrinsic motivation (Luttmer and Singhal, 2014). In contrast, we examine whether messages that shift corruption perceptions—rather than prizes, threatening penalties or invoking duty—can increase tax compliance.

2. Hypotheses

Following standard practice in the political psychology literature, we define beliefs as the initial intuitions and feelings that serve as a baseline for processing new information. These beliefs—shaped by personal experience, social context, and elite cues—structure how citizens predispose themselves to accept new information as true or false, favorable or unfavorable. Rather than serving as neutral baselines, priors function as cognitive filters: they guide attention, influence evaluation, and often distort or discount counter-attitudinal information (Zaller, 1992; Taber and Lodge, 2006).

This presents a particular challenge for anticorruption campaigns. Priors are frequently cited as the “strongest obstacles to political persuasion” (Cotter et al., 2020, p. 69), and in environments where corruption is seen as pervasive, even reform-oriented messages may be met with skepticism or indifference.

This concern is central to a growing body of experimental research showing that anti-corruption messages can backfire when they inadvertently prime negative associations. In a survey experiment in Nigeria, Peiffer (2020) finds that even “positive” messages—those highlighting government success in curbing corruption—may reinforce negative perceptions among pessimistic respondents. As she notes, “if a positively toned message inad-

vertently primes a negative issue, the message can lead individuals to adopt more negative views about the issue” (p. 7). Similarly, Cheeseman and Peiffer (2022) show that messages describing corruption as widespread or harmful often reduce civic confidence and may increase tolerance for bribery.

Despite this, other strands of the literature suggest that informational interventions can shift beliefs and behavior in the expected direction—particularly when messages emphasize positive trends, rely on norm-based framings, or come from sources perceived as trustworthy (e.g., Köbis et al., 2015; Blair et al., 2019; Köbis et al., 2022; Agerberg, 2022; Erlich and Gans-Morse, 2025). These studies suggest that message effects depend not only on tone or salience, but on whether recipients view the information as credible and belief-congruent.

Our belief-shifting approach builds on this existing research by suggesting that individuals interpret messages through reference points about what is typical or socially normative. For instance, Schultz et al. (2007) demonstrate that norm-based messages can either reduce or reinforce undesirable behavior depending on where individuals perceive themselves relative to the described norm. From this perspective, backfire effects may reflect norm reinforcement—where anticorruption messages inadvertently signal that corruption remains common or socially accepted—rather than priming-based explanations and models of motivated reasoning, which emphasize how messages activate existing associations and elicit affective or directional biases.²

While our sample is overwhelmingly pessimistic, it is useful to consider the mechanisms that might operate across the full distribution of priors. Among respondents who believe corruption is relatively rare (i.e., optimistic priors), a message that emphasizes reform may actually increase perceived corruption by inadvertently revealing that wrongdoing was more common than assumed—reinforcing a more negative normative baseline. By contrast, among respondents with pessimistic priors, the same reform message may confirm their expectations but fail to persuade, as claims of progress are interpreted as implausible given their belief that corruption is the norm. In both cases, the reform message may underperform—but for distinct psychological reasons: norm reinforcement for optimists, and credibility skepticism for pessimists.

Following this logic, our approach begins by examining how anticorruption messages shape beliefs. In contexts of institutional distrust, even well-intentioned reform messages may be discounted or ignored. We hypothesize that effective persuasion requires a two-step approach: first, confronting priors by reframing expectations, and only then

²For a discussion of competing theoretical accounts that predict backlash, see Guess and Coppock (2020).

introducing new information. Shifting these priors can make citizens more receptive to institutional claims that might otherwise be met with skepticism. We test whether anchoring beliefs in credible, comparative data can reduce the risk of backlash. This mechanism draws on political psychology research showing that prior attitudes operate as interpretive anchors and that reference points critically shape how new information is evaluated (Cialdini et al., 2006; Schultz et al., 2007; Taber et al., 2009).

Our “debiasing” strategy follows this logic: We examine whether introducing a neutral, third-party benchmark—Latinobarómetro data showing that perceived corruption in Honduras’s tax agency is relatively low by regional standards—can recalibrate expectations and enhance the credibility of subsequent institutional claims. By embedding this comparative information within the message, we provide a credible external benchmark designed to weaken pessimism and create space for belief revision.³ We hypothesize that only after expectations are reframed will messages about institutional improvement meaningfully influence beliefs and behavior.

This approach contrasts with interventions that ask respondents to “take the perspective of others” (e.g., Enríquez et al., 2025) or “keep an open mind” (e.g., Barabas, 2004), or those that rely on contextual moderators such as monetary incentives for accuracy (Bullock et al., 2015). Our design avoids relying on explicit instructions about how to process information, offering an alternative framework for how individuals might encounter and interpret anti-corruption campaign material in everyday contexts.

Based on these insights, the pre-analysis plan proposed the following hypotheses:⁴

H1: Reform Treatment. Communicating government efforts to combat corruption might have a counterproductive effect, worsening perceptions of the Honduran tax authority and increasing the perceived likelihood that tax officials engage in corrupt practices relative to a neutral message (control).

H2: Debiasing Treatment. Presenting information emphasizing that the average Honduran perceives corruption in tax administration to be relatively low compared to citizens in the rest of Latin America might positively influence individuals’ perceptions of the Honduran tax authority and reduce the perceived likelihood that tax officials engage in corrupt practices relative to a neutral message (control).

H3: Combined Treatment. The combination of changing beliefs about corruption

³Research in political communication highlights the persuasive strength of “in-group” messengers—such as average citizens or everyday people—who are often perceived as more trustworthy than elite or institutional sources (Hewitt et al., 2024).

⁴Our full pre-analysis plan can be found in Appendix Section E. We have changed the labels of the treatments from the pre-analysis plan as follows: “perception” has been renamed to “debiasing” and “purge” to “reform.”

(“debiasing”) followed by a message emphasizing government actions (“reform”) might improve more effectively perceptions of the Honduran tax authority and reduce the perceived likelihood that tax officials engage in corrupt practices relative to a neutral message.

H4: Combined Treatment versus Reform Treatment. The “combined” treatment might significantly outperform just emphasizing government actions (“reform”) in terms of improving corruption perceptions of the Honduran tax authority and reducing the perceived likelihood that tax officials engage in corrupt practices relative to a neutral message. This hypothesis reflects the idea that the same reform message can have divergent effects depending on the normative reference point it activates. When preceded by a credible benchmark that shifts priors, the reform message is more likely to be interpreted as evidence of progress. When delivered on its own, it may reinforce existing beliefs or be dismissed as implausible.

3. Case Selection: Tax Administration Reform

Honduras is a lower-middle-income country in Latin America with GDP per capita of \$7,178 PPP in 2023. It ranks 173 out of 213 countries in the World Bank’s Control of Corruption Index. Tax revenues account for 18 percent of GDP and tax collection is heavily reliant on taxes on goods and services, making up over half of the total tax burden.

Starting in 2014, the Honduran tax administration underwent a series of reforms and institutional changes aimed at strengthening the country’s tax system. These included improvements in operational management, a new billing regime, and the adoption of new technologies, among others. The key salient dimension of the reform that we exploit in our experiment was an effort to restore the tax administration’s corporate image through a new human resource policy, motivated in part by the need to address corruption inside the tax administration.

Prior to the reform, the tax administration faced challenges such as insufficient professionalization of its workforce,⁵ lack of transparency and integrity in crucial areas of the organization,⁶ inefficient tax processes, and weak information systems that hampered attempts at evasion control and contributed to poor tax collection performance (e.g.,

⁵For instance, only a third of the tax administration personnel held a college degree by the time of the reform onset. See IDB (2015).

⁶Prior to the reform, out of a sample of more than 800 employees who took a polygraph test, 30 percent failed it. The tax administration received, on average, 90 complaints per month related to issues of lack of transparency in job performance. See IDB (2015).

tax revenues hovered at 15 percent of GDP by the time of the reform onset). Citizen perceptions of corruption in the tax administration remained widespread: more than 40 percent of Honduran citizens believed that all or almost all tax officials were involved in corrupt practices (Latinobarómetro 2016).

Recognizing the need for a comprehensive overhaul, between 2014 and 2019 the government implemented a major institutional reform of the tax administration, with support from multilateral organizations. The reform involved the dissolution of the existing tax administration by executive decree, including the dismissal of 1,500 employees, constituting 85 percent of existing personnel.⁷ Approximately 300 employees were retained to safeguard the tax base during the transition period. The recruitment and selection process of personnel was conducted under the guidelines of a new human resource policy, including: (i) a tax career path with respective job profiles, (ii) competitive salary scales and performance evaluation procedures, and (iii) promotion schemes. New hires were subject to several tests prior to selection, including polygraph evaluations.

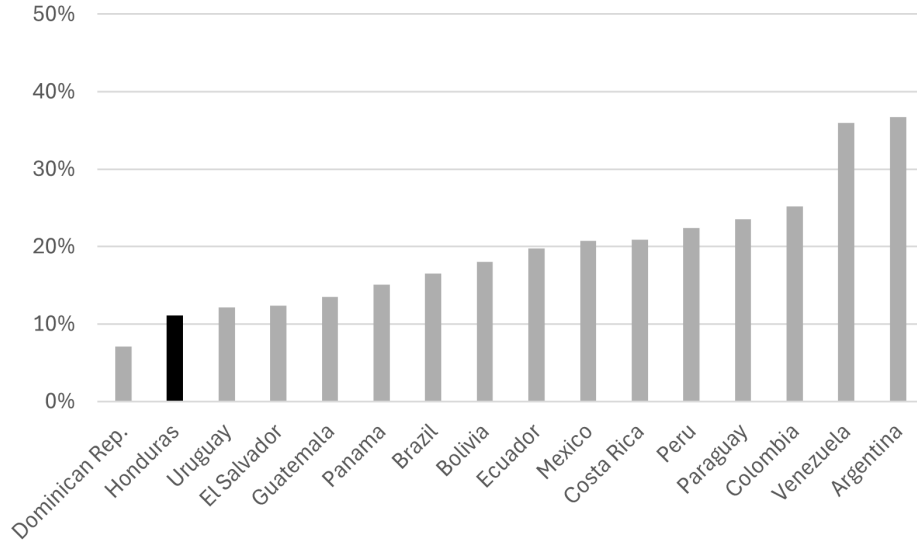
By January 2017, the new tax administration, *Servicio de Administración de Rentas* (SAR), started operations with around 500 workers, a combination of employees from the transition period and new hires. The transition occurred abruptly, without a formal media rollout or broader communication strategy. Public knowledge of the reform was therefore limited, and it is unlikely that respondents had pre-formed views or affective priors regarding its content or implications.⁸

Substantively, the reform brought significant institutional changes. It expanded the tax base, tripling the number of active taxpayers; it streamlined compliance by increasing electronic filing rates for major taxes from under 50% to over 95%; and it improved staffing, with the share of tax personnel holding college degrees more than doubling. These changes contributed to a sustained rise in tax collection—20% above pre-reform levels until the COVID-19 crisis. Notably, despite the limited public attention at the time of implementation, perceptions of corruption within the tax authority have improved over time. Nationally representative surveys now rank Honduras’s tax agency among the least corrupt in Latin America (see Figure 1).

⁷A special commission was appointed to proceed with termination of existing contracts. Additionally, 900 employees from the customs area were transferred to the Ministry of Finance.

⁸We conducted a review of Google Trends, media coverage and public information campaigns surrounding the restructuring of the *Dirección Ejecutiva de Ingresos* (DEI). Beyond official decrees and a single technocratic blog post, we found scant evidence that the reform received sustained public attention.

Figure 1: Corruption Perceptions of Tax Administration - Latinobarometro



Notes: The figure shows the answer to the question “How many of the following people do you think are involved in corruption, or haven’t you heard enough about them to say?” for the year 2020 of the Latinobarometro survey, showing the response for the tax administration. The figure shows the proportion of people who believe that the tax administration is corrupt by country, these being ordered from lowest perceptions of corruption to highest.

Alt text: Bar chart of perceived corruption in tax administrations across Latin America. Bars show the share of citizens viewing each national tax agency as corrupt. Honduras appears among the lowest, indicating relatively positive perceptions.

4. The Survey Experiment

To assess how individuals react to anticorruption efforts by the tax administration, we designed a survey experiment with data collected through an online survey between October 2, 2023, and October 4, 2023. We recruited 2,372 participants through Facebook ads (2,000 targeted), but excluded from our main sample those individuals who did not pass our attention check, who had their IP address duplicated within the sample, or who had their IP address duplicated with our pilots, resulting in our final sample of 1,411 observations. Participants were recruited from the population of Honduran Facebook users, and we used quota sampling to match the distribution of age, gender and region in the national census of 2013. While the sample is not fully representative of the general population of Honduras, it approximates key demographic benchmarks and provides meaningful variation to test our hypotheses.

4.1. Treatments

Respondents in our survey were assigned to one of four information treatments:⁹

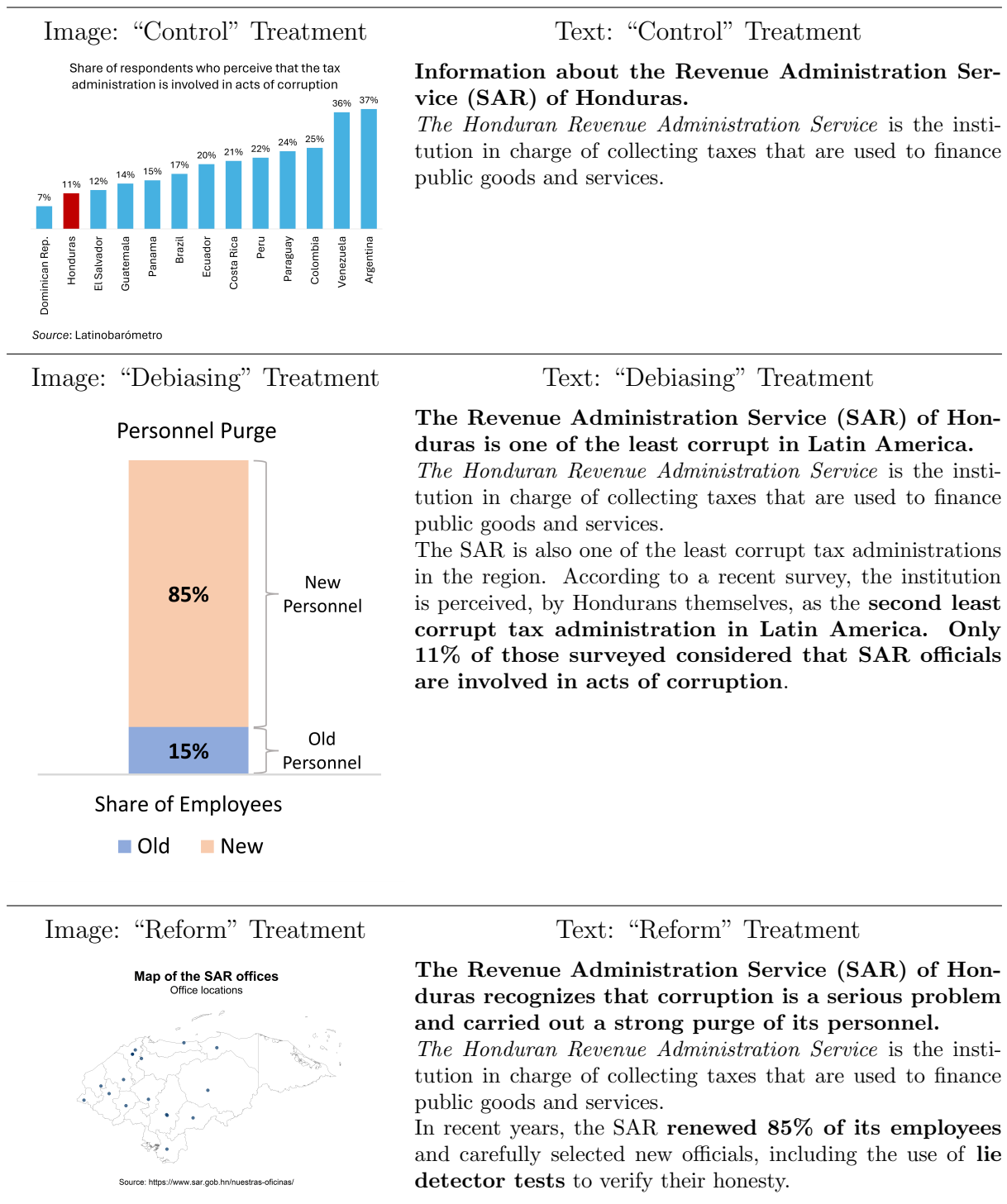
1. The control group, which received basic information about the SAR’s role in tax collection, along with an image of the locations of the SAR offices in Honduras.
2. The “reform” treatment arm, which transmits the SAR’s recognition of corruption as a significant issue and recent efforts to combat it. The message highlights the personnel renewal done by the government for the purpose of reducing corruption within the tax agency and includes a simple figure to illustrate the magnitude of the policy reform.
3. The “debiasing” treatment arm, which provides basic information about the SAR’s role in tax collection and information about the SAR’s reputation as one of the least corrupt tax authorities in Latin America, backed by a graph using survey data indicating it as the second most honest tax authority institution in the Latin American region.¹⁰
4. The “combined” treatment arm, which combines the two previous information treatments, with the “debiasing” message shown first, followed by the “reform” treatment.

Participants were randomly assigned to each group with equal probability (25 percent to each group). They were also informed about the anonymity of their answers. Figure 2 presents the treatments displayed to each respondent. Appendix Table B1 presents balance check among the treatment groups based on our pre-treatment variables. Appendix Table B2 presents the descriptive statistics for our outcomes and main pre-treatment variables.

⁹See Appendix Section A for the full treatment layouts.

¹⁰While Honduras’s high ranking in this figure may reflect genuine institutional reforms—such as the dissolution of the old tax agency, personnel turnover, and new enforcement protocols—we do not claim to identify the cause of the improved perception; rather, we use the 2020 figure as a plausible and credible third-party signal. The logic of the treatment hinges not on the figure capturing “truth,” but on its perceived credibility and its capacity to re-anchor expectations in a low-trust environment.

Figure 2: Treatment Messages: Survey Experiment (Translated)



Note: Each row shows one treatment from the survey experiment. Left panels display the visual content shown to participants; right panels present the text received. The “combined” treatment shows first the “debiasing” and then the “reform” treatment. A more detailed display of the treatments—including minor textual variations in the “combined” treatment and the exact sequence in which texts and images were presented—is available in Appendix Section A.

Alt text: Multi-panel figure displaying the four information treatments used in the survey experiment. Each row pairs an image on the left with translated message text on the right. The control message provides basic information about the tax authority. The debiasing message highlights that Honduras’s tax agency is among the least corrupt in Latin America. The reform message describes the replacement of 85 percent of staff and use of lie-detector tests. The combined treatment presents the debiasing message first and the reform message immediately after, illustrating the sequential design.

4.2. Outcomes

The primary focus of this analysis is respondents’ beliefs and perceptions regarding the tax authority and pervasive corruption. Respondents rated their beliefs on a scale from 0 (not likely at all) to 10 (very likely) on the likelihood of SAR officials engaging in corruption (*Perception of Corruption of the SAR*) and the chance that a SAR official might seek gifts or money for resolving or preventing tax-related issues (*Probability That a SAR Agent Will Accept a Bribe*). Aligning with standard practices for multiple outcome analysis, we developed the *Beliefs Index* combining these two outcomes. Since both outcomes are measured on the same scale, this index is simply the average of both outcomes.

Further, we examine secondary outcomes concerning individuals’ willingness to evade taxes. To counter potential biases in self-reported measures of tax evasion, we employed a crosswise measure. This involved respondents reporting the truthfulness of two statements: one on their willingness to evade taxes and the other, a non-sensitive question about the birth month of their mother. With the known probability distribution of the latter, unbiased prevalence estimates on the stated willingness to evade taxes were obtained, following Jann et al. (2012). We alter our outcome in the following way:

$$\tilde{R} = (R + p - 1)/(2p - 1) \quad (1)$$

Where p is the probability that an individual’s mother was born in November or December, which on average is $p = 2/12$, and R is a dummy variable equal to 1 if individuals have responded that both statements (the one about their willingness to evade taxes and the non-sensitive statement) are true or that both statements are false, while the dummy is equal to 0 if respondents reported only one of the statements is true.¹¹

We also evaluated respondents’ reactions to a hypothetical scenario where a store owner offers a VAT-exempt discount, probing their readiness to participate in tax evasion. Respondents were presented with a scenario in which a store owner offered a 10 percent discount in exchange for a sale without VAT during a shopping transaction. Respondents indicated whether they were willing to accept this discount, implying a willingness to evade the corresponding tax.¹²

Additionally, we incorporated a series of questions to probe our belief-updating mech-

¹¹For more information about this transformation see Appendix Section C.1.

¹²Another possible implementation of a crosswise analysis through a *joint model* (Gingerich et al., 2016) which simultaneously incorporates the binary indicator of whether the individual accepts the discount (a direct measure of willingness to evade) is presented in Appendix Section C.2.

anism. Specifically, we asked them to evaluate the *SAR’s Capability*—their assessment of the tax administration’s ability to collect taxes owed by taxpayers. To examine broader effects beyond taxation, we included two additional measures: *Corruption of City Residents*, which captured perceptions of fellow citizens’ involvement in corrupt activities, and a behavioral test, *Lies in Dice Game*. In this game, participants had the opportunity to misreport their dice rolls to improve their lottery odds, serving as a proxy for their propensity to provide dishonest information. Importantly, the virtual die was heavily biased toward the value 1, thereby maximizing incentives for participants to misreport. Unlike traditional dice games, the true outcome was visible to the investigators; therefore, we can directly observe whether a respondent lied. For further details see Appendix Section C.3, which includes the detailed explanation of the dice game implementation.¹³

4.3. Estimation

We estimate the following linear regression model by Ordinary Least Squares for each of our outcomes:

$$Y_i = \alpha + \beta_1 \text{Reform}_i + \beta_2 \text{Debiasing}_i + \beta_3 \text{Combined}_i + X_i \beta_4 + \varepsilon_i \quad (2)$$

The specification includes indicator variables for each treatment arm with the control group as the reference category. Y_i is the outcome analyzed, while X_i is a control vector. We include covariates to improve the precision of treatment effect estimates, following best practices in the analysis of randomized experiments (Gerber and Green, 2012). The control set is selected using a double-selection lasso procedure.¹⁴ ε_i is the error term, estimated using Huber-White standard errors to account for the potential heteroskedasticity. We report the coefficients β_1 , β_2 , and β_3 for each of our regressions, while also including the mean of the control group for comparison. Finally, we also include for each outcome two different tests in which we compare whether the coefficient of the “combined” treatment is the same as that of the “debiasing” or “reform” treatment, respectively.¹⁵

¹³These outcomes were pre-registered as secondary hypotheses.

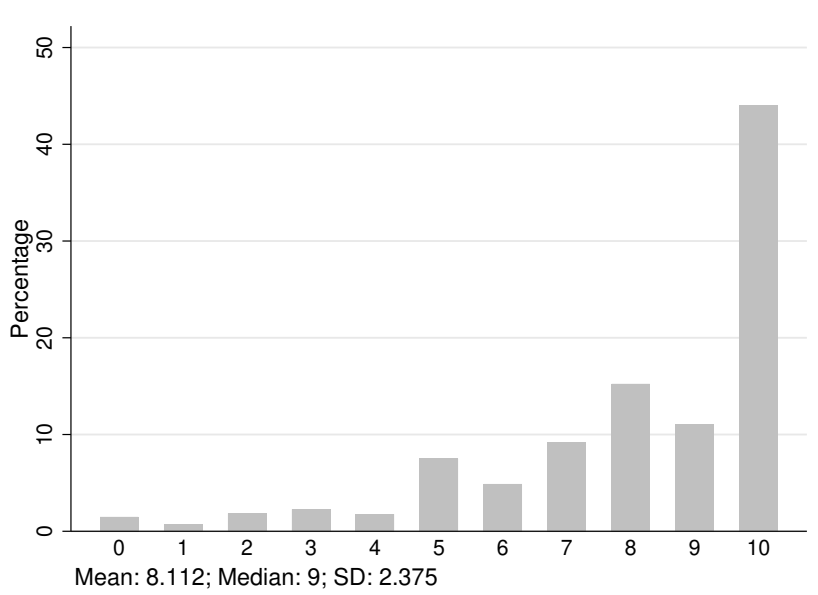
¹⁴The potential controls, ultimately selected or excluded by lasso, included respondent characteristics such as sex, aggregated age groups, ideological self-placement, education level, confidence in institutions (e.g., church, congress, executive power, and police), prior perceptions of corruption among administrative officials and Hondurans in general, and region.

¹⁵Appendix Table B3 tests for potential experimenter demand effects by examining whether treatments influence two outcomes: the survey rating (number of stars) and the probability of leaving a comment different than “none” or similar. Results show no significant effects of the treatments on either outcome, indicating that experimenter demand effects are unlikely.

4.4. Main Results

We first document extremely pessimistic prior beliefs about generalized corruption. As depicted in the Histogram of Pre-Treatment Corruption Beliefs (Figure 3), around 70% of respondents assign an 8 or more to the likelihood that public administration officials engage in corrupt practices, with nearly 45% of the sample assigning the maximum likelihood. These beliefs reveal widespread skepticism about public-sector integrity, establishing an important baseline for assessing whether our interventions can shift entrenched perceptions.

Figure 3: Histogram of Pre-Treatment Corruption Beliefs of the Public Administration



Notes: The figure shows a histogram of the responses to the question “On a scale from 0 to 10, where 0 is “not likely at all” and 10 is “very likely”, how likely do you think public officials in Honduras are involved in acts of corruption?” The mean, median and standard deviation of the variable are also included.

Alt text: Histogram of respondents’ beliefs about public corruption on a 0–10 scale. Most answers cluster near 8 to 10, showing strong pessimism about officials’ honesty.

We then assess the effect of SAR’s messages on our primary outcomes, beginning with the *Beliefs Index*. As displayed in column (1) of Table 1, the impact of the treatments on individual beliefs is nuanced. The “reform” treatment showed a positive but statistically insignificant effect on perceptions of corruption. Conversely, the “debiasing” treatment—which highlighted SAR’s standing as one of the least corrupt tax agencies in the region—produced a statistically significant reduction in corruption beliefs, lowering the *Beliefs Index* by an average of 0.43 units. With a baseline belief of 6.2 units in the control group, this effect translates to a 7 percent decrease in perceived corruption. The most

pronounced impact came from the “combined” treatment, which provides the “debiasing” and “reform” messages to respondents. This approach resulted in a significant reduction in the *Beliefs Index*, decreasing it by an average of 0.78 units, representing a notable 13 percent decline in corruption beliefs.¹⁶

Table 1: Beliefs Index and Willingness to Pay Outcomes - Survey Experiment

Variables	(1) Beliefs Index	(2) Willing to Evade Taxes	(3) Accepts Discount
Reform	0.263 (0.169)	2.173 (5.572)	2.680 (3.440)
Debiasing	-0.433** (0.182)	-7.027 (5.644)	-0.604 (3.397)
Combined	-0.778*** (0.179)	-11.159** (5.449)	-2.642 (3.267)
Observation	1,411	1,411	1,411
Control Group Mean	6.150	51.4%	33.3%
LASSO controls	Yes	Yes	Yes
P-value: Reform = Combined	0.000	0.017	0.125
P-value: Debiasing = Combined	0.068	0.467	0.556

Notes: Robust standard errors in parentheses. “Beliefs Index” outcome is the mean between the *Belief of Corruption of the SAR* outcome and the *Probability that a SAR Agent will Accept a Bribe* outcome, which varies from 0 (not likely at all) to 10 (very likely). “Willing to Evade Taxes” outcome is a dummy variable, which was transformed according to Jann et al. (2012) in order to get the unbiased estimates of the willingness of evading taxes from the crosswise initial measure. Both outcomes varies from 0 (not likely at all) to 10 (very likely). “Accepts Discount” outcome equal to 1 if the respondent is willing to accept buying without VAT included. Controls were chosen using a double-selection lasso linear regression from the following possibilities: respondent’s sex, aggregated age groups, ideological self-placement, education level, confidence in institutions (church, congress, executive power, police), prior perceptions of corruption (administrative officials and Hondurans in general), and region. Dummy variables were multiplied by 100 so that the interpretation of the coefficients in percentage points is direct. The last two rows present the p-values for two tests of coefficient equality: the first compares the coefficient for the “reform” treatment with that for the “combined” treatment, while the second compares the coefficient for the “debiasing” treatment with the “combined” treatment. P-values: ***p<0.01, **p<0.05, *p<0.1.

The last two rows show significant differences between the combined treatment and other treatments. The row that tests the “reform” effect against the “combined” one shows a large and significant difference, which confirms our hypothesis that changing individuals’ initial corruption perceptions leads to improvements in the reception of anti-

¹⁶Results without controls are shown in Appendix Table B4.

corruption messages. We also show that the difference between the “combined” and “debiasing” treatment arm is statistically significant, however only at a 10% significance level. Appendix Table B5 present these results for each component of the *Beliefs Index* separately.

Table 1 also examines the *Willingness to Evade Taxes* outcome (column (2)). The coefficient associated with the “reform” message is positive but not statistically significant at conventional levels. In turn, the “debiasing” treatment consistently had a negative effect on people’s willingness to evade taxes; however, again, none of the coefficients are statistically significant at the conventional level. Combining both messages, however, had a substantial, statistically and economically significant impact on the individuals’ willingness to evade taxes, reducing it by approximately 11 percentage points on average. Given a prevalence estimate of 51 percent in the control group, the treatment effect is able to reduce the share of evaders by around 21 percent. Consistent with the results on beliefs, the difference between treatments is significant when comparing the “combined” and “reform” treatment arms at a 5% significance level.¹⁷

These findings show that the “combined” treatment—first recalibrating expectations, then highlighting reform—effectively improves both corruption perceptions and attitudes toward tax evasion. As shown in the final row of Table 1, the same reform message has significantly different effects depending on whether it is preceded by a belief-shifting benchmark. This supports our core hypothesis: messages are more persuasive when they first engage with informational priors using credible reference points. Rather than being dismissed or triggering skepticism, institutional claims gain traction once expectations have been reframed. In contrast to earlier studies that warned against invoking the state in anticorruption messaging due to backfire risks (Cheeseman and Peiffer, 2023, p. 1047), our results show that such messages can be effective—so long as they are sequenced in a way that makes progress seem plausible (e.g., Erlich and Gans-Morse, 2025).

Not all evidence, however, aligns with our hypotheses. We also included a question assessing respondents’ likelihood of colluding with a seller to avoid paying sales tax. The analysis of the *Accepts Discount* question, presented in column (3) of Table 1, reveals no significant average effects from our treatments on respondents’ willingness to accept a VAT-free discount. This null result may stem from the nature of the outcome itself: self-reported behaviors are prone to social desirability bias, where respondents report what they perceive as morally acceptable rather than their actual behaviors.¹⁸ Nevertheless,

¹⁷Appendix Table C1 presents the results using a logit model instead of a linear probability model, being consistent with those shown in Table 1.

¹⁸Following Gingerich et al. (2016) and Corbacho et al. (2016), we estimate a joint model that combines responses from a crosswise question and a related direct question to infer both the prevalence of the

the field experiment analyzed below demonstrates that our messages did influence real-world behavior, leading to substantial improvements in tax compliance.

4.5. Further analysis

In previous sections, we argued that benchmarking citizens’ beliefs—by showing that Hondurans, on average, perceive their tax agency more favorably than residents of other Latin American countries—could recalibrate expectations and encourage more compliant tax behavior. To examine this mechanism more closely, we pre-registered additional hypotheses that tested the relationship between belief updating, institutional messaging, and tax evasion. Three key findings support this belief-based explanation.

First, our treatments influenced perceptions of the tax authority’s enforcement capacity. Column (1) of Table 2 shows that all messages led to improvements, with the “combined” treatment generating the largest increase—about 0.23 units, or a 7 percent gain relative to control.¹⁹ Although the “reform” message alone produced a weaker and only marginally significant effect, differences across treatment arms were not statistically distinguishable. These findings are consistent with earlier results: the “debiasing” and “combined” treatments improved beliefs about corruption within SAR and reduced perceptions that its agents accept bribes.

Column (2), in contrast, shows no treatment effects on beliefs about corruption among other city residents. This null result strengthens our interpretation: since tax compliance increased even though perceptions of peer behavior remained unchanged, the observed behavioral effects are more plausibly driven by updated beliefs about the state’s ability to detect tax evasion—rather than shifts in social norms or conformity pressures.

To further explore behavioral responses, we included a virtual dice game designed to incentivize dishonest reporting. The game was biased toward rolling a “1,” so overreporting higher values implied misrepresentation. Column (3) shows no significant treatment effects on this measure of dishonest behavior, with one exception: the “reform” treatment increased dishonesty by about 4 percentage points relative to the 8 percent baseline,

sensitive trait and patterns of misreporting. The joint model leverages the protection offered by the crosswise design to assume truthful responding there, while allowing for lying under direct questioning, under the assumption of one-sided lying (only those with the sensitive trait may misreport). As Appendix Table C2 shows, estimates suggest that among respondents willing to evade taxes, approximately 58.5% ($\hat{\lambda}_1^T = 0.585$) truthfully report their willingness under direct questioning, while around 41.5% ($\hat{\lambda}_1^L = 0.415$) misreport. The qualitative results, using this model, remain the same as the ones of Table 1. Although truthful reporting appears more frequent, misreporting is substantial. However, these results should be interpreted cautiously. First, the direct question pertains to willingness to accept purchases without a value-added tax (VAT) invoice, which, while related, is not identical to general tax evasion. Second, the identification relies on the one-sided lying assumption, which may not fully hold in practice.

¹⁹Estimates without covariates appear in Appendix Table B6.

Table 2: Beliefs About SAR’s Capability, Perceptions of City’s Corruption, and Dishonest Behavior Outcomes

Variables	(1) SAR’s Capability	(2) Corruption of city’s residents	(3) Lies in Dice Game
Reform	0.154* (0.086)	0.179 (0.176)	3.957* (2.182)
Debiasing	0.177** (0.088)	0.248 (0.182)	1.226 (2.143)
Combined	0.230*** (0.086)	0.200 (0.173)	0.098 (2.041)
Observation	1,411	1,411	1,411
Control Group Mean	3.143	5.520	8.4%
LASSO controls	Yes	Yes	Yes
P-value: Reform = Combined	0.384	0.903	0.088
P-value: Debiasing = Combined	0.558	0.784	0.605

Notes: Robust standard errors in parentheses. “Corruption of city’s residents” outcome is the respondent belief that individuals of his/her city are prone to corruption, which varies from 0 (not likely at all) to 10 (very likely). “SAR’s Capability” outcome corresponds to the respondents’ belief that the SAR has the capability of collecting the amounts owed by individuals with tax obligations, which varies from 1 (not capable at all) to 5 (highly capable). “Lies in Dice Game” variable equal to 1 if individual lied on the dice game. Controls were chosen using a double-selection lasso linear regression from the following possibilities: respondent’s sex, aggregated age groups, ideological self-placement, education level, confidence in institutions (church, congress, executive power, police), prior perceptions of corruption (administrative officials and Hondurans in general), and region. Dummy variables were multiplied by 100 so that the interpretation of the coefficients in percentage points is direct. The last two rows present the p-values for two tests of coefficient equality: the first compares the coefficient for the “reform” treatment with that for the “combined” treatment, while the second compares the coefficient for the “debiasing” treatment with the “combined” treatment. P-values: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

although this effect was only marginally significant.²⁰ These findings suggest that belief-based interventions can shift behavior that is closely tied to the treatment content—such as tax attitudes—but may have limited effects on unrelated ethical choices like general honesty.

Second, we find that the belief-updating effects of the debiasing and combined messages were especially pronounced among respondents with more negative priors. To explore this heterogeneity, we interacted treatment indicators with a dummy for whether individuals’ pre-treatment corruption beliefs exceeded the sample median.²¹ This allows

²⁰Differences between treatments were not statistically significant at conventional thresholds.

²¹The relevant survey item asked: “On a scale from 0 to 10, where 0 is ‘not likely at all’ and 10 is

us to formally test whether baseline pessimism moderates treatment effects. The regression specification is as follows:

$$\begin{aligned}
Y_i = & \alpha + \beta_1 \text{Reform}_i + \beta_2 \text{Debiasing}_i + \beta_3 \text{Combined}_i + \beta_4 \text{Prior}_i^+ \\
& + \beta_5 \text{Prior}_i^+ \cdot \text{Reform}_i + \beta_6 \text{Prior}_i^+ \cdot \text{Debiasing}_i + \beta_7 \text{Prior}_i^+ \cdot \text{Combined}_i \\
& + X_i \beta_8 + \varepsilon_i
\end{aligned} \tag{3}$$

Here, Prior_i^+ is a binary variable equal to one if the respondent’s pre-treatment belief score exceeded the median value of 9, as shown in Figure 3. The model retains the same covariates as equation (2). This specification allows us to distinguish how treatment effects differ between respondents who entered the study with relatively pessimistic versus moderately negative views about public corruption. Balance tables for each subgroup appear in Appendix Tables B7 and B8.

Table 3 presents estimates for two outcomes: the Beliefs Index and Willingness to Evade Taxes. To facilitate interpretation, the table reports linear combinations of treatment and interaction terms for each subgroup. Among respondents with less pessimistic priors (i.e., corruption belief scores below the median), the “reform” message backfired, increasing perceived corruption by 0.44 points. The other treatments yielded no statistically significant effects for either outcome. For the Willingness to Evade Taxes outcome, none of the messages had a detectable impact among this subgroup.

The pattern differs sharply among those with more pessimistic priors. Here, both the “debiasing” and “combined” treatments significantly reduced perceived corruption. The “debiasing” message lowered the *Beliefs Index* by 0.68 points, while the “combined” treatment produced the largest change: a 1.52-point reduction in perceived corruption, equivalent to a 25 percent decrease relative to the control group mean. These effects suggest that combining a benchmark with reform content magnifies impact among the most skeptical respondents.

For tax-related attitudes, the “combined” treatment again had the strongest effect, reducing *Willingness to Evade Taxes* by 14.3 percentage points. These results reinforce the idea that engaging priors directly is particularly important for persuading individuals who start with entrenched distrust.²²

These patterns challenge conventional priming accounts, which predict greater backfire among pessimists. Instead, we find that skeptical respondents are more receptive

‘very likely’, how likely do you think public officials in Honduras are involved in acts of corruption?”

²²Because the experiment was not stratified by prior beliefs, we report estimates with covariate adjustment but not without controls.

Table 3: Heterogeneity Analysis of Beliefs Index and Willingness to Evade Outcomes

Variables	(1) Beliefs Index	(2) Willing to Evade Taxes
<i>Low Pre-Treatment Corruption</i>		
Reform - Low Corruption	0.444** (0.222)	11.095 (7.69)
Debiasing - Low Corruption	-0.205 (0.238)	-3.436 (7.649)
Combined - Low Corruption	-0.244 (0.218)	-6.189 (7.280)
P-value: Debiasing = Combined - Low Corrup.	0.864	0.714
P-value: Reform = Combined - Low Corrup.	0.001	0.022
<i>High Pre-Treatment Corruption</i>		
Reform - High Corruption	0.065 (0.262)	-6.187 (8.266)
Debiasing - High Corruption	-0.676** (0.285)	-10.353 (8.548)
Combined - High Corruption	-1.524*** (0.303)	-14.347* (8.459)
P-value: Debiasing = Combined - High Corrup.	0.010	0.661
P-value: Reform = Combined - High Corrup.	0.000	0.347
<i>Tests Low Corruption=High Corruption Coefficients</i>		
P-value: Reform Low = Reform High	0.267	0.126
P-value: Debiasing Low = Debiasing High	0.201	0.547
P-value: Combined Low = Combined High	0.001	0.466
Observation	1,411	1,411
LASSO controls	Yes	Yes

Notes: Robust standard errors in parentheses. “Beliefs Index” outcome is the mean between the *Belief of Corruption of the SAR* outcome and the *Probability that a SAR Agent will Accept a Bribe* outcome, which varies from 0 (not likely at all) to 10 (very likely). “Willing to Evade Taxes” outcome is a dummy variable, which was transformed according to Jann et al. (2012) in order to get the unbiased estimates of the willingness of evading taxes from the crosswise initial measure. Dummy variables were multiplied by 100 so that the interpretation of the coefficients in percentage points is direct. Controls were chosen using a double-selection lasso linear regression from the following possibilities: respondent’s sex, aggregated age groups, ideological self-placement, education level, confidence in institutions (church, congress, executive power, police), prior perceptions of corruption (administrative officials and Hondurans in general), and region. Section *Low Pre-Treatment Corruption* of the table shows the effects of our treatments for those individuals below the median pre-treatment corruption beliefs (the distribution of these beliefs can be seen in Figure 3), while also including tests that compare the equality of our treatments. Section *High Pre-Treatment Corruption* of the table shows the effects of our treatments for those individuals above the median pre-treatment corruption beliefs, while also including tests that compare the equality of our treatments. Finally, section *Tests Low Corruption=High Corruption Coefficients* compares the equality of the coefficients for each treatment between those individuals above and below the median pre-treatment corruption beliefs. P-values: ***p<0.01, **p<0.05, *p<0.1.

when exposed to credible benchmarks, while limited evidence of backfire emerges only among optimists. This suggests that backfire effects are not refuted but are contingent on prior beliefs and message framing.

5. Field Experiment

Even though our survey experiment mitigates social desirability bias through the self-reported willingness-to-pay measure, it cannot answer whether changes in beliefs and willingness to pay affect actual tax behavior. To address external validity concerns, we analyze results from a field experiment conducted by the Honduran Tax Administration (SAR) before the 2023 income tax filing deadline. Building on our survey findings, SAR randomly assigned similar treatment messages to individuals and firms subject to the income tax to test whether our experimental results would translate into tax compliance behavior.

SAR sent emails to a representative sample of 45,000 taxpayers in two rounds: three weeks and one week prior to the income tax filing deadline for fiscal year 2023 (April 30, 2024). All emails included a reminder about the filing deadline. Although SAR initially contacted 45,000 firms or individuals, only around 33,000 taxpayers submitted their tax returns during this round.

Although the government did not pre-register this experiment, and we did not participate in its execution, the evidence we analyzed from administrative records—kindly provided by the tax administration—supports the random assignment of the treatment (see Appendix Table B9). This allows us to analyze the data similarly to how researchers approach natural experiments with observational data (Dunning, 2012). The model emails used by the tax administration are available in the Appendix Section D.

5.1. Treatments

The control group received an email with information about the locations of SAR offices across Honduras, while taxpayers in the treatment group received either the “de-biasing” message or the “combined” message. This experiment deliberately excluded the “reform” treatment due to its potential to backfire—a concern highlighted by our initial survey results. The differences in the content of messages between our survey experiment and SAR’s intervention are subtle and can be found in the Appendix Section D.²³

²³In SAR’s experiment, neither group received basic information about SAR’s role in tax collection; the “de-biasing” treatment used a relatively shorter message, and the “combined” treatment did not include details about the process used to select new employees (e.g., lie detector tests).

5.2. Outcomes

The primary outcome of this analysis is the amount of income declared by taxpayers to the tax authority for the 2023 fiscal period, measured in thousands of lempiras. This variable captures the intensive margin of tax compliance—how much income individuals report—and provides a direct measure of the behavioral response to the experimental interventions. For context, the median declared income in our sample was 418,795 lempiras (37,814 USD PPP) in 2023, compared to 310,691 lempiras (29,338 USD PPP) in 2021.²⁴

In addition to this primary outcome, we also examine the extensive margin of compliance—specifically, whether individuals filed a tax return at all. We construct a binary indicator equal to one if the individual filed a personal income tax return for 2023. This outcome allows us to assess whether the treatments affected the decision to engage with the tax system in the first place. Analyzing the extensive margin is also important for ruling out selective filing (or attrition) as a potential source of bias in the intensive-margin results.

5.3. Estimation

Similarly to the approach taken in Equation (2), we estimate the following linear regression model using Ordinary Least Squares (OLS) to assess the effects of the experimental intervention on tax compliance:

$$Y_i = \alpha + \beta_1 \text{Combined}_i + \beta_2 \text{Debiasing}_i + \beta_3 X_i + \varepsilon_i \quad (4)$$

Here, Y_i represents the declared income, Combined_i and Debiasing_i are the two treatments implemented by the tax authority, and X_i denotes a control vector selected through a double-selection regression applied to the outcome.²⁵ The error term, ε_i , is estimated using Huber-White standard errors to account for potential heteroskedasticity. We report the coefficients β_1 and β_2 for each regression and include the mean of the control group as a reference point, along with the p-values to test whether the coefficient for the “combined” treatment differs from that of the “debiasing” treatment. We excluded the top 1% of payers to increase the power of our analysis, as the distribution is highly right-skewed, approximating a log-normal distribution and therefore substantially reducing the

²⁴These PPP-adjusted income figures refer to formal taxpayers with income tax filing obligations, who constitute a higher income group than the general population. As such, they do not reflect median national income.

²⁵The potential controls, ultimately selected or excluded by lasso, encompassed taxpayers’ baseline characteristics, including their region, their risk of non-compliance level, their type (natural or legal entity), the source of information for inclusion in the sample (e.g., previous tax returns, third-party information), and their size.

likelihood of detecting effects, following what was done in other experiments with a high presence of outliers (De Mel et al., 2008; Blattman et al., 2014).

We estimate Equation (4) on two samples: the full set of taxpayers and a subsample classified as high risk for non-compliance, according to the SAR’s internal *Modelo de Gestión de Riesgo*. This model, used to guide enforcement strategy, flags individuals and firms based on discrepancies between declared income and third-party data, as well as deviations from behavioral norms within comparable taxpayer groups (for further details see De Gouvea Scot de Arruda et al. 2022).²⁶

Although risk classification is not a direct proxy for attitudes toward the state, prior research shows that low institutional trust is consistently associated with higher levels of tax evasion and under-reporting (Luttmer and Singhal, 2014; Besley and Dray, 2024). Given our theoretical focus on belief updating and the importance of priors, the high-risk group is substantively relevant: these taxpayers may be more likely to hold skeptical views of the tax authority and thus may be more sensitive to belief-updating interventions. If the “combined” message succeeds in shifting priors about competence and integrity, we expect the largest behavioral effects to emerge among those whose past behavior already signals institutional distrust. Analyzing this group allows us to assess whether persuasion effects extend to the taxpayers most relevant to the state’s revenue base—and most likely to discount official narratives absent belief recalibration.

5.4. Main Results

As shown in Table 4, the “combined” treatment proved to be effective, resulting in an increase of 180 thousand lempiras (16 thousand USD) in the reported income (p-value < 0.1), representing a 7 percent increase relative to the control group. The impact is more pronounced for the subsample including only high-risk taxpayers, with an increase of 875 thousand lempiras (79 thousand USD) in taxable income, as shown in column (2). On the other hand, the “debiasing” treatment shows no statistically significant impact across specifications. The p-values testing equality between the effects of the “debiasing” and “combined” treatments suggest that the differences are significant in all cases (although only at a 10% significance level). These findings align with our hypothesis that the “combined” treatment reinforces the “debiasing” treatment without negatively influencing it through a priming or confirmation bias effect, as happened with the “reform” treatment in the survey experiment (Table 2).²⁷

²⁶While the full specifications of the model are confidential, it serves as a central component of SAR’s audit infrastructure.

²⁷Results without controls are shown in Appendix Table B10.

Table 4: WTP Outcome - Field Experiment

Variables	(1) Reported Income	(2) Reported Income	(3) Filled Tax Declaration	(4) Filled Tax Declaration
<i>Results in Lempiras or p.p., by outcome</i>				
Debiasing	-33.48 (99.93)	3.00 (423.44)	-0.34 (0.45)	-0.72 (0.96)
Combined	180.43* (104.60)	874.58** (443.95)	-0.54 (0.46)	-0.71 (0.95)
Control Group Mean	2,639	7,614	74.25%	86.70%
<i>Results in USD PPP</i>				
Debiasing	<i>-3.02 (9.02)</i>	<i>0.27 (38.23)</i>		
Combined	<i>16.29* (9.45)</i>	<i>78.97** (40.09)</i>		
Control Group Mean (USD PPP)	<i>238.3</i>	<i>687.5</i>		
Observations	32,938	6,142	44,489	7,113
LASSO controls	Yes	Yes	Yes	Yes
Low-Risk of Payment Excluded	No	Yes	No	Yes
P-value: Debiasing = Combined	0.074	0.086	0.701	0.992

Notes: Robust standard errors in parentheses. Observations above the 99th percentile of the declared income were excluded from the regressions. “Reported Income” is the amount of declared income subject to the income tax corresponding to the 2023 period, divided by 1,000, while “Filled Tax Declaration” outcome is a dummy variable equal to 100 if the respondent filled his income declaration. Dummy variables were multiplied by 100 so that the interpretation of the coefficients in percentage points is direct. Results in italics are expressed in 2023 USD PPP. Columns (2) and (4) exclude low risk taxpayers. Controls were chosen using a double-selection lasso linear regression from the following possibilities: region, taxpayer risk level, type of taxpayer (natural or legal entity), the source of information for inclusion in the sample (e.g., previous tax returns, third-party information), and taxpayer size. The last row presents the p-value for the test of coefficient equality between the “debiasing” treatment with that for the “combined” treatment. P-values: ***p<0.01, **p<0.05, *p<0.1.

In addition, columns (3) and (4) of Table 4 examine the impact of the treatments on the extensive margin—that is, whether individuals submitted an income tax return. Across both the full sample and a restricted sample that excludes individuals with a low ex-ante probability of tax liability, we find no statistically significant effects. This suggests that the “combined” treatment did not expand the pool of filers, but instead influenced behavior among those already inclined to comply. This null result is informative for two reasons. First, it helps rule out treatment-induced attrition as a threat to the internal validity of our main (intensive-margin) estimates. Second, it suggests that while the treatment can shape how individuals report, it does not meaningfully affect whether they choose to participate in the tax system. This pattern may reflect high baseline filing

rates, more rigid beliefs or structural barriers at the filing stage, or lower message salience among non-filers.

Overall, the average treatment effects of the communication intervention on actual tax behavior are consistent with results stemming from our self-reported willingness to pay outcome of Table 1, thus reinforcing the validity of our survey experiment measures as proxies for the willingness to evade taxes in the real world.

6. Conclusions and Policy Implications

This study shows that in low-trust contexts, the main obstacle to effective anticorruption communication is often not backlash, but disbelief. Drawing on evidence from a survey and a field experiment in Honduras, we find that well-designed informational messages—especially when structured to first improve pessimistic priors and then highlight progress—can meaningfully shift perceptions, influence tax compliance behavior and, even more, translate into real-world behavioral outcomes by increasing declared income among taxpayers.

While we find some suggestive evidence of backlash in specific subgroups, particularly among less pessimistic respondents, our overall results are more consistent with recent findings from political psychology and behavioral science. They suggest that strong backfire effects are relatively uncommon. The greater challenge lies in limited receptivity: messages that contradict deeply held beliefs about institutional failure are often discounted or ignored. By contrast, messages that start by reshaping what people view as “normal” or possible are more likely to be persuasive.

Our findings carry significant implications for policymakers. In low-trust settings, exaggerated beliefs about corruption and skepticism about government intentions can cause anticorruption messages to fall flat or be dismissed altogether. These challenges are particularly acute in developing countries, where weak institutions and widespread tax evasion undermine governance. For instance, an estimated 25% of total sales in African and Latin American countries go unreported for tax purposes, compared to just 7% in OECD countries (Goyette and Kouame, 2018). Addressing pessimistic beliefs can help rebuild trust in tax administrations, foster voluntary compliance, and strengthen fiscal capacity.

Effective communication in anticorruption campaigns thus requires more than moral appeals or generic calls to action. Our study focuses on a specific and policy-relevant form of messaging: informational signals intended to update beliefs about public sector institutional integrity. By carefully combining these messages—first debiasing, then sig-

naling progress—governments can induce more positive views of institutional competence and reduce tendencies toward tax evasion or corruption.

We do not claim to evaluate the full spectrum of anticorruption messaging, but by identifying how messages interact with prior skepticism—and showing that neutral benchmarks can mitigate disbelief—this study contributes to a growing literature on strategic communication in low-trust environments. Anticorruption campaigns, in this view, are not just about delivering facts, they are about resetting expectations.

References

- Acemoglu, D. and Verdier, T. (2000). The choice between market failures and corruption. *American Economic Review*, 91(1):194–211.
- Agerberg, M. (2022). Messaging about corruption: The power of social norms. *Governance*, 35(3):929–950.
- Ajzenman, N. (2021). The power of example: Corruption spurs corruption. *American Economic Journal: Applied Economics*, 13(2):230–257.
- Barabas, J. (2004). How deliberation affects policy opinions. *American Political Science Review*, 98(4):687–701.
- Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, 182:104123.
- Baum, A., Gupta, S., Kimani, E., and Tapsoba, S. (2024). Citizens perceptions of tax authorities and tax efficiency. *IMF Working Papers*, 2024(234):1.
- Beesley, C. and Hawkins, D. (2022). Corruption, institutional trust and political engagement in Peru. *World Development*, 151:105743.
- Besley, T. and Dray, S. (2024). Trust and state effectiveness: The political economy of compliance. *The Economic Journal*, 134(662):2225–2251.
- Blair, G., Littman, R., and Paluck, E. L. (2019). Motivating the adoption of new community-minded behaviors: An empirical test in Nigeria. *Science Advances*, 5(3):eaau5175.
- Blattman, C., Fiala, N., and Martinez, S. (2014). Generating skilled self-employment in developing countries: Experimental evidence from Uganda. *The Quarterly Journal of Economics*, 129(2):697–752.
- Boudreau, C. and MacKenzie, S. A. (2014). Informing the electorate? How party cues and policy information affect public opinion about initiatives. *American Journal of Political Science*, 58(1):48–62.
- Broockman, D. E. and Kalla, J. L. (2023). When and why are campaigns’ persuasive effects small? Evidence from the 2020 US presidential election. *American Journal of Political Science*, 67(4):833–849.

- Bullock, J. G. (2009). Partisan bias and the bayesian ideal in the study of public opinion. *The Journal of Politics*, 71(3):1109–1124.
- Bullock, J. G. (2011). Elite influence on public opinion in an informed electorate. *American Political Science Review*, 105(3):496–515.
- Bullock, J. G., Gerber, A. S., Hill, S. J., and Huber, G. A. (2015). Partisan bias in factual beliefs about politics. *Quarterly Journal of Political Science*, 10(4):519–578.
- Cheeseman, N. and Peiffer, C. (2022). The curse of good intentions: Why anticorruption messaging can encourage bribery. *American Political Science Review*, 116(3):1081–1095.
- Cheeseman, N. and Peiffer, C. (2023). Why efforts to fight corruption can undermine the social contract: Lessons from a survey experiment in Nigeria. *Governance*, 36(4):1045–1061.
- Chong, A., De La O, A. L., Karlan, D., and Wantchekon, L. (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.
- Christensen, L. (2023). Optimal persuasion under confirmation bias: Theory and evidence from a registered report. *Journal of Experimental Political Science*, 10(1):4–20.
- Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., and Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, 1(1):3–15.
- Corbacho, A., Gingerich, D. W., Oliveros, V., and Ruiz-Vega, M. (2016). Corruption as a self-fulfilling prophecy: Evidence from a survey experiment in Costa Rica. *American Journal of Political Science*, 60(4):1077–1092.
- Cotter, R. G., Lodge, M., and Vidigal, R. (2020). The boundary conditions of motivated reasoning. In et al., E. S., editor, *The Oxford Handbook of Electoral Persuasion*, pages 66–87. Oxford University Press.
- Daniele, G., Aassve, A., and Le Moglie, M. (2023). Never forget the first time: The persistent effects of corruption and the rise of populism in Italy. *The Journal of Politics*, 85(2):468–483.
- De Gouvea Scot de Arruda, T., Del Carmen, G., and Espinal Hernandez, E. E. (2022). Targeting in tax compliance interventions: Experimental evidence from honduras. Policy research working paper, World Bank Group, Washington, D.C. World Bank Policy Research Working Paper 9967.
- De Mel, S., McKenzie, D., and Woodruff, C. (2008). Returns to capital in microenterprises: Evidence from a field experiment. *The quarterly journal of Economics*, 123(4):1329–1372.
- Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge University Press.

- Dunning, T., Monestier, F., Piñeiro-Rodríguez, R., Rosenblatt, F., and Tuñón, G. (2025). Disrupting compliance: The impact of a randomized tax holiday in uruguay. *The Journal of Politics*, 87(3):000–000.
- Enríquez, J. R., Larreguy, H., Marshall, J., and Simpser, A. (2025). Accountability under polarization. Available at SSRN: <https://ssrn.com/abstract=5161375> or <http://dx.doi.org/10.2139/ssrn.5161375>.
- Epley, N. and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40.
- Erlich, A. and Gans-Morse, J. (2025). Can norm-based information campaigns reduce corruption? *American Journal of Political Science*, pages 1–20.
- Fisman, R. and Svensson, J. (2007). Are corruption and taxation really harmful to growth? Firm level evidence. *Journal of Development Economics*, 83(1):63–75.
- Flynn, D. J., Nyhan, B., and Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.
- Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton & Company, New York.
- Gingerich, D. W., Oliveros, V., Corbacho, A., and Ruiz-Vega, M. (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.
- Goyette, J. and Kouame, W. A. K. (2018). Tax evasion in Africa and Latin America: The role of distortionary infrastructures and policies. Policy research working paper, World Bank Group, Washington, D.C. World Bank Policy Research Working Paper 8522.
- Guess, A. and Coppock, A. (2020). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 50(4):1497–1515.
- Gulino, G. and Masera, F. (2023). Contagious dishonesty: Corruption scandals and supermarket theft. *American Economic Journal: Applied Economics*, 15(4):218–251.
- Hewitt, L., Broockman, D., Coppock, A., Tappin, B. M., Slezak, J., Coffman, V., Lubin, N., and Hamidian, M. (2024). How experiments help campaigns persuade voters: Evidence from a large archive of campaigns’ own experiments. *American Political Science Review*, pages 1–19.
- IDB (2015). Tax Administration Institutional and Operational Strengthening. Technical report, Inter-American Development Bank.
- Iyengar, S. and Kinder, D. R. (2010). *News that matters: Television and American opinion*. University of Chicago Press.

- Jann, B., Jerke, J., and Krumpal, I. (2012). Asking sensitive questions using the cross-wise model: An experimental survey measuring plagiarism. *Public Opinion Quarterly*, 76(1):32–49.
- Köbis, N. C., Troost, M., Brandt, C. O., and Soraperra, I. (2022). Social norms of corruption in the field: Social nudges on posters can help to reduce bribery. *Behavioural Public Policy*, 6(4):597–624.
- Köbis, N. C., van Prooijen, J.-W., Righetti, F., and Van Lange, P. A. (2015). “Who Doesn’t?”—The Impact of Descriptive Norms on Corruption. *PLOS ONE*, 10(6):e0131830.
- Levi, M., Sacks, A., and Tyler, T. (2009). Conceptualizing legitimacy, measuring legitimating beliefs. *American Behavioral Scientist*, 53(3):354–375.
- Little, A. T., Schnakenberg, K. E., and Turner, I. R. (2022). Motivated reasoning and democratic accountability. *American Political Science Review*, 116(2):751–767.
- Luttmer, E. F. P. and Singhal, M. (2014). Tax morale. *Journal of Economic Perspectives*, 28(4):149–68.
- Martin, L. E. (2023). *Strategic Taxation: Fiscal Capacity and Accountability in African States*. Oxford University Press.
- Mauro, P. (1995). Corruption and growth. *The Quarterly Journal of Economics*, 110(3):681–712.
- Nicholson, S. P. (2012). Polarizing cues. *American Journal of Political Science*, 56(1):52–66.
- Nyhan, B. (2020). Facts and myths about misperceptions. *Journal of Economic Perspectives*, 34(3):220–236.
- Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15):e1912440117.
- Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.
- Okunogbe, O. and Tourek, G. (2024). How can lower-income countries collect more taxes? The role of technology, tax agents, and politics. *Journal of Economic Perspectives*, 38(1):81–106.
- Olken, B. A. (2006). Corruption and the costs of redistribution: Micro evidence from Indonesia. *Journal of Public Economics*, 90(4):853–870.
- Peiffer, C. (2020). Message received? Experimental findings on how messages about corruption shape perceptions. *British Journal of Political Science*, 50(3):1207–1215.
- Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

- Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature*, 57(4):904–54.
- Taber, C. S., Cann, D., and Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, 31:137–155.
- Taber, C. S. and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769.
- Wood, T. and Porter, E. (2019). The elusive backfire effect: Mass attitudes’ steadfast factual adherence. *Political Behavior*, 41(1):135–163.
- Yamou, T., Thomas, A. H., and Cai, K. (2024). Citizens perceptions of tax authorities and tax efficiency. *IMF Working Papers*, 2024(234):1.
- Zaller, J. (1992). The nature and origins of mass opinion. *Cambridge University*.