Northwestern | INSTITUTE FOR POLICY RESEARCH

RESEARCH EXCELLENCE • POLICY IMPACT

IPR Working Paper Series

WP-23-33

Can Norm-Based Information Campaigns Reduce Corruption?

Aaron Erlich
McGill University

Jordan Gans-Morse
Northwestern University and IPR

Version: August 8, 2023

DRAFT

Abstract

Can norm-based information campaigns reduce corruption? Such campaigns use messaging about how people typically behave (descriptive norms) or ought to behave (injunctive norms). Drawing on survey and lab experiments in Ukraine, Erlich and Gans-Morse unpack and evaluate the distinct effects of these two types of social norms. Four findings emerge: First, injunctive-norm campaigns produce consistent but modest, temporary effects. These may serve as moderately effective, low-cost anti-corruption tools but are unlikely to inspire large-scale behavioral transformations. Second, contrary to recent studies, the researchers find no evidence that either type of norm-based messaging "backfires" by inadvertently encouraging corruption. Third, descriptive-norm campaigns emphasizing corruption's decline produce relatively large and long-lasting effects — but only among subjects who find these messages credible. Fourth, both types of norm-based messaging have a substantially larger effect on younger citizens. These findings have broader implications for messaging campaigns, especially those targeting social problems that, like corruption, require mitigation of collective action dilemmas.

The authors thank Taylor Boas, Donald Bowser, John Bullock, James Druckman, Daniel Gingerich, Henry Hale, Daniel Hidalgo, Kelly McMann, Daniel Molden, Simeon Nichter, Daniel O'Keefe, Jason Seawright, Joshua Tucker, and Susanne Wengle, as well as participants in Northwestern University's Institute for Policy Research Colloquium, the University of Virginia's Quantitative Collaborative speaker series, the 2022 Midwest Eurasian Political Economy Workshop, the 2018 Annual PONARS Eurasia Policy Conference, and the "Politics and Corruption" panel at the 2017 Annual Meeting of the American Political Science Association, for insights and advice. They also thank Ian Woodward, Yevhen Barshchevskyi, and Liudmyla Struk for facilitating data collection for Study 1; Dmitry Roy for overseeing onsite data collection for Study 2; and Annie Chen, Anna Marukhnyak, and Maya Novak-Herzog for outstanding research assistance on Study 3.

Corruption creates barriers to entrepreneurial activity, reduces the quality of public services, and undermines political institutions' legitimacy (Fisman and Golden 2017, ch. 4; Olken and Pande 2012, 491–495; Svensson 2005, 36–39). But while corruption's harms are well understood, we know less about how to reduce it. Curtailing corrupt practices is particularly challenging when corruption is widespread. In such societies, even well-intentioned citizens face incentives to pay bribes, hindering efforts to disrupt entrenched expectations and social norms.

Educational and informational campaigns provide a potential tool for transforming social norms. As early as the 1970s, Hong Kong's Independent Commission Against Corruption (ICAC) employed campaigns to combat corruption (Klitgaard 1988). However, only after the United Nations Convention Against Corruption codified the importance of raising public awareness in 2004 did anti-corruption billboards, posters, and television advertisements become ubiquitous in many countries (Peiffer 2020, 1207). In Ukraine alone, the site of our studies, at least three organizations initiated campaigns as part of anti-corruption efforts launched after the 2014 Revolution of Dignity.[1]

Do anti-corruption information campaigns reduce corruption? Rigorous studies have only begun to emerge, and the available evidence is unpromising. Indeed, a high-profile study by Cheeseman and Peiffer (2022) finds that anti-corruption messaging may *increase* citizens' willingness to offer bribes.

This article offers a more nuanced and optimistic appraisal of anti-corruption information campaigns and emphasizes the need to re-conceptualize their role: While campaigns are unlikely to produce large-scale transformations in entrenched beliefs and expectations, they can be an effective tool for modifying short-term intentions and choices about ethical behavior. To be effective, however, requires those implementing campaigns to devote attention to messaging's content, credibility, location, timing, and target audience.

We base these conclusions on three pre-registered experiments conducted between 2017

1. Author interviews in Kyiv with representatives of Transparency International Ukraine (March 21, 2017) and the UNDP (April 28, 2017). The National Anti-Corruption Bureau of Ukraine also initiated campaigns.

and 2021 in Ukraine: a survey experiment embedded in a nationally representative survey (Study 1), a laboratory experiment with university students at a top legal academy (Study 2), and a multi-wave large-scale survey experiment with subjects recruited via Facebook (Study 3). Our multi-study approach demonstrates our findings' consistency and robustness in the larger, more representative survey experiment samples and when using incentivized behavioral measures of corruption in the more controlled laboratory experiment setting.

Our studies focus specifically on "norm-based messaging": information campaigns that employ messages about how people typically behave or ought to behave. Research has shown norm-based messaging's efficacy in public health and environmental campaigns (e.g., Schultz et al. 2007, 429), but only recently have social norm frameworks attracted anti-corruption practitioners' attention.[2] Our research design unpacks the distinct impacts of injunctive and descriptive social norms. The former are norms prescribing what people ought to do; the latter are norms referring to what people typically do (Cialdini et al. 1990; Tankard and Paluck 2016). Study 3, our most comprehensive study, integrates a 2×3 factorial design that manipulates exposure to combinations of injunctive and descriptive-norm messages with a repeated measures design examining treatment effects' durability. Studies 1 and 2 employ a subset of Study 3's six experimental arms. All studies used appropriate tools to mitigate potential social desirability bias concerns, but we also emphasize that bribe-giving in contexts with widespread corruption is a far-from-taboo subject about which many citizens speak freely and frequently.

We highlight four main findings. First, we find consistent yet modest, short-lived effects for all treatments emphasizing injunctive norms about the importance of fighting corruption, regardless of whether they also emphasized descriptive norms. Second, contrary to Cheeseman and Peiffer (2022), we find no evidence of "backfire" effects in which anti-corruption messaging increases subjects' willingness to bribe, even for messages or subgroups often considered susceptible to such effects. Third, we find that treatments combining injunctive

2. See, for example, Scharbatke-Churck and Hathaway (2017) and Scharbatke-Church and Chigas (2019).

norms with descriptive norms about decreasing corruption — messaging that the existing social norms literature predicts should be most effective — produce relatively large and durable effects, but only among the limited group of subjects who find these messages credible. Finally, our pre-registered analysis of heterogeneous treatment effects finds consistently larger effects among younger subjects.

Together, these findings suggest that anti-corruption messaging can be effective if attention is devoted to the scope of campaigns' ambitions and to messaging's content, location, timing, and target audience. Given that messaging's most common effect may be of small magnitude and short duration, campaigns are unlikely to catalyze shifts in the beliefs and expectations underlying deeply rooted social norms. But if strategically placed to attract attention in critical moments preceding potential bribe transactions, messaging may be an effective, low-cost tool for making injunctive norms salient at the right time and place. This could reduce bribe-giving at the margins in ways that produce large aggregate effects. Meanwhile, our findings indicate that reformers with the more ambitious goal of transforming social norms should employ messages combining injunctive norms with descriptive norms about undesirable behaviors' decline, while targeting citizens likely to perceive such messages as credible. In Ukraine, we show that these individuals are young, women, and score higher on psychological indices of susceptibility to persuasion.

Our studies contribute to an emerging literature that has produced mixed results.[3] On the one hand, several experiments find that information about corruption's prevalence may increase citizens' willingness to engage in corruption, raising concerns that anti-corruption messaging may "backfire" (Abbink et al. 2018; Cheeseman and Peiffer 2022; Corbacho et al. 2016).[4] On the other hand, several studies have demonstrated that information about low corruption levels, or society's disapproval of corruption, can reduce corrupt behavior

3. A related literature has investigated anti-clientelism messaging's effects on citizens' willingness to sell votes (see, e.g., Blattman et al. 2019; Erlich 2020; Hicken et al. 2018).

4. It should be noted that Abbink et al.'s (2018) and Corbacho et al.'s (2016) primary focus is the rise or persistence of high-corruption equilibria, not the effectiveness of anti-corruption measures.

(Agerberg 2022; Köbis et al. 2015; Köbis et al. 2019). A similar divide exists among studies evaluating anti-corruption messaging's effects on outcomes other than willingness to engage in corruption, including awareness about corruption's harms, willingness to report corruption to authorities, or levels of institutional trust.[5] In part, these countervailing findings reflect a distinction the social norms literature has long recognized: Campaigns may produce the opposite of their intended effects if they convey — perhaps inadvertently — descriptive-norm messaging about undesirable behaviors' prevalence, even if the messaging explicitly invokes an injunctive norm about abstaining from such behavior. By contrast, descriptive-norm messaging conveying the rarity of undesirable behaviors may reinforce injunctive-norm messaging, contributing to campaigns' overall effectiveness.

Our studies advance the existing literature in five ways. First, our factorial research design facilitates analysis of injunctive and descriptive norms' effects, both in isolation and in interaction with each other, whereas existing studies primarily focus on one type of norm or the other. Second, ours is among the first anti-corruption messaging studies to evaluate treatment effects' duration. Third, we offer novel analyses about how subjects' perceptions of messages' credibility may affect messaging's effectiveness. Fourth, to facilitate more targeted campaigns, we conduct the most comprehensive analysis of pre-registered heterogeneous effects hypotheses to date. Fifth, our multiple studies in a single country at different points in time allow for examining our findings' robustness to a greater degree than earlier works. Moreover, given scholars' interest in norm-based messaging as a tool for mitigating the collective action problems that undermine responses to challenges ranging from climate change to pandemics (see, e.g., Raymond et al. 2023), our findings offer insights with relevance beyond the corruption literature.

5. Peiffer and Walton (2019), Peiffer (2020), and Denisova-Schmidt et al. (2015) report either null effects or backfire effects. By contrast, Blair et al. (2019) demonstrate that messaging can increase citizens' propensity to report corruption.

Harnessing Social Norms to Fight Corruption

Briefly examining why systemic corruption proves so resistant to reform offers insights into norm-based messaging's potential for combating corruption. Recently, scholars and policy-makers have applied collective action frameworks to make sense of corruption's persistence by analyzing how individuals' propensity to act corruptly depends on beliefs about others' behavior (e.g., Persson et al. 2013; Corbacho et al. 2016; Stephenson 2020). From a collective action perspective, both the probability of punishment and the search costs of finding partners for a corrupt transaction decrease as corruption increases. Moreover, with widespread corruption, even citizens who find corruption reprehensible are compelled to participate or else face a disadvantage vis-à-vis peers who exploit bribe-facilitated shortcuts. Finally, if only a small minority of citizens eschew corruption, they are likely to perceive anti-corruption efforts as hopeless endeavors. These dynamics produce vicious cycles of expectations and social norms that become entrenched.

Norm-Based Messaging Campaigns

If expectations and social norms sustain corruption, then successful anti-corruption efforts must transform these norms and expectations. Social norms are the "shared understandings about actions that are obligatory, permitted, or forbidden within a society" (Ostrom 2000, 143–144) or, in simpler terms, "mutual expectations about the right way to behave" (Scharbatke-Church and Chigas 2019). Following Cialdini et al.'s (1990) influential work, we distinguish between *descriptive* and *injunctive* norms. Descriptive norms consist of expectations and beliefs about what others are likely to do in a given situation. They influence behavior by facilitating mental shortcuts for making decisions: In many contexts, the most efficient or effective behavior will be the one in which other people also engage. Injunctive norms consist of expectations and beliefs about what people ought to do. They influence behavior by providing cues about which types of behavior likely lead to social sanctions.

Numerous studies have demonstrated norm-based messaging's effectiveness, particularly

with respect to environmental and public health goals. Such campaigns have produced declines in binge drinking, littering, and energy consumption and increases in water conservation and recycling (e.g., Farrow et al. 2017; Lewis and Neighbors 2006). Often these campaigns convey information about descriptive norms that "reduce the occurrence of deleterious behaviors by correcting targets' misperceptions regarding the behaviors' prevalence'' (Schultz et al. 2007, 429), but ample evidence also points to the effectiveness of messages that draw attention to injunctive norms (see, e.g., Cialdini et al. 2006; Reno et al. 1993). As recognition of corruption as a collective action problem has grown, norm-based messaging has attracted practitioners' interest (Bicchieri 2016; Hoffmann and Patel 2017; Scharbatke-Church and Chigas 2019), and studies evaluating its effectiveness have begun to emerge. As with research on other policy domains, initial findings indicate that messaging about others' unwillingness to engage in corruption reduces the propensity for bribe-giving (Köbis et al. 2015; Köbis et al. 2019), and that correcting misperceptions about others' injunctive norms via messaging about widespread disapproval of corrupt behavior can have a similarly beneficial effect (Agerberg 2022).

While messaging that emphasizes injunctive norms may be broadly applicable, social norms scholars have cautioned that drawing attention to descriptive norms in societies where socially undesirable behavior is prevalent may prove counterproductive (Cialdini et al. 1990; Reno et al. 1993). In such situations, campaigns that increase descriptive norms' salience are likely to produce more of the salient behavior, creating the "backfire" or "boomerang" effects discussed below. But, whereas earlier studies suggested avoiding descriptive-norm messaging altogether when unwanted behavior is widespread, recent studies advocate for messaging about normative *trends*. Redirecting focus from a socially harmful behavior's current *levels* to the behavior's decreasing prevalence encourages individuals to anticipate how norms are likely to evolve in the future and adjust present-day behavior accordingly (Mortensen et al. 2019; Sparkman and Walton 2017). Consonant with these studies, our experiments' descriptive-norm treatments focus on corruption trends rather than levels. We

nevertheless recognize — and empirically evaluate — the possibility that a limitation to normative-trend messaging is that citizens in societies with endemic corruption may be unlikely to perceive messages about declining corruption as credible (Agerberg 2022, 932).

To summarize, existing research suggests that injunctive-norm messaging may be broadly effective and descriptive-norm messaging may be effective when undesirable behavior is rare, or when focus is diverted to trends about harmful behaviors' decline. What, however, should be expected when campaigns convey messages about both descriptive and injunctive norms? In real-world campaigns, this blending of messages occurs frequently, even if it is often inadvertent. According to the social norms literature, which norm will affect behavior depends on the degree to which situational context "activates'' (i.e., makes salient) each type of norm (Cialdini et al. 1990). It follows that campaigns should either focus audiences' attention on norms that are most likely to produce the intended effect or should combine injunctive and descriptive norms with care. Consider, for example, an anti-corruption campaign conveying a straightforward injunctive norm such as, "We must fight corruption!" A seemingly trivial addition of the type commonly employed in anti-corruption campaigns to raise awareness or add urgency — e.g., "*Bribery is increasing!* We must fight corruption!" — presents descriptive-norm information at odds with the injunctive norm. Despite the explicit appeal to not engage in corruption, the message makes clear that an increasing number of fellow citizens do, in fact, act corruptly, an implicit recognition that bribery is to some extent socially acceptable. At best, countervailing norms might dampen each other's influence, leading to no aggregate effects; at worst, the counterproductive descriptive norm's influence may overwhelm the injunctive norm's intended effects and increase undesirable behavior (Cialdini et al. 2006; Schultz et al. 2007).

According to social psychologists, optimal norm-based messaging consists of well-aligned descriptive and injunctive-norm messages that "work in tandem rather than in competition with one another," thereby "conveying to recipients that the desired activity is widely performed and roundly approved, whereas the unwanted activity is relatively rare and roundly

Table 1: Relative Effectiveness of Injunctive and Descriptive-Norm Messaging Predicted by the Social Norms Literature

Type of Message	Predicted Effectiveness
Injunctive + Descriptive Norm About ↓ Corruption	Most effective
Injunctive Norm	Effective
Descriptive Norm About ↓ Corruption	Effective
Injunctive + Descriptive Norm About ↑ Corruption	Ineffective, risk of backfire effect
Descriptive Norm About ↑ Corruption	Most likely to backfire

disapproved. Such a line of attack unites the power of two independent sources of normative motivation and can provide a highly successful approach to social influence'' (Cialdini et al. 2006, 13; see also Smith et al. 2012). In the anti-corruption context, a statement conveying some initial success (e.g.,"*Bribery is decreasing!* We must fight corruption!"), offers a descriptive-norm message that reinforces the injunctive norm.

The predictions summarized in Table 1 draw on the social norms literature's insights. We focus on high-corruption contexts, as anti-corruption campaigns usually are launched in countries facing systemic corruption. First, messages combining an injunctive norm with a descriptive norm about decreasing corruption are most likely to reduce corrupt behavior. Second, while the existing literature does not offer generalizable predictions about injunctive-norm messages' effectiveness relative to descriptive-norm messages about decreasing corruption, citizens are unlikely to perceive information about declining corruption as credible when corruption is widespread. Accordingly, injunctive-norm messaging might be preferable. Third, combining injunctive-norm messaging with clashing descriptive-norm messaging about increasing corruption is likely to dampen both messages' effects. Finally, descriptive-norm messaging focused exclusively on increasing corruption is not only unlikely to prove effective, but may run the risk of backfiring.[6]

The Risk of "Backfire" Effects

That information campaigns to curtail socially harmful behavior can potentially "backfire" or "boomerang" and produce the opposite of intended effects has been documented in the

6. Online Appendix H discusses how these predictions map to our three studies' pre-analysis plans.

social norms literature (Cialdini et al. 2006; Schultz et al. 2007).[7] As indicated above, the underlying mechanisms are also well-understood: They likely result from the inadvertent activation of descriptive norms conveying the prevalence of the unwanted behavior. The problem emanates from what Cialdini et al. (2006, 4) describe as the "understandable but misguided tendency of public officials to try to mobilise action against socially disapproved conduct by depicting it as regrettably frequent, thereby inadvertently installing a counter-productive descriptive norm in the minds of their audiences."

These insights are especially concerning for anti-corruption efforts. Since anti-corruption organizations' rise in the mid-1990s, and particularly following the 2004 United Nations Convention Against Corruption, anti-corruption efforts have focused on drawing attention to corruption's prevalence. As Cheeseman and Peiffer (2022, 1081) note, the widely accepted logic is that "because graft 'lives in the shadows,' a critical step to fighting corruption is to illuminate and bring popular awareness to it." This "knowledge-deficit" model of behavior assumes that individuals would reduce their role in harmful activity, and potentially work to pressure others to do so as well, if only they had sufficient information. Little evidence supports this behavioral change model, in part because it overlooks the motivations and incentives underlying harmful behavior (Schultz et al. 2007, 5).

While several earlier studies offered evidence that information about bribery's prevalence or increasing levels makes subjects themselves more willing to act corruptly (Abbink et al. 2018; Corbacho et al. 2016), Cheeseman and Peiffer's (2022) influential recent article suggests that backfire-effect risks are even greater than the social norms literature posits. Based on a survey and experimental game conducted in Lagos, Nigeria, they find backfire effects not only from a descriptive-norm message emphasizing corruption's prevalence but also from other messages unrelated to social norms. Though some evidence of backfire effects emerges in their full sample, these effects are particularly pronounced among citizens with strong prior beliefs about corruption's prevalence, a group they refer to as "pessimistic

7. A distinct literature investigates "backfire" effects from efforts to correct individuals' misperceptions, though these effects are rarer than initially assumed (see Guess and Coppock 2020; Wood and Porter 2019).

perceivers." They argue that backfiring occurs because nearly *any* message about corruption — not just messages predicted by the social norms literature to lead to backfire effects — "primes" such subjects and counterproductively makes salient their beliefs about corruption's ubiquity. Considering that beliefs about corruption's prevalence are widespread, they warn that "traditional anticorruption campaigns may increase the willingness of most of the population to pay bribes" (Cheeseman and Peiffer 2022, 1090).

The risk of unintended consequences undoubtedly deserves attention, but the existing evidence should be properly contextualized. First, the social norm treatments employed by Abbink et al. (2018), Cheeseman and Peiffer (2022), and Corbacho et al. (2016) emphasize only descriptive norms about corruption's prevalence or rising rates. While awareness-raising efforts sometimes convey only such information, messaging focused on fighting corruption nearly always combines information about descriptive *and* injunctive norms. As Schultz et al. (2007) have shown in studies about water conservation, the inclusion of injunctive norm messaging even as subtle as smiley or sad faces can mitigate descriptive norms' backfire effects. Second, Cheeseman and Peiffer's (2022, pp. 1089-1091) findings are based largely on the subgroup effects they identify among citizens who believe corruption to be widespread. Yet, this subgroup hypothesis appears not to have been pre-registered, and these analyses rely on post-treatment measures of "prior beliefs" to classify pessimistic perceivers.[8] Until further evidence emerges, it may be premature to conclude that a wide range of messages, not just those that draw attention to unwanted behavior's prevalence, are likely to backfire.

Our studies nevertheless take seriously concerns about unintended effects. We both evaluate messages considered most likely to backfire and collect extensive data on pre-treatment corruption beliefs to analyze subgroups of subjects who might be particularly susceptible to boomerang effects.

8. Although Cheeseman and Peiffer (2022, 1090) report statistically insignificant results when regressing these post-treatment corruption attitude variables on their treatment indicator, Montgomery et al. (2018, 772–773) have shown such empirical tests to be insufficient for diagnosing post-treatment bias.

Making Messaging Campaigns More Targeted

Contemporary political advertising employs fine-grained data to engage in targeted, customized messaging via online platforms (Wakefield 2018). Commercial advertisers deploy similarly sophisticated targeting (Zuboff 2015). Yet, many anti-corruption messaging campaigns continue to operate under the potentially dangerous assumption that any campaign raising awareness about corruption has value,[9] rather than adopting targeted approaches. To offer insights into how anti-corruption messages might better target those most likely to be positively affected, we examine three sets of variables: prior beliefs, psychological traits, and demographic characteristics. Given our results below primarily emphasize our finding about the importance of age, we offer an abbreviated discussion of subgroup analyses and provide a more comprehensive overview in Online Appendix D.

As Cheeseman and Peiffer's (2022) previously discussed findings suggest, prior beliefs about corruption's prevalence may affect individuals' reactions to anti-corruption messages. The same message may increase a behavior among citizens who had previously underestimated this behavior's prevalence while decreasing the behavior among those who had overestimated its prevalence (Schultz et al. 2007, 430). We accordingly collected data on respondents' pre-treatment beliefs about corruption trends and about the percentage of other citizens who give bribes. Additionally, given that pre-treatment exeriences can influence information experiments' effects (Druckman and Leeper 2012), we asked subjects to rate their prior familiarity with anti-corruption campaigns.[10]

To consider how psychological traits might affect susceptibility to social norm messaging, we employed three indices. The first, Kaptein et al.'s (2012) susceptibility to persuasion strategies (STPS) scale, measures individuals' likeliness of responding to each of Cialdini's (2001) six influence strategies. We utilize scale items pertaining specifically to social norm

9. In Ukraine, for example, advertising agencies developed some anti-corruption campaigns *pro bono*, with little input from corruption experts. Other campaigns were created via crowdsourcing. Author interviews in Kyiv with representatives of TI-Ukraine (March 21, 2017) and the UNDP (April 28, 2017).

10. To mitigate the risk of priming effects, we separated these corruption beliefs questions from the survey's experimental section with several long batteries of questions unrelated to corruption.

strategies and hypothesized that treatment effects should be larger for subjects with higher STPS scores. Second, we employed Lins de Holanda Coelho et al.'s (2020) six-item version of Cacioppo et al.'s (1984) Need for Cognition (NfC) scale, which measures the extent to which individuals feel a need to analyze and elaborate on information received. Following Kaptein (2012, 23–24), we hypothesized that subjects with high NfC are more likely to scrutinize arguments with which they are presented and, consequently, less susceptible to social influence strategies. Third, we presented subjects with Gosling et al.'s (2003) 10-item questionnaire for measuring Big Five personality dimensions, predicting that subjects with higher agreeableness and lower openness scores may be more susceptible to messaging.

Finally, with respect to demographic characteristics, we hypothesized that younger subjects might have more malleable beliefs and be more responsive to anti-corruption messaging. Similar assumptions about the malleability of youth motivated Hong Kong's successful anti-corruption educational campaigns targeting school-age children in the 1970s (Klitgaard 1988, 116–117). Additionally, in countries such as Ukraine, the communist collapse may have exacerbated age-based differences. Citizens who spent formative years under communism were socialized in ways that created distinct and enduring social and political attitudes (Mishler and Rose 2007; Pop-Eleches and Tucker 2017), potentially making them less responsive to anti-corruption messaging relative to those raised in the post-communist era.

Country Selection: Anti-Corruption Efforts in Ukraine

Following the 2014 Euromaidan revolution, Ukraine initiated a high-profile anti-corruption campaign, making it a fitting context for our studies. Despite some notable successes, Ukraine still ranked 122nd out of 180 countries on Transparency International's Corruption Perceptions Index in the year prior to Russia's 2022 full-scale invasion.[11] Accordingly, many of our studies' subjects likely had direct familiarity with the scenarios in the vignettes we use to measure bribe intentions, or indirect familiarity from family members' and friends' experi-

11. See www.transparency.org/cpi.

ences. Meanwhile, non-governmental organizations, government agencies, and international organizations in Ukraine have utilized anti-corruption messaging resembling the information experiments we conduct, lending our studies a significant degree of experimental realism. For example, Figure 1a shows an anti-corruption poster that at one point greeted travelers at Kyiv Borispol Airport, announcing the goal of "Ukraine without corruption!" and reminding passersby that bribery "is a criminal offense."[12] Campaigns in Ukraine also illustrate how such efforts may inadvertently emphasize corruption's ubiquity, thereby enforcing descriptive norms about corruption's prevalence. Figure 1b, taken from an online university anti-corruption training course, shows Ukraine's corruption scores worsening over time.

Figure 1: Anti-Corruption Messaging in Ukraine

(a) Anti-Corruption Poster in Kyiv Airport

(b) Flyer from Anti-Corruption Training Course

Ukraine was also a fitting site for our studies because it presents a particularly tough test of anti-corruption messaging's effectiveness. As Corbacho et al. (2016, 1090) emphasize, citizens are most susceptible to the types of messages we evaluate in countries where beliefs about corruption are in flux. Their survey experiment, for instance, was conducted in Costa

12. Photo from "V 'Borispole' inostrantsam napominaeyut, chto Ukraina svobodna ot korruptsii" [At "Borispol" foreigners are reminded that Ukraine is free of corruption], ukranews.com (March 4, 2016).

Rica, a country with a longstanding reputation for integrity that recently had endured rising corruption. By contrast, where perceptions about corruption are entrenched — as in most of the former Soviet Union — it may prove difficult for informational campaigns to shift beliefs or practices. Consequently, if anti-corruption messaging proves effective in the context of deeply rooted systemic corruption, it likely would prove at least as effective in less formidable circumstances.

Research Design & Implementation

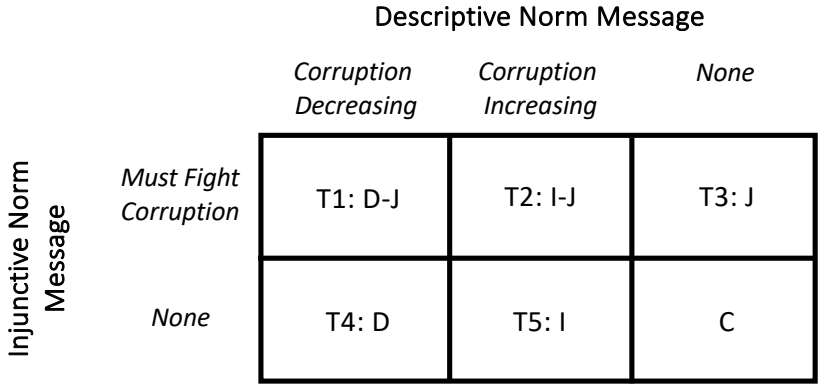
Our three studies employed a similar research design and nearly identical information treatments. We, therefore, first introduce the studies' common elements before turning to details about how each was implemented and the specific indicators each used to measure willingness to engage in corruption. All studies were approved by the authors' university Institutional Review Boards and conducted in accordance with APSA's Principles and Guidance on Human Subjects Research (see Online Appendix A).

Information Treatments

Our studies employ a 2×3 factorial design, with three levels of descriptive-norm and two levels of injunctive-norm messaging, per Figure 2. The two descriptive-norm treatments expose subjects to information about other Ukrainians' behavior (see Figure 3). Meanwhile, our injunctive-norm treatment, which exhorts citizens to "Keep fighting — stop corruption!", draws on Cialdini et al.'s (1990) influential Focus Theory of Normative Conduct.[13] Research in this tradition has shown how exhortations about what should or should not be done, or even subtle cues (e.g., smiley or frowny faces) assigned in response to behaviors such as energy consumption, elicit a focus on societal approval or disapproval. In turn, this focus raises injunctive norms' salience and increases their likeliness to shape behavior (Allcott 2011; Bhanot 2021; Cialdini et al. 2006; Schultz et al. 2007).

13. For example, Cialdini et al.'s (2006) classic study of norm-based messaging's impact on visitors' willingness to remove artifacts from a national park utilized the injunctive norm treatment "Please don't remove the petrified wood from the park," along with descriptive-norm treatments such as "Many past visitors have removed the petrified wood from the park, changing the state of the Petrified Forest."

Figure 2: Factorial Design for Anti-Corruption Messaging Treatments



Our first study, a survey experiment embedded in a representative national sample of Ukrainians, included treatments $T1$ and $T2$, which combine messages about injunctive and descriptive norms, and a control group. Subjects in treatment arm D-J were shown anti-corruption flyers conveying an injunctive norm invoking the need to fight corruption and a descriptive norm emphasizing decreasing corruption. Subjects in treatment arm I-J were shown the same injunctive norm but with a descriptive norm emphasizing increasing corruption. These two treatments are most directly relevant for real-world anti-corruption campaigns, which nearly always invoke a combination of injunctive and descriptive norms.

Our second study consisted of an information experiment conducted as part of a laboratory bribery game. This study's smaller sample size required us to focus on a single treatment and a control group to maintain sufficient statistical power. Per our theoretical expectations that the reinforcing descriptive and injunctive-norm messages of Treatment D-J should prove most effective — an expectation that, as discussed below, was supported by Study 1 — we focused on treatment $T1$.

For our third study, we recruited a large sample of Ukrainians via Facebook to comprehensively examine the full set of five treatment arms in Figure 2. In addition to treatments D-J and I-J, Study 3 included treatment arms focused exclusively on the injunctive norm (Treatment J) or the descriptive norm messages (Treatments D and I). As discussed below, Study 3 combined this 2×3 factorial design with a repeated measures design, in which

Table 2: Treatment Arms Included in Each Study

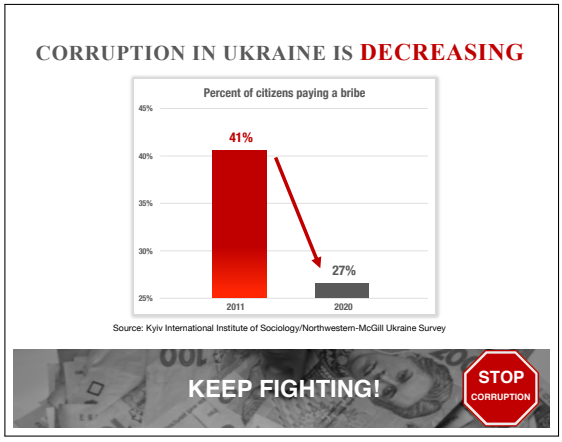
	Decreasing-Injunctive (D-J)	Increasing-Injunctive (I-J)	Decreasing (D)	Increasing (I)	Injunctive (J)	Control (C)
Study 1	✓	✓				✓
Study 2	✓					✓
Study 3 ($W1$)	✓	✓	✓	✓	✓	✓
Study 3 ($W2$)	✓					✓

we deployed a second survey wave to analyze treatment effects' duration and the potential effects of repeated exposure to the treatment posters. Given attrition concerns, the second wave focused again, as in Study 2, only on the treatment expected to have the largest effect, Treatment D-J. Table 2 provides an overview of the treatments employed in each study.

Figure 3 shows English-language versions of our treatment messages.[14] Treatments D-J and I-J have identical formatting and emphasize the same injunctive norm: an anti-corruption message at the bottom of the flyer calling on citizens to "Keep fighting — stop corruption." They differ in that Treatment D-J calls citizens to action by emphasizing success in fighting corruption (i.e., a descriptive norm conveying that others are engaging in less bribery). By contrast, Treatment I-J invokes action by emphasizing that corruption is spreading (i.e., a descriptive norm conveying that others are engaging in more corruption). Treatments D and I show the same descriptive-norm information used in Treatments D-J and I-J but without an injunctive norm message; Treatment J shows a larger version of the injunctive norm message from Treatments D-J and I-J but without a descriptive norm message. None of the treatment flyers employs deception. They rely on real data but draw on different time periods and surveys to emphasize distinct messages.

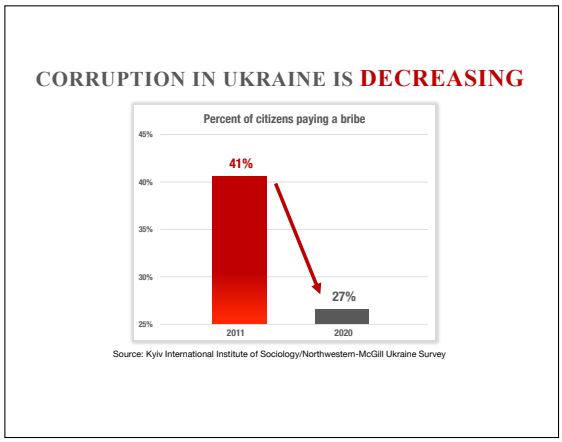
14. Figure 3 shows our Study 3 treatment flyers. Flyers for the first two studies were nearly identical. Appendix G.1 includes links to Ukrainian and Russian versions.

Figure 3: Anti-Corruption Information Treatments

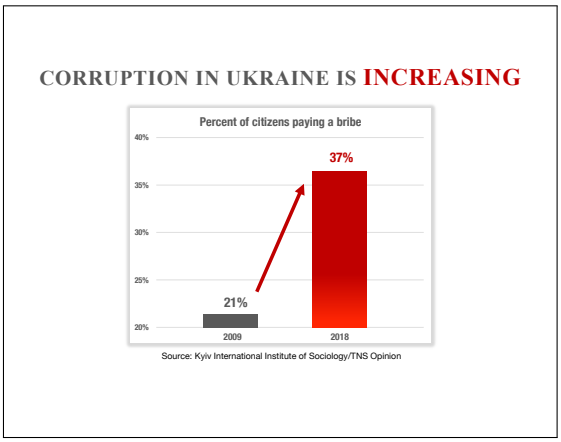


(a) Treatment D-J

(b) Treatment I-J



(c) Treatment D



(d) Treatment I

(e) Treatment J

Implementation

Study 1

Our pre-registered Study 1 experiment was embedded in a large, nationally representative survey of 6,926 Ukrainians conducted face-to-face by the Kyiv International Institute of Sociology (KIIS) between July 3 and July 31, 2017.[15] Respondents chose to participate in either Ukrainian or Russian. Subjects were assigned with equal probability to Treatment *D-J*, Treatment *I-J*, or a control group.[16] Interviewers handed treatment group respondents a flyer and asked them to familiarize themselves with the image's key messages. No time limit was placed on this process. On average, respondents examined flyers for 48 seconds.

Given the challenge of observing real-world bribes, especially in a context conducive to experimental manipulation, Studies 1 and 3 followed the approach used in other recent anti-corruption messaging studies (e.g., Agerberg 2022; Corbacho et al. 2016) and employed an outcome variable measuring intention to bribe. Study 1's measure of intention to bribe is based on respondents' binary "yes" or "no" answers to the question: "Would you be willing to pay a bribe to receive a driver's license more quickly or easily?"[17] We focused on driver's licenses for three reasons. First, most Ukrainians have personal experience obtaining a license. Second, in years prior to our study, Ukrainians regularly rated the agency that grants licenses among Ukraine's most corrupt institutions (KIIS 2015, 33). Third, bribery's use to obtain or expedite receipt of driver's licenses figures prominently in the corruption literature (see, e.g., Bertrand et al. 2007; Ryvkin and Serra 2020).

To assess potential risks to inference posed by social desirability bias, we additionally estimated intent to bribe using a crosswise model, a sensitive survey technique for eliciting

15. The survey included booster samples for several cities and regions overlaid on a representative sample. Per Miratrix et al. (2018), we use unweighted data in our analyses below but show in Online Appendix F.5 that results are nearly identical when using population weights.

16. Online Appendix F.1 provides evidence that randomization achieved balance for all three studies.

17. Our approach builds on Corbacho et al. (2016), but we use a driver's license in place of their traffic police scenario because Ukraine at the time of the study had recently undertaken ambitious police reforms.

honesty (Tan et al. 2009). Online Appendix F.3 shows that the crosswise model and direct question produced similar and statistically indistinguishable estimates.

Study 2

Study 2 consisted of a pre-registered experimental laboratory bribery game. Although a laboratory experiment necessitated a smaller sample size and, accordingly, limited the number of treatment arms we could include, it facilitated the use of a behavioral measure similar to that utilized in several other recent anti-corruption messaging studies (Cheeseman and Peiffer 2022; Köbis et al. 2015; Köbis et al. 2019). Such games use incentive payments to induce subjects to reveal preferences via observable choices made in response to decisions with real-world implications for how much money they receive (Abbink 2006).

The study was conducted with undergraduate and master's students at a top Ukrainian legal academy from October 25 to November 3, 2017. The choice of a legal academy was motivated by an unrelated study we conducted with an overlapping student sample. With the university administration's assistance, we created a sample frame based on enrollment data and conducted stratified random sampling by class year and department. Research assistants visited classrooms and requested the optional participation of sampled students, who were informed they could earn money. When students were absent, they were replaced with the next person on the list until quotas for each department and class year were filled.[18] Research assistants then led participants to the computer lab and invited them to read the instructions on the computer screens.[19] Consonant with our pre-registration, 695 students participated. Online Appendix B.1 contains additional demographic data on the sample.

The study's bribery game builds on Barr and Serra (2010). All participants were given 35 hryvnia at the game's outset and randomly assigned to the role of citizen or bureaucrat. The citizen then was presented with a scenario in which she could receive an additional 45

18. Departmental response rates varied from 14 to 41 percent (with a mean of 27 percent). Low rates nearly always reflected student absences or incorrect enrollment information rather than refusal to participate.

19. To ensure uniformity in the experimental setting, we transported students from one off-campus department to the main campus.

hryvnia by obtaining a permit. She is automatically denied when seeking the permit but given a chance to offer a bribe ranging between 5 to 35 hryvnia to the bureaucrat. Online Appendix G.2 details how this setup models the risk of punishment incurred when engaging in bribery and the harm that corruption inflicts on society at large, the latter operationalized by two randomly chosen subjects' loss of 5 hryvnia for each instance in which a citizen offers and the bureaucrat accepts a bribe.

Payoffs were constructed such that an easy-to-calculate equilibrium exists in which the citizen can offer a bribe that, from a purely strategic perspective, will raise payoffs for both her and the bureaucrat should he accept. Both subjects, however, must consider whether they are willing to reduce other participants' earnings and to engage in an act explicitly labeled a "bribe." If the bureaucrat considers factors other than financial payoffs and rejects the citizen's offer, the citizen is strictly worse off, receiving a payoff lower than the endowment with which she began the game. Study 2's primary outcome of interest was whether an individual offers (in the role of citizen) or accepts (in the role of bureaucrat) a bribe.

Study 3

For Study 3, conducted between July 15 and September 8, 2021, we recruited 7,901 subjects via Facebook ads, consonant with our pre-registered target of 7500. With approximately two-thirds of adults as registered users, Facebook, at the time of our study, was by far Ukraine's most popular social media application, making it an appropriate tool for recruiting research participants.[20] After careful consideration, we chose Facebook over alternative platforms. MTurk lacks a sizable pool of Ukrainian workers. Lucid does not guarantee a more representative or attentive sample pool (see, e.g., Ternovski and Orr 2022) and costs approximately three times more per subject. We incentivized participation with a raffle for an Apple Watch (for wave one) and an iPad (for wave two). The survey was administered via Qualtrics, and respondents could participate in either Ukrainian or Russian.

20. In July 2021, FB Ads reported that ads reached 23.3 million unique adult (18+) users out of an adult population of 33.6 million. For FB's share of Ukraine's social media market, see: https://bit.ly/3ykYlvL.

Table 3: Experimental Conditions of Combined Factorial and Repeated Measures Designs

	Treatment at $t1$	**Treatment at** $t2$	**Target** N $(W1/W2)$
(1) *D-J* $\rightarrow$ *D-J*	*D-J*	*D-J*	1500/500
(2) *D-J* $\rightarrow$ *C*	*D-J*	*PLACEBO EMAILS*	1500/500
(3) *C* $\rightarrow$ *C*	*NONE*	*PLACEBO EMAILS*	1500/500
(4) *I-J* $\rightarrow$ *No W2*	*I-J*	–	750
(5) *D* $\rightarrow$ *No W2*	*D*	–	750
(6) *I* $\rightarrow$ *No W2*	*I*	–	750
(7) *J* $\rightarrow$ *No W2*	*J*	–	750

Note: $W1$ refers to the first survey wave; $W2$, to the second. Time Period 1 ($t1$) refers to the treatment condition during $W1$. Time Period 2 ($t2$) refers to the treatment condition during the period between $W1$ and $W2$.

To analyze treatment effects' duration and the effects of multiple exposures to messaging, Study 3 combined the 2×3 factorial design discussed above with a repeated measures design. We deployed a second survey wave ($W2$) seven days after the initial survey ($W1$). Of subjects exposed to a $W1$ treatment flyer, half were randomly assigned to receive additional exposure via email to the same flyer one day and four days after the $W1$ survey. The other half of treated subjects, as well as subjects assigned to the $W1$ control group, received placebo emails reminding them they would receive a link via email to participate in a follow-up survey seven days after the $W1$ survey. This combination of the $W1$ factorial design and repeated measures design produces a total of seven experimental arms, per Table 3.

To ensure sufficient statistical power, we limited our focus for $W2$, as in Study 2, to the messaging we expected to be most effective, Treatment *D-J*. We accordingly assigned additional subjects to this treatment arm and the control group to account for attrition between survey waves. Based on pilot tests, we expected approximately one in three subjects from eligible treatment arms to participate in the second wave.[21] To obtain our pre-registered

21. The actual rate was 38.8 percent, resulting in a sample of size of 1488 for $W2$. As shown in Online Appendix F.2, there is no evidence of differential attrition across experimental arms.

target of 1500 subjects for $W2$, divided equally among subjects exposed to Treatment D-J only once during $W1$, subjects exposed multiple times to Treatment D-J, and subjects never exposed to any treatment flyer, we set the probability of assignment to the D-J treatment in $W1$ at four times the rate of other treatment arms and the probability of assignment to the control group at two times the rate (see the Target N column in Table 3).

Using a Facebook-recruited sample allowed us to recruit a large number of subjects cost-efficiently, personally manage treatment assignments in a complex research design, and implement a two-wave survey with some of the subjects participating in wave 2 subsequent to additional treatment exposures. Although we make no claims about our sample's representativeness of the overall Ukrainian population, the sample demographics show a wide range of diversity by age, gender, and geographic location, including respondents from all of Ukraine's regional administrative units (*oblasts*) other than Russian-occupied Crimea, making us confident that our results generalize beyond a narrow slice of Ukrainians.[22]

To measure bribe intent in Study 3, we expanded on the approach used in Study 1 but employed multiple scenarios to ensure results were not contingent on specific types or sectors of corruption. In both survey waves, respondents were presented (in randomized order) with the following five questions, each with a dichotomous choice of "yes" or "no":

1. Imagine that you are driving your car when the police pull you over for speeding. In order to avoid paying a traffic ticket, would you be willing to offer the police officer a bribe?

2. Imagine you are applying for a driver's license. At the Ministry of Internal Affairs service center, you learn that the process will require a considerable wait time. In order to receive the license more quickly, would you be willing to offer an official a bribe?

3. Imagine that you have a child who is applying to university. In order to improve your child's chances of admission, would you be willing to offer a university representative a bribe?

4. Imagine that you end up in the hospital with a painful but not life-threatening ailment and learn there will be a long wait for treatment. In order to receive treatment more quickly, would you be willing to offer a doctor a bribe?

22. The results when weighting the sample to correspond with Ukrainian census data are consistent with our main results but less robust. This lesser robustness reflects one of our key findings: Messaging has a significantly larger effect on younger individuals, and younger individuals are disproportionately represented in our sample relative to the general population. Online Appendix F.5 documents that if we consider only subjects 50 years of age and under, then weighted and unweighted analyses produce nearly identical results.

5. Imagine that you need a permit to open a new business and learn that there will be a long wait. In order to receive the permit more quickly, would you be willing to offer a bribe to an official at [the registry office of] the Ministry of Justice?

Per our pre-analysis plan, after conducting factor analysis on these outcome variables to confirm they load cleanly onto a single factor, we created an additive index of the number of scenarios for which subjects indicated willingness to bribe, ranging from 0 to 5. We additionally conducted two list experiments, one focused on the traffic police scenario and one on the medical scenario. As with the crosswise model in Study 1, neither list experiment produced evidence of social desirability bias (see Online Appendix F.4).[23]

External Validity & Experimenter Demand Effects

Given the challenges of observing bribery, we follow existing anti-corruption messaging studies in our use of measures of intent (Agerberg 2022; Corbacho et al. 2016) and experimental behavioral measures (Abbink et al. 2018; Cheeseman and Peiffer 2022; Köbis et al. 2015; Köbis et al. 2019). For the measures of intent employed in Studies 1 and 3, psychology studies have produced extensive evidence that experimentally induced changes in intentions are robust predictors of behavioral changes (see, e.g., the meta-analysis in Webb and Sheeran 2006). Accordingly, even prominent messaging studies of less difficult phenomena to measure, such as vaccine uptake, regularly employ measures of "vaccine intent" (e.g., Nyhan et al. 2014; Nyhan and Reifler 2015). This approach is even more appropriate when studying sensitive and often unobservable phenomena such as corruption.

We nevertheless supplement measures of intent with the incentivized behavioral measures employed in Study 2. Evidence of these types of measures' external validity is robust: Subjects exhibiting traits ranging from dishonesty to altruism in laboratory games have been shown to exhibit similar traits in real-world decision-making (e.g., Benz and Meier 2008; Hanna and Wang 2017).

23. We turned to list experiments for Study 3 due to concerns about implementing crosswise models in a self-administered online study.

A second potential research design question pertains to experimenter demand effects, the concern that subjects seek to conform with researchers' expectations. However, Mummolo and Peterson's (2019) rigorous empirical analyses have shown that demand effects are exceedingly rare in survey experiments, particularly when conducted online, as in Study 3. In part, this is because online studies do not expose subjects to subtle cues via in-person interaction with researchers, but they also find that even explicitly exposing subjects to researchers' hypotheses rarely affects the results of online experiments. As discussed above, our Study 2 lab experiment also was implemented as an online module. Researcher-subject interaction was limited to guiding subjects to the university computer lab, and this task was performed by local university students who themselves had only minimal information about the study's content or purpose.

Results

Table 4 presents all three studies' primary results. The table's first row presents the control group means. Both Study 1's and 2's outcome variables are dichotomous; therefore, Study 1's baseline estimate represents the percentage of subjects (17%) in the control group indicating willingness to pay a bribe to expedite receipt of a driver's license. Similarly, Study 2's estimate shows that 32% of subjects in the control group engaged in a corrupt transaction in the experimental bribery game.

The baseline for Study 3's index represents the mean number of the five bribe scenarios to which subjects in the control group responded affirmatively. In wave 1, respondents, on average, indicated bribe intention for approximately 1.1 of the 5 scenarios; for wave 2, the corresponding figure was 1.0. Although we rely on this index for our primary analyses of treatment effects, per our pre-analysis plan, Table 4 also includes disaggregated analyses for the five dichotomous bribe scenarios. For these five scenarios, the baseline figures, as in Studies 1 and 2, represent the percentage of control group subjects indicating willingness to bribe. The variation across scenarios stands out, ranging from a low 15% expressing

Table 4: Main Effects

	Study 1	**Study 2**	**Study 3** ($W1$)						**Study 3** ($W2$)
Treatment			Index	*Police*	*License*	*Education*	*Medical*	*Business*	Index
(1) Baseline (C)	0.165***	0.322***	1.129***	0.153***	0.149***	0.225***	0.413***	0.187***	1.004***
	(0.008)	(0.025)	(0.033)	(0.009)	(0.009)	(0.010)	(0.012)	(0.010)	(0.055)
(2) Decreasing (D)			−0.073	−0.020	−0.003	−0.041*	0.019	−0.029+	
			(0.057)	(0.015)	(0.015)	(0.017)	(0.021)	(0.016)	
(3) Increasing (I)			−0.023	−0.020	0.008	−0.016	−0.010	0.016	
			(0.060)	(0.015)	(0.016)	(0.018)	(0.021)	(0.017)	
(4) Injunctive (J)			−0.222***	−0.037**	−0.025+	−0.062***	−0.062**	−0.033*	
			(0.055)	(0.014)	(0.015)	(0.017)	(0.021)	(0.016)	
(5) Decreasing-Injunctive (D-J)	−0.022+	−0.073*	−0.135***	−0.035***	−0.020+	−0.050***	0.010	−0.038***	
	(0.012)	(0.034)	(0.040)	(0.011)	(0.011)	(0.012)	(0.015)	(0.011)	
(6) Increasing-Injunctive (I-J)	−0.012		−0.143*	−0.042**	−0.003	−0.052**	−0.019	−0.029+	
	(0.012)		(0.056)	(0.014)	(0.015)	(0.017)	(0.021)	(0.016)	
(7) t1: D-J/ t2: C									−0.072
									(0.077)
(8) t1: D-J/ t2: D-J									−0.077
									(0.080)
N	5,670	693	7,829	7,884	7,883	7,878	7,881	7,880	1,505

Notes: Baseline refers to control group means. Indices measure the mean number of bribe scenarios to which subjects responded affirmatively, on a scale of 0 to 5. All other outcome variables are dichotomous, and so the baseline figures represent the proportion of subjects reporting intention to bribe (Study 1 and Study 3) or engaging in a bribe transaction in the bribery game (Study 2). Robust standard errors in parentheses. Multiple significance thresholds shown, where $^{+}p < .10$, $^{*}p < .05$, $^{**}p < 0.01$, and $^{***}p < .001$, but .05 is the pre-registered hypothesis test level.

readiness to bribe to expedite a driver's license (an estimate similar to Study 1) to a high of 41% expressing readiness to bribe to expedite medical treatment.

Table 4 shows clear evidence of the first of our four key findings: Treatments including injunctive norm messages consistently reduced subjects' propensity to engage in bribery. Contrary to the social norms literature's predictions presented above (see Table 1), row 4 in Table 4 shows that the Study 3 treatment consisting of just the injunctive norm message (Treatment J) — rather than the combination of an injunctive and a descriptive norm about decreasing corruption (Treatment D-J) — had the largest effect, a decline of 0.23 points on the 5 points scale (0.18 of a standard deviation), or an approximately 20 percent decrease relative to the control group baseline. The D-J treatment, shown in row 5, also decreased bribe intention in Study 3, but the effect was smaller: 0.14 points on the 5-point scale (0.11 of

a standard deviation), a 12 percent decline relative to the control group. This effect also was no longer statistically significant one week after the initial treatment, even for subjects who received additional exposure to the treatment message multiple times between the two survey waves (rows 7 and 8). While the *D*-*J* treatment in Study 2 did produce a robust decrease of over seven percentage points in subjects' likeliness of engaging in a bribe transaction in the laboratory corruption game, a 22 percent decrease relative to the control group, in Study 1 this treatment caused a more modest decline in bribe intention of just over 2.2 percentage points, or a 13 percent decrease relative to the control group.

Our findings on injunctive norms diverged from the social norms literature's predictions in two additional ways: As shown in row 6, treatments in Study 3 combining injunctive norms with descriptive norms about increasing corruption (Treatment *I*-*J*) also produced a statistically significant decrease in bribe intention, contrary to expectations that this treatment's countervailing messages would mitigate each other's effects. Meanwhile, the treatment consisting exclusively of a descriptive norm about decreasing corruption (Treatment *D*) confounded predictions by failing to decrease propensity to bribe, per row 2.

Turning to our second key finding, Table 4 shows no evidence of backfire effects. Contrary to Abbink et al. (2018), Cheeseman and Peiffer (2022), and Corbacho et al. (2016), the message in Study 3 consisting exclusively of a descriptive norm about increasing corruption (Treatment *I*) did not increase subjects' bribe intention, as seen in row 3. And, as noted above, messaging about increasing corruption combined with an anti-corruption injunctive norm (Treatment *I*-*J*) *decreased* subjects' propensity for corruption in Study 3 and produced no statistically significant effect in Study 1. Moreover, we find no evidence of backfire effects even when examining subgroups whose prior beliefs about corruption's prevalence could make them particularly susceptible to backfire effects, as we discuss below. We return to the issue of why our results diverge from those of earlier studies in the concluding section.

In summary, two key findings emerge from our primary analyses: First, messaging based on injunctive norms, both in isolation and in combination with various descriptive norms,

produced consistent, albeit moderate and temporary, reductions in willingness to engage in bribery. Second, none of our studies, including those with treatment messages conveying descriptive norms about increasing corruption, produced evidence of backfire effects.

Accounting for Perceived Credibility of Treatment Messages

Our third key finding pertains to the perceived credibility of treatment messages — particularly those that convey information about decreasing corruption. A potential challenge of using messaging about decreases in undesirable behavior in societies where such behavior is widespread is that subjects may respond skeptically even to factual information. To examine this concern, we presented subjects in Study 3 with both traditional manipulation checks and with survey items evaluating the extent to which they found the descriptive-norm information believable. Online Appendix C.1 shows that, for all treatments containing descriptive norms, over 90 percent of subjects passed manipulation checks asking them to recall treatment message details. But, as Table 5 highlights, subjects were far more likely to believe messages about upward than about downward corruption trends. Only approximately a quarter of respondents perceived Treatments D and D-J as credible, as measured by a rating of four or higher on a six-point credibility scale, compared to around 60 percent or more who perceived Treatments I and I-J as credible.

Given the non-random distribution of subjects' credibility perceptions, a simple comparison of bribe intent among those who were treated *and* found the treatment credible to subjects from the control group would likely produce biased estimates. Therefore, for each of the four treatments that included descriptive-norm messaging, we divided subjects into two sub-groups according to whether they perceived messages to be credible. We then employed entropy balancing (Hainmueller 2012) to re-weight the control group such that its pre-treatment covariate means and other sample moments matched those in our sub-groups

Table 5: Believability of Treatment Message

	Mean Rating	% Perceiving as Credible
Treatment *D*	2.41	23.8
Treatment *I*	3.89	64.6
Treatment *D*-*J*	2.47	25.2
Treatment *I*-*J*	3.74	59.3

Notes: Rated on 6-point scale, where 1 is "absolutely do not believe" and 6 is "absolutely believe." Percent perceiving as credible refers to rating of 4 or higher.

of treated subjects.[24] We then estimated treatment effects conditional on perceptions of treatment messages' credibility.

The analyses presented in Table 6 were not pre-registered and should therefore be considered exploratory.[25] But the results indicate that among subjects who perceived the decreasing corruption message to be credible, Treatment *D*-*J* produced an effect as large as any of the unconditional effects shown in Table 4 — a 0.23 decline on the five-point bribery index (0.18 of a standard deviation) — as shown in row 4, column 6. By contrast, the effect among subjects who did not perceive the *D*-*J* message as credible (row 4, column 7) is approximately two times smaller. Moreover, the effect of repeated exposure to Treatment *D*-*J* among those perceiving the message as credible proved durable: A week later, these effects not only remained statistically significant but were nearly double in magnitude: a 0.42 point decline on the five-point bribery index (0.34 of a standard deviation) (row 7, column 2).

Accounting for treatment messages' perceived credibility offers suggestive evidence that messages incorporating descriptive norms about declining corruption might produce relatively large and sustainable treatment effects. However, the challenge facing such campaigns

24. Prior to entropy balancing, we multiply imputed missing values for covariates using van Buuren and Groothuis-Oudshoorn's (2011) `mice` algorithm. Analyses shown here use four on the six-point scale as the cutpoint for perceived credibility, but results are robust to other thresholds and multiple ways of estimating the average treatment effect on the treated (ATT). See Online Appendix C.2 for details.

25. Entropy balancing achieved balance on a large and varied set of pre-treatment covariates (see Online Appendix C.2), but like all matching techniques, it cannot ensure balance on unobservables. Nevertheless, absent a technique for experimentally manipulating credibility perceptions, we believe our approach contributes insights into a substantively important but challenging-to-estimate quantity: treatment effects adjusted for whether subjects find treatment messages believable.

Table 6: Study 3 Effects Conditional on Perceived Credibility of Treatment Messages (Entropy-Weighted ATT Estimates)

	Main Effects (1)	Credible (2)	Not Credible (3)	Credible (4)	Not Credible (5)	Credible (6)	Not Credible (7)	Credible (8)	Not Credible (9)
Wave 1:									
(1) Decreasing (D)	−0.073	0.009	−0.068						
	(0.057)	(0.125)	(0.062)						
(2) Increasing (I)	−0.023			−0.035	−0.016				
	(0.060)			(0.087)	(0.076)				
(3) Injunctive (J)	−0.222***								
	(0.055)								
(4) Decreasing-Injunctive (D-J)	−0.135***					−0.225***	−0.119**		
	(0.040)					(0.067)	(0.043)		
(5) Increasing-Injunctive (I-J)	−0.143*							−0.141+	−0.158*
	(0.056)							(0.081)	(0.068)
N	7,829	1,730	2,318	1,979	2,037	2,006	4,235	1,941	2,072
Wave 2:									
(6) t1: D-J/ t2: C	−0.072			−0.100	−0.077				
	(0.077)			(0.136)	(0.083)				
(7) t1: D-J/ t2: D-J	−0.077	−0.420***	−0.026						
	(0.080)	(0.126)	(0.087)						
N	1,505	652	1,353	652	1,353				

Notes: Dependent variable for all models is the additive bribe index measuring the mean number of bribe scenarios to which subjects responded affirmatively, on a scale of 0 to 5. Perceived credibility refers to a rating of four or higher on the six-point credibility scale. Missing values for pre-treatment covariates multiply imputed for columns 2-9 (van Buuren and Groothuis-Oudshoorn 2011). Robust standard errors in parentheses. $^{+}$ $p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

in countries such as Ukraine is that only a minority of individuals find them believable. Fortunately, our analyses indicate that it is feasible to identify and thereby help anti-corruption campaigns target the types of citizens most likely to be persuaded by messages similar to Treatment D-J. Online Appendix Table C-3 provides evidence that women, younger citizens, and subjects who score high on the susceptibility to persuasion (STPS) scale are more likely to find messages about decreasing corruption credible.

Heterogeneous Treatment Effects

Our final set of findings concerns heterogeneous treatment effects across psychological traits, prior beliefs about corruption levels and trends, and demographic characteristics. This section highlights three sets of results; Online Appendix D provides additional details.

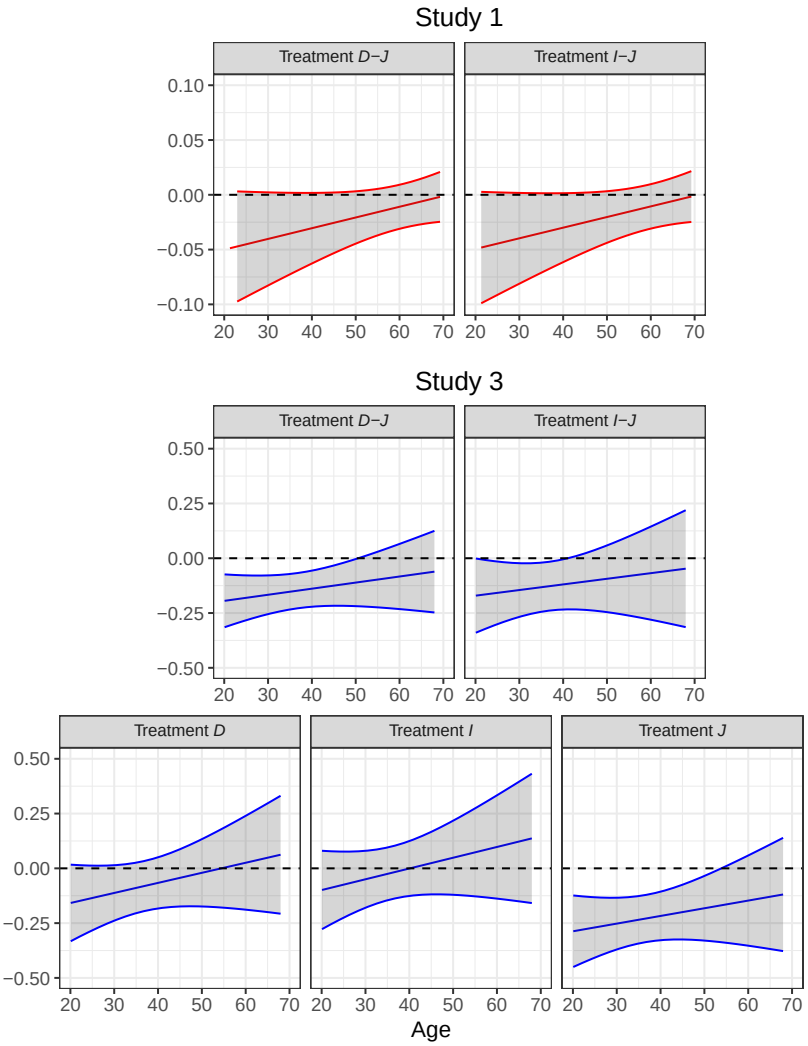
First, we find little evidence that prior beliefs make certain citizens more susceptible to backfire effects. Whereas Cheeseman and Peiffer's (2022) previously-discussed priming theory suggests messaging should be particularly likely to backfire among individuals who believe corruption to be widespread, in both Studies 1 and 3, we find messaging more likely to *reduce* willingness to bribe among "pessimistic perceivers" (see Online Appendices D.1 and E). This finding corresponds with the social norms literature's prediction that accurate information about descriptive norms may induce the largest declines in undesirable behavior among individuals who previously overestimated the behavior's prevalence (see, e.g., Schultz et al. 2007). We also find suggestive evidence that anti-corruption messaging is more effective among individuals with prior beliefs that corruption is declining.

Second, concordant with our hypotheses, we find larger treatment effects among subjects with low Need for Cognition (NfC), high agreeableness, and low openness scores across nearly all treatment arms (see Online Appendix D.2). We do not, however, find support for the hypothesis that messaging should have a greater effect on subjects with high susceptibility to persuasion strategies (STPS) scores.

Third, across our three studies, we find that effects are consistently larger among younger subjects. Figure 4 shows that this pattern holds for all five treatment conditions in Study 3.[26] For subjects under 30 years-old, treatment effects are statistically significant and approximately 1.5 times larger than the effect for the overall sample in four of the five treatment arms, the exception being the Treatment *I* message. Meanwhile, for subjects over 50, effects across all treatment arms are statistically indistinguishable from zero. As seen in Figure 4, similar patterns hold for both treatment arms in Study 1. Finally, consistent with the proposition that anti-corruption messaging has a greater impact on younger citizens, the robust treatment effects found in Study 2 (see row 5 in Table 4) are based solely on university-age students. The conclusion discusses how these findings have significant implications for how to effectively target anti-corruption campaigns.

26. Diagnostics developed by Hainmueller et al. (2019) produce no evidence of non-linear interaction effects for age and other subgroup effects discussed in Online Appendix D.

Figure 4: Marginal Effects Conditional On Age



Notes: These figures plot treatment effects by age. Effects are consistently larger among younger subjects. Analyses control for gender, education level, language preference (Ukrainian vs. Russian), and region.

Discussion

Drawing on three studies conducted between 2017 and 2021 in Ukraine, this article offers novel insights regarding anti-corruption information campaigns' effectiveness. Extending the recently emerging literature, our factorial design facilitates a far more comprehensive analysis of descriptive and injunctive norms than existing studies. Ours is also the first to investigate duration effects, treatment messages' perceived credibility, and an extensive set of pre-registered hypotheses about heterogeneous effects.

We emphasize four main findings. First, across our multiple studies, injunctive-norms messaging consistently reduced bribe intentions, but effect sizes were relatively modest and of short duration. Second, none of our studies produced robust evidence of "backfire" effects, neither in the overall sample nor in subgroups that might be particularly susceptible to such effects. Third, while we find evidence that combinations of injunctive norms with descriptive norms about decreasing corruption — the messaging the social norms literature predicts should be most effective — can produce relatively large and durable effects, we also find that these effects are concentrated among respondents who perceive information about declining corruption as credible. In countries with widespread corruption, only a minority of citizens are likely to believe such messaging. Finally, our analyses point to the potential for anti-corruption messaging to have the greatest impact among younger citizens.

These findings offer more cause for optimism than previous anti-corruption messaging studies. Still, they suggest the need to reconceptualize the objectives of at least some types of anti-corruption information campaigns. While social norm messaging's advocates often focus on transforming entrenched beliefs and expectations, we suggest that anti-corruption advertising should be deployed to modify short-term intentions and choices. Moreover, effective campaigns will require attention to messaging's content, perceived credibility, location and timing of placement, and targeting. For example, materials emphasizing injunctive-norm messaging should be placed at strategic locations (e.g., entrances to public service agencies) with the goal of affecting bribe-making decisions on the margin. Meanwhile, messaging combining injunctive and descriptive norms about declining corruption could be targeted via social media at those most likely to find such messaging credible (younger people and women) with the more ambitious goal of transforming longer-term beliefs and expectations.

These findings raise important questions for future research. First, while existing studies on anti-corruption messaging focus predominantly on visual messaging with which citizens interact fleetingly, little rigorous evidence exists about the effects of more intensive approaches such as anti-corruption training for public officials or educational programs in schools or uni-

versities. Tools like these may hold greater promise for reformers seeking to transform social norms fundamentally. That said, the cumulative effects of anti-corruption advertising on many smaller decisions to eschew bribery should not be underestimated. Consider a country such as Ukraine in the years before Russia's 2022 invasion, with an adult population of more than 30 million. Approximately 30 percent of those who sought public services in the past 12 months reported paying at least one bribe. Therefore, a single percentage point decline in the bribe rate would reduce the number of yearly bribe-payers by 150,000. A treatment effect of 20 percent reduction from the baseline — an effect similar to that of our studies' more effective messages — would result in nearly a million fewer citizens paying bribes annually.[27] Moreover, when individuals' choices depend on expectations about others' choices, recent analyses find that behavioral shifts among as little as 25 percent of a population can create tipping points that spark large-scale changes (Centola et al. 2018). Accordingly, a better understanding of how citizens' individual yet interconnected decisions about engaging in corruption affect aggregate corruption levels is a second promising line of future research.

A third area for future research concerns our findings about anti-corruption messaging's strong effect on younger citizens. Future investigations could offer insights into whether these results reflect younger individuals' generally more malleable social norms, or whether older Ukrainians' specific experiences during the late Soviet or early post-Soviet periods have shaped mentalities about corruption in ways unique to the post-communist region.

Finally, better understanding the extent to which lessons about messaging from one part of the world generalize to other regions, and which findings generalize from campaigns aimed at one social problem to those targeting others, will be critical to address as the literature on anti-corruption campaigns grows. The extent of generalizations is particularly important with respect to the risk of backfire effects. We are skeptical that a wide range of anti-corruption messages can prime citizens to engage in corruption simply by raising the

27. This estimate that 30% of citizens pay bribes is based on the surveys used in our treatment messages (Figure 3), which show annual bribe rates ranging from 20% to 40% in the decade prior to 2022. Based on estimates from the same surveys, our calculations assume that around half of citizens sought access to public services annually, meaning that approximately 15 million faced a potential bribe decision in any given year.

topic's salience. But with the exception of our studies, extensive evidence indicates that information about the prevalence or rising levels of unwanted behavior can produce adverse effects, suggesting that practitioners should avoid such messaging. That our treatments with descriptive-norm messages about increasing corruption did not increase subjects' bribe intentions raises questions about why our findings diverge from those of earlier studies. One possibility is that unlike studies conducted in countries such as Costa Rica, where increasing bribery contrasts sharply with traditionally low levels of corruption (e.g, Corbacho et al. 2016), information about rising bribery is likely unsurprising to subjects in countries with endemic corruption. Although strong beliefs about corruption's prevalence may limit the effectiveness of normative-trend messaging about declining corruption, the silver lining of such beliefs may be that they inoculate citizens against potential backfire effects from poorly designed anti-corruption messaging.

While much remains unknown about anti-corruption messaging's effects, our studies offer a nuanced yet guardedly optimistic assessment of the prospects for harnessing social norms to reduce corruption. Anti-corruption campaigns are, to be sure, no silver bullet, and poorly designed messaging may produce more harm than good. But with realistic objectives, careful message design, strategic deployment, and sophisticated targeting, these campaigns have the potential to serve as a valuable tool in the fight against corruption.

References

Abbink, Klaus. 2006. "Laboratory experiments on corruption." In *International Handbook on the Economics of Corruption,* edited by Susan Rose-Ackerman. ElgarOnline.

Abbink, Klaus, Esteban Freidin, Lata Gangadharan, and Rodrigo Moro. 2018. "The effect of social norms on bribe offers." *The Journal of Law, Economics, and Organization* 34 (3): 457–474.

Agerberg, Mattias. 2022. "Messaging about corruption: The power of social norms." *Governance* 35 (3): 929–950.

Allcott, Hunt. 2011. "Social norms and energy conservation." *Journal of Public Economics* 95 (9-10): 1082–1095.

Barr, Abigail, and Danila Serra. 2010. "Corruption and culture: An experimental analysis." *Journal of Public Economics* 94 (11): 862–869.

Benz, Matthias, and Stephan Meier. 2008. "Do people behave in experiments as in the field? Evidence from donations." *Experimental Economics* 11 (3): 268–281.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan. 2007. "Obtaining a driver's license in India: An experimental approach to studying corruption." *The Quarterly Journal of Economics* 122 (4): 1639–1676.

Bhanot, Syon P. 2021. "Isolating the effect of injunctive norms on conservation behavior: New evidence from a field experiment in California." *Organizational Behavior and Human Decision Processes* 163:30–42.

Bicchieri, Cristina. 2016. *Norms in the wild: How to diagnose, measure, and change social norms.* Oxford University Press.

Blair, Graeme, Rebecca Littman, and Elizabeth Levy Paluck. 2019. "Motivating the adoption of new community-minded behaviors: An empirical test in Nigeria." *Science Advances* 5 (3): eaau5175.

Blattman, Christopher, Horacio Larreguy, Benjamin Marx, and Otis R. Reid. 2019. "Eat widely, vote wisely? Lessons from a campaign against vote buying in Uganda." National Bureau of Economic Research Working Paper 26293.

Cacioppo, John T., Richard E. Petty, and Chuan Feng Kao. 1984. "The efficient assessment of need for cognition." *Journal of Personality Assessment* 48 (3): 306–307.

Centola, Damon, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. 2018. "Experimental evidence for tipping points in social convention." *Science* 360 (6393): 1116–1119.

Cheeseman, Nic, and Caryn Peiffer. 2022. "The curse of good intentions: Why anticorruption messaging can encourage bribery." *American Political Science Review* 116 (3): 1081–1095.

Cialdini, Robert B. 2001. *Influence: Science and practice.* Boston: Allyn & Bacon.

Cialdini, Robert B., Linda J. Demaine, Brad J. Sagarin, Daniel W. Barrett, Kelton Rhoads, and Patricia L. Winter. 2006. "Managing social norms for persuasive impact." *Social Influence* 1 (1): 3–15.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. 1990. "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places." *Journal of Personality and Social Psychology* 58 (6): 1015–1026.

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.

Denisova-Schmidt, Elena, Martin Huber, and Yaroslav Prytula. 2015. "An experimental evaluation of an anti-corruption intervention among Ukrainian university students." *Eurasian Geography and Economics* 56 (6): 713–734.

Druckman, James N., and Thomas J. Leeper. 2012. "Learning more from political communication experiments: Pretreatment and its effects." *American Journal of Political Science* 56 (4): 875–896.

Erlich, Aaron. 2020. "Can information campaigns impact preferences toward vote selling? Theory and evidence from Kenya." *International Political Science Review* 41 (3): 419–435.

Farrow, Katherine, Gilles Grolleau, and Lisette Ibanez. 2017. "Social norms and pro-environmental behavior: A review of the evidence." *Ecological Economics* 140:1–13.

Fisman, Raymond, and Miriam A. Golden. 2017. *Corruption: What everyone needs to know.* Oxford University Press.

Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. 2003. "A very brief measure of the Big-Five personality domains." *Journal of Research in Personality* 37 (6): 504–528.

Guess, Andrew, and Alexander Coppock. 2020. "Does counter-attitudinal information cause backlash? Results from three large survey experiments." *British Journal of Political Science* 50 (4): 1497–1515.

Hainmueller, Jens. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* 20 (1): 25–46.

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.

Hanna, Rema, and Shing-Yi Wang. 2017. "Dishonesty and selection into public service: Evidence from India." *American Economic Journal: Economic Policy* 9 (3): 262–290.

Hicken, Allen, Stephen Leider, Nico Ravanilla, and Dean Yang. 2018. "Temptation in vote-selling: Evidence from a field experiment in the Philippines." *Journal of Development Economics* 131:1–14.

Hoffmann, Leena Koni, and Raj Navanit Patel. 2017. *Collective action on corruption in Nigeria: A social norms approach to connecting society and institutions.* Chatham House Report (May 17). Available at: https://www.chathamhouse.org/2017/05/collective-action-corruption-nigeria.

Kaptein, Maurits. 2012. "Personalized persuasion in ambient intelligence," Eindhoven University of Technology.

Kaptein, Maurits, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. 2012. "Adaptive persuasive systems: A study of tailored persuasive text messages to reduce snacking." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2 (2): 1–25.

KIIS. 2015. *Corruption in Ukraine: Comparative analysis of national surveys: 2007, 2009, 2011, and 2015.* Technical report. Available at: https://kiis.com.ua/materials/pr/20161602_corruption/Corruption%20in%20Ukraine%202015%20ENG.pdf.

Klitgaard, Robert. 1988. *Controlling corruption.* University of California Press.

Köbis, Nils C., Marleen Troost, Cyril O. Brandt, and Ivan Soraperra. 2019. "Social norms of corruption in the field: Social nudges on posters can help to reduce bribery." *Behavioural Public Policy,* 1–28.

Köbis, Nils C., Jan-Willem Van Prooijen, Francesca Righetti, and Paul Van Lange. 2015. "'Who doesn't?' – The impact of descriptive norms on corruption." *PloS One* 10 (6): 1–14.

Lewis, Melissa A., and Clayton Neighbors. 2006. "Social norms approaches using descriptive drinking norms education: A review of the research on personalized normative feedback." *Journal of American College Health* 54 (4): 213–218.

Lins de Holanda Coelho, Gabriel, Paul Hanel, and Lukas Wolf. 2020. "The very efficient assessment of need for cognition: Developing a six-item version." *Assessment* 27 (8): 1870–1885.

Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. 2018. "Worth weighting? How to think about and use weights in survey experiments." *Political Analysis* 26 (3): 275–291.

Mishler, William, and Richard Rose. 2007. "Generation, age, and time: The dynamics of political learning during Russia's transformation." *American Journal of Political Science* 51 (4): 822–834.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62 (3): 760–775.

Mortensen, Chad R., Rebecca Neel, Robert B. Cialdini, Christine M. Jaeger, Ryan P. Jacobson, and Megan M. Ringel. 2019. "Trending norms: A lever for encouraging behaviors performed by the minority." *Social Psychological and Personality Science* 10 (2): 201–210.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113 (2): 517–529.

Nyhan, Brendan, and Jason Reifler. 2015. "Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information." *Vaccine* 33 (3): 459–464.

Nyhan, Brendan, Jason Reifler, Sean Richey, and Gary Freed. 2014. "Effective messages in vaccine promotion: a randomized trial." *Pediatrics* 133 (4): 835–842.

Olken, Benjamin A., and Rohini Pande. 2012. "Corruption in developing countries." *Annual Review of Economics* 4 (1): 479–509.

Ostrom, Elinor. 2000. "Collective action and the evolution of social norms." *Journal of Economic Perspectives* 14 (3): 137–158.

Peiffer, Caryn. 2020. "Message received? Experimental findings on how messages about corruption shape perceptions." *British Journal of Political Science* 50 (3): 1207–1215.

Peiffer, Caryn, and Grant Walton. 2019. *Overcoming collective action problems through anti-corruption messages.* Development Policy Centre Discussion Paper No. 77. Available at https://devpolicy.org/publications/discussion_papers/DP77-Overcoming.collective.action.problems.pdf.

Persson, Anna, Bo Rothstein, and Jan Teorell. 2013. "Why anticorruption reforms fail – Systemic corruption as a collective action problem." *Governance* 26 (3): 449–471.

Pop-Eleches, Grigore, and Joshua Tucker. 2017. *Communism's shadow: Historical legacies and contemporary political attitudes.* Princeton University Press.

Raymond, Leigh, Daniel Kelly, and Erin P. Hennes. 2023. "Norm-based governance for severe collective action problems: Lessons from climate change and COVID-19." *Perspectives on Politics* 21 (2): 519–532.

Reno, Raymond R., Robert B. Cialdini, and Carl A. Kallgren. 1993. "The transsituational influence of social norms." *Journal of Personality and Social Psychology* 64 (1): 104–112.

Ryvkin, Dmitry, and Danila Serra. 2020. "Corruption and competition among bureaucrats: An experimental study." *Journal of Economic Behavior & Organization* 175:439–451.

Scharbatke-Church, Cheyanne, and Diana Chigas. 2019. *Understanding social norms: A reference guide for policy and practice.* Corruption, Justice & Legitimacy Project at the Fletcher School of Law & Diplomacy Henry J. Leir Institute. Available at: https://sites.tufts.edu/ihs/files/2019/10/SN_CorruptionRefGuide_AUG2019-linked.MR_.pdf.

Scharbatke-Churck, Cheyanne, and Russell Hathaway. 2017. "Are social norms an important missing link in anti-corruption programming?" *CDA Collaborative* (March 21). Available at: https://www.cdacollaborative.org/blog/social-norms-important-missing-link-anti-corruption-programming/.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. "The constructive, destructive, and reconstructive power of social norms." *Psychological Science* 18 (5): 429–434.

Smith, Joanne R., Winnifred R. Louis, Deborah J. Terry, Katharine H. Greenaway, Miranda R. Clarke, and Xiaoliang Cheng. 2012. "Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions." *Journal of Environmental Psychology* 32 (4): 353–361.

Sparkman, Gregg, and Gregory M. Walton. 2017. "Dynamic norms promote sustainable behavior, even if it is counternormative." *Psychological Science* 28 (11): 1663–1674.

Stephenson, Matthew C. 2020. "Corruption as a self-reinforcing trap: Implications for reform strategy." *The World Bank Research Observer* 35 (2): 192–226.

Svensson, Jakob. 2005. "Eight questions about corruption." *Journal of Economic Perspectives* 19 (3): 19–42.

Tan, Ming T., Guo-Liang Tian, and Man-Lai Tang. 2009. "Sample surveys with sensitive questions: A nonrandomized response approach." *The American Statistician* 63 (1): 9–16.

Tankard, Margaret E., and Elizabeth Levy Paluck. 2016. "Norm perception as a vehicle for social change." *Social Issues and Policy Review* 10 (1): 181–211.

Ternovski, John, and Lilla Orr. 2022. "A note on increases in inattentive online survey-takers since 2020." *Journal of Quantitative Description: Digital Media* 2.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate imputation by chained equations in R." *Journal of Statistical Software* 45 (3).

Wakefield, Jane. 2018. "Cambridge Analytica: Can targeted online ads really change a voter's behaviour?" *BBC News* (March 30).

Webb, Thomas L., and Paschal Sheeran. 2006. "Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence." *Psychological Bulletin* 132 (2): 249–268.

Wood, Thomas, and Ethan Porter. 2019. "The elusive backfire effect: Mass attitudes' steadfast factual adherence." *Political Behavior* 41 (1): 135–163.

Zuboff, Shoshana. 2015. "Big other: Surveillance capitalism and the prospects of an information civilization." *Journal of Information Technology* 30 (1): 75–89.

Online Appendix

Can Norm-Based Information Campaigns Reduce Corruption?

Appendices

Contents

A Human Subjects Research & Ethical Research Practices **1**

B Overview of Survey Samples **2**
B.1 Study 1 & Study 2 Sample Descriptive Statistics 2
B.2 Study 3 Sample Descriptive Statistics & Population Benchmarking 3

C Perceived Credibility of Treatment Messages in Study 3 **5**
C.1 Manipulation Checks 5
C.2 Effects Conditional on Perceived Credibility with Alternative Thresholds . . 5
C.3 Predictors of Perceived Credibility 7

D Heterogeneous Treatment Effects **9**
D.1 Prior Beliefs About Corruption 9
D.2 Psychological Traits & Susceptibility to Persuasion 9
D.3 Demographics 10

E Supplementary Analyses of Potential Backfire Effects **11**

F Robustness Checks **13**
F.1 Randomization Checks 13
F.2 Non-Response Rates by Treatment Arm for Study 1 & Study 3 13
F.3 Crosswise Model Used in Study 1 14
F.4 List Experiments Used in Study 3 15
F.5 Regressions with Covariate Adjustment & Weights 16

G Supplementary Information About Research Instruments **18**
G.1 Ukrainian & Russian Versions of Information Treatments 18
G.2 Scripts for Bribery Game for Study 2 18
G.3 Social Psychology Indices Used in Study 3 18

H Mapping to Pre-Analysis Plans **19**

A Human Subjects Research & Ethical Research Practices

All three studies were approved by the authors' university Institutional Review Boards and conducted in accordance with APSA's Principles and Guidance on Human Subjects Research:

Power: All participation was voluntary and subjects could opt out at any time. No coercion or pressure was used to persuade subjects to participate.

Consent: For Study 2 and Study 3, the initial screen of the online survey instrument served as a consent form. It included all items recommended by APSA's Principles and Guidance for consent. Informed consent was indicated by clicking the "next page" button to start the survey. For Study 1, the consent process was managed by one of Ukraine's top survey firms, the Kyiv International Institute of Sociology (KIIS).

Deception: None of the studies employed deception.

Harm and Trauma: The studies involved minimal risk of harm or trauma to participants. The topic of corruption is regularly discussed in Ukraine, and sensitive survey question techniques suggest no signs of discomfort, as evidenced by subjects' high levels of honesty when answering direct questions about corruption (see Online Appendices F.3 and F.4).

Confidentiality: The survey instruments recorded no identifying information about research participants. For the purpose of compensation, cell phone numbers and emails were collected in Study 2 and Study 3, respectively. All phone and email data was stored separately from survey answers, and proper data management practices were employed to keep this information secure.

Impact: The studies had no direct effect on and in no way compromised the integrity of local political processes.

Laws, Regulations, and Prospective Review: All studies were approved by the authors' university Institutional Review Boards and were conducted in compliance with Ukraine's laws. The authors also adhered to ethical research practices beyond IRB review, including reliance on input from local researchers and experts to ensure that all research practices complied with informal social norms.

Compensation: For Study 1, the survey firm conducting the survey compensated subjects at their standard rate. For Study 2, compensation depended on the play of experimental games, as detailed in the article text. For Study 3, participants in the first survey wave were entered into a lottery to win one of three Apple Watches; participants in the second survey wave were entered into an additional lottery to win an iPad.

Shared Responsibility: The authors recognize the responsibility of all parties for upholding ethical standards for research practices and have sought to adhere to the principle of shared responsibility, as described by APSA's Principles and Guidance for Human Subjects Research.

B Overview of Survey Samples

B.1 Study 1 & Study 2 Sample Descriptive Statistics

Study 1 was embedded in a face-to-face survey that employed multi-stage cluster random sampling and included booster samples. Descriptive statistics are representative of the adult Ukrainian population when weights are applied. Study 2 was conducted with Ukrainian university students. To construct the sample, we conducted stratified random sampling by class year and department using a sample frame based on university enrollment data. Tables B-1 and B-2 present descriptive statistics from Study 1 and Study 2, respectively.

Table B-1: Study 1 Descriptive Statistics

	Unweighted	**Weighted**
	N = 6,926[1]	**N = 30,283,502**[2]
Age	53 (18)	47 (18)
Woman	4,596 (66%)	(55%)
Home Language		
Ukrainian	2,428 (35%)	(48%)
Russian	2,852 (41%)	(31%)
Both	1,379 (20%)	(19%)
Other	243 (3.5%)	(1.6%)
Survey in Russian	4,243 (61%)	(44%)
Higher Education	2,196 (32%)	(32%)
City Size/Settlement Type		
Village	2,034 (29%)	(34%)
Urban type village	294 (4.2%)	(5.5%)
Small town (up to 20K)	307 (4.4%)	(6.4%)
Town with population 20-49K	316 (4.6%)	(7.9%)
Town with population 50-99K	199 (2.9%)	(5.1%)
City with population 100-499K	1,251 (18%)	(20%)
Large city 500K+	2,525 (36%)	(21%)
Region		
Central	1,462 (21%)	(35%)
East	1,928 (28%)	(26%)
South	1,908 (28%)	(12%)
West	1,628 (24%)	(27%)

[1]Mean (SD); N (%). [2]Mean (SD); (%).

Table B-2: Study 2 Descriptive Statistics

	N = 695[1]
Age	19.30 (1.92)
Woman	447 (64%)
Home Language	
Ukrainian	277 (40%)
Russian	269 (39%)
Both	131 (19%)
Other	17 (2.4%)
Home region	
Central	182 (27%)
East	82 (12%)
South	291 (43%)
West	127 (19%)

[1]Mean (SD); N (%)

B.2 Study 3 Sample Descriptive Statistics & Population Benchmarking

Study 3 was conducted with a convenience sample of Ukrainians recruited via Facebook. While we make no claims regarding representativeness, our sample includes a wide range of demographic groups. This section provides descriptive statistics for our survey sample and, where possible, population benchmarks from 2021 census extrapolations estimated by the Census Bureau of the State Statistics Service of Ukraine.[1]

The sample is 40% male versus 59% female, and all age groups between the ages 18 and 70 are well-represented (see Table B-3). The overrepresentation of under 30-year-olds relative to the Ukrainian population reflects our intentional oversampling of this age group in order to have sufficient power for evaluating our pre-registered hypothesis concerning heterogeneous effects by age (see additional details in Online Appendix D.3). The sample also displays considerable geographic diversity, with the majority of the sample reporting place of residence as outside of the capital (Kyiv) or other large cities – 35% in small cities or towns and 25% in a village or rural area (see Table B.2). The sample also includes respondents from all of Ukraine's regional administrative units (*oblasts*), with the exception of Russian-occupied Crimea. While there are disproportionately fewer respondents from the east (see Table B-3), this may reflect the inability of recent census extrapolations to properly account for internal migration resulting from the war in Donbas that has been ongoing since 2014.

1. We rely on extrapolations because Ukraine has not conducted a census since 2001.

Table B-3: Benchmarking Survey Sample to Census Data

	Census		Survey	
	N	Pct	N	Pct
Gender				
Men	15,353,209	45.2%	3,183	40.3%
Women	18,605,831	54.8%	4,717	59.7%
Age groups				
[18,30)	5,300,461	15.6%	2,813	37.8%
[30,40)	6,849,855	20.2%	1,727	23.2%
[40,50)	6,079,955	17.9%	993	13.3%
[50,60)	5,607,145	16.5%	982	13.2%
[60,70)	5,286,715	15.6%	764	10.3%
[70,Inf]	4,834,909	14.2%	161	2.2%
Regions				
East	11,747,035	40.3%	1,344	17.3%
North/Central	8,317,520	28.5%	2,904	37.4%
South	2,996,401	10.3%	707	9.1%
West	6,078,390	20.9%	2,810	36.2%

Notes: Census data are from the January 1, 2021 census extrapolations conducted by the Census Bureau of the State Statistics Service of Ukraine.

Table B-4: Additional Descriptive Statistics for Study 3

	Unweighted	**Weighted**
	N = 7,901[1]	**N = 29,139,346**[2]
Survey in Russian	1,851 (23%)	(36%)
Higher Education	5,293 (67%)	(72%)
Home Language		
Ukrainian	5,449 (69%)	(54%)
Russian	700 (8.9%)	(16%)
Both	1,700 (22%)	(30%)
Other	46 (0.6%)	(0.7%)
Employed	5,551 (71%)	(65%)
City Size/Settlement Type		
Kyiv	780 (9.9%)	(7.7%)
Big City (not Kyiv)	2,383 (30%)	(38%)
Small City	2,738 (35%)	(35%)
Rural/Village/Town	1,984 (25%)	(19%)

[1]N (%); Mean (SD). [2](%); Mean (SD). For robustness checks presented in Online Appendix F.5, we applied post-stratification weights to match population margins. The second column in the table above includes descriptive statistics for these weighted data.

C Perceived Credibility of Treatment Messages in Study 3

C.1 Manipulation Checks

We use factual manipulation checks after each of our treatment messages, a practice recommended by Kane and Barabas (2019). For all respondents not assigned to the control group, we asked "What was depicted on the flyer you were just shown?" Respondents were then presented with five options (and were allowed to choose more than one):

- An infographic showing that the level of corruption in Ukraine is increasing
- An infographic showing that the level of corruption in Ukraine remains unchanged
- An infographic showing that the level corruption in Ukraine is decreasing
- A stop sign with the phrase "stop corruption"
- Don't know / Can't remember

For subjects assigned to messages with both a descriptive and injunctive norm, we considered the manipulation check passed if they indicated the correct trend in corruption and/or the stop sign. For subjects assigned to messages with only a descriptive or injunctive norm, we considered the manipulation check passed only if they indicated the correct corruption trend or the stop sign image, respectively. As seen in Table C-1, passage rates were high across all treatment conditions. The injunctive norm treatment exhibited the lowest rate, largely because subjects often inferred a corruption trend even though the flyer showed only a stop sign with the phrase "stop corruption."

Table C-1: Manipulation Check Passage Rates by Treatment Arm

Treatment Arm	Pct Passed	Total N
Decreasing (D)	91.0%	721
Increasing (I)	93.8%	727
Injunctive (J)	79.6%	633
Decreasing-Injunctive (D-J)	90.9%	2787
Increasing-Injunctive (I-J)	97.0%	747

C.2 Effects Conditional on Perceived Credibility with Alternative Thresholds

As discussed in the article's subsection "Accounting for Perceived Credibility of Treatment Messages," subjects' perception of a message's credibility is non-randomly distributed. Accordingly, comparing willingness to bribe among those who were treated and found the treatment credible to subjects from the control group would likely produce biased estimates. We instead divided subjects in each of the four treatment arms that included descriptive-norm messaging about corruption trends into two sub-groups, one for subjects who report that they perceived the messages as credible and the other for those who did not. After multiply imputing missing values for pre-treatment covariates using the algorithm developed by van Buuren and Groothuis-Oudshoorn (2011), we employed Hainmueller's (2012) entropy balancing technique to re-weight the control group such that its pre-treatment covariate means and other sample moments match those in our sub-groups of treated subjects who found the treatment message credible.

We achieve balance on a large and varied set of pre-treatment covariates, including demographic variables for age, education, income, employment status, region, and size of home city, town, or village; attitudinal variables measuring pre-treatment beliefs about economic conditions, corruption levels, and corruption trends, as well as prior exposure to anti-corruption information; and scores on the Need for Cognition and Susceptibility to Persuasion Strategies indices. Nevertheless, as with all matching techniques, entropy balancing cannot ensure balance on unobservables. It does, however, minimize imbalances across subgroups with respect to the first, second, and potentially higher moments of the covariate distributions, thereby reducing model dependence for treatment effect estimates utilizing the weighted data (Hainmueller 2012). Absent a technique to experimentally manipulate perceptions of credibility, we believe this approach presents the most informative estimates of treatment effects adjusted for whether subjects found the treatment messages believable.

Table C-2: Study 3 (Wave 1) Entropy-Weighted Treatment Estimates with Alternative Thresholds

	ATT			
	Thresh = 2	Thresh = 3	Thresh =4	Omit 3/4
Control (C)	1.019***	0.982***	0.951***	0.951***
	(0.035)	(0.036)	(0.038)	(0.038)
Decreasing (D) – Credible	−0.059	0.014	0.112	0.112
	(0.081)	(0.102)	(0.159)	(0.159)
Decreasing (D) – Noncredible	−0.074	−0.060	−0.014	−0.065
	(0.080)	(0.073)	(0.071)	(0.091)
Increasing (I) – Credible	−0.057	−0.059	−0.108	−0.108
	(0.066)	(0.074)	(0.092)	(0.092)
Increasing (I) – Noncredible	−0.132	−0.096	−0.088	−0.160
	(0.113)	(0.095)	(0.078)	(0.119)
Decreasing-Injunctive (D-J) – Credible	−0.206***	−0.207***	−0.220***	−0.220***
	(0.047)	(0.053)	(0.066)	(0.066)
Decreasing-Injunctive (D-J) – Noncredible	−0.050	−0.073	−0.097*	−0.045
	(0.059)	(0.053)	(0.049)	(0.072)
Increasing-Injunctive (I-J) – Credible	−0.071	−0.030	−0.010	−0.010
	(0.066)	(0.080)	(0.097)	(0.097)
Increasing-Injunctive (I-J) – Noncredible	−0.140	−0.119	−0.012	−0.118
	(0.104)	(0.083)	(0.081)	(0.114)
Num.Obs.	7021	7021	7021	5497
Num.Imp.	5	5	5	5

Notes: Perceptions of credibility are based on a 6-point scale, where 1 refers to "absolutely do not believe" and 6 refers to "absolutely believe." Estimates presented in Table 6 of the main article text assign subjects reporting ratings of 4 or higher to the sub-groups of those who found the messages credible. The table above shows that results are robust if we instead use thresholds of 2 or higher, 3 or higher, or omit subjects with ratings of 3 and 4 and assign subjects reporting ratings of 5 or 6 (reporting ratings of 1 or 2) to the sub-groups of those who found the messages credible (not credible). $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Perceptions of credibility are based on a 6-point scale, where 1 refers to "absolutely do not believe" and 6 refers to "absolutely believe." Estimates presented in Table 6 of the main article text assign subjects reporting ratings of 4 or higher to the sub-group of those who found the messages credible. Table C-2 demonstrates that these findings — particularly the

most substantively noteworthy finding about the effect size of Treatment *D-J* on subjects perceiving the message as credible – are robust across a range of alternative thresholds.[2]

C.3 Predictors of Perceived Credibility

Although our findings suggest that Treatment *D-J* can be effective in contexts with endemic corruption among a subset of citizens who perceive the message as credible, the fact that this subgroup of citizens represents a minority of our sample attests to the limitations of anti-corruption campaigns relying on this messaging. However, Table C-3 provides evidence that campaigns could potentially target citizens most likely to perceive such messaging as credible. For the *D-J* message, it is clear that female and younger subjects — two demographic variables that are readily observable — fall into this category. Subjects with high scores on the susceptibility to persuasive strategies (STPS) index also are more likely to perceive the *D-J* message as credible. While psychological traits are more difficult to observe, it is increasingly feasible to identify attitudinal dispositions based on social media profiles. As anti-corruption information campaigns become more adept at social media strategies, nuanced targeting of appropriate messaging likely will become a viable part of their toolkits.

2. To compactly present results, we simultaneously estimate ATTs for all treatment arms using the *D-J* (Credible) sub-group as the focal treatment arm. This approach differs slightly from the approach used in the main text, but the two approaches produce nearly identical treatment effect estimates.

Table C-3: Predictors of Perceived Credibility of Treatment Messages in Study 3

	Demographic Covariates				All Covariates			
	Decreasing		Increasing		Decreasing		Increasing	
	Dec.	Dec. Inj.	Inc.	Inc. Inj.	Dec.	Dec. Inj.	Inc.	Inc. Inj.
Intercept	3.004***	3.307***	3.699***	3.739***	1.093+	2.853***	2.378**	2.414**
	(0.196)	(0.108)	(0.228)	(0.229)	(0.589)	(0.341)	(0.784)	(0.750)
Age	−0.018***	−0.019***	0.003	−0.008+	−0.019***	−0.019***	0.003	−0.007
	(0.003)	(0.002)	(0.004)	(0.004)	(0.003)	(0.002)	(0.004)	(0.005)
Woman	0.269**	0.227***	0.052	0.007	0.149	0.239***	−0.005	−0.043
	(0.101)	(0.053)	(0.119)	(0.122)	(0.107)	(0.057)	(0.128)	(0.128)
Higher Ed	0.054	−0.089	0.019	0.182	0.144	−0.075	0.037	0.170
	(0.111)	(0.058)	(0.126)	(0.128)	(0.113)	(0.061)	(0.133)	(0.132)
Survey in Rus.	−0.390**	−0.089	0.127	0.168	−0.342*	−0.090	0.082	0.187
	(0.149)	(0.084)	(0.196)	(0.207)	(0.150)	(0.088)	(0.196)	(0.211)
Home Lang: Ukr.	-	-	-	-	-	-	-	-
Home Lang: Rus.	−0.434*	−0.490***	−0.159	0.279	−0.299+	−0.446***	0.010	0.362
	(0.172)	(0.110)	(0.269)	(0.284)	(0.172)	(0.116)	(0.270)	(0.295)
Home Lang: Both	−0.165	−0.203*	−0.339+	0.409*	−0.100	−0.191*	−0.280	0.431*
	(0.149)	(0.079)	(0.189)	(0.183)	(0.150)	(0.082)	(0.191)	(0.188)
Home Lang: Other	−0.327	−0.421	0.496		0.044	−0.368	0.749+	
	(0.228)	(0.363)	(0.361)		(0.289)	(0.402)	(0.410)	
STPS Index					0.238***	0.108**	0.210**	0.225**
					(0.069)	(0.037)	(0.079)	(0.084)
NfC Index					0.030	0.039	0.044	−0.148
					(0.099)	(0.052)	(0.118)	(0.112)
Big 5-C					−0.008	−0.030	0.158	0.198*
					(0.068)	(0.041)	(0.104)	(0.090)
Big 5-N					0.020	−0.075*	0.080	−0.062
					(0.061)	(0.032)	(0.068)	(0.072)
Big 5-E					0.059	0.081*	0.069	0.046
					(0.061)	(0.034)	(0.078)	(0.071)
Big 5-A					0.341***	0.040	−0.025	0.065
					(0.069)	(0.038)	(0.080)	(0.081)
Big 5-O					−0.062	−0.020	−0.096	0.052
					(0.065)	(0.036)	(0.082)	(0.074)
Region: Central	-	-	-	-	-	-	-	-
Region: East	−0.023	−0.061	0.058	−0.125	−0.048	−0.080	0.000	−0.139
	(0.137)	(0.075)	(0.186)	(0.195)	(0.134)	(0.076)	(0.183)	(0.196)
Region: South	0.194	0.010	0.138	0.112	0.175	0.029	0.097	0.050
	(0.204)	(0.100)	(0.216)	(0.194)	(0.209)	(0.103)	(0.222)	(0.188)
Region: West	0.080	−0.069	0.215+	0.137	0.111	−0.089	0.189	0.121
	(0.122)	(0.063)	(0.129)	(0.135)	(0.122)	(0.064)	(0.130)	(0.137)
N	742	2816	724	706	709	2700	694	676
R^2	0.106	0.079	0.012	0.020	0.162	0.085	0.033	0.045
R^2 Adj.	0.092	0.076	−0.003	0.006	0.140	0.079	0.007	0.021

OLS regressions with robust standard errors in parentheses. $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

D Heterogeneous Treatment Effects

In the article's main text, we focus primarily on heterogeneous effects by age and, to a lesser extent, by prior beliefs about corruption. In this section we present a more comprehensive overview of our pre-registered subgroup hypotheses, the logic underlying them, and our findings. Table D-1 provides an overview of our results; marginal effects plots are not included here because of space constraints but are available upon request.

D.1 Prior Beliefs About Corruption

Our first set of pre-registered heterogeneous effects hypotheses concerned prior beliefs about corruption. Proceeding from the premise that respondents learn and update their beliefs when presented with new information (see, e.g., Lenz 2009), we predicted that information treatments should have a larger effect on individuals whose prior beliefs are at odds with the newly received evidence. We accordingly proposed that subjects who believe corruption levels are high (low) should experience greater effects when assigned to Treatments D or D-J (Treatments I or I-J) (S3-H5.1). (Enumeration of hypotheses follows our pre-analysis plan for Study 3.) We also distinguished between beliefs about corruption levels and beliefs about corruption trends, proposing that subjects who believe corruption is increasing (decreasing) should experience greater effects when assigned to Treatments D or D-J (Treatments I or I-J) (S3-H5.2). Finally, we analyzed the role of prior exposure to anti-corruption info, given Druckman and Leeper's (2012) finding that prior exposure to relevant information can dampen treatment effects (S3-H5.3). Table D-1 shows that we found mixed evidence for H5.1 and H5.2. (see also Online Appendix E). Meanwhile, subgroups effects were apparent for H5.3, but in the opposite of the predicted direction: Effects were larger for subjects who reported more prior exposure to anti-corruption information campaigns.

D.2 Psychological Traits & Susceptibility to Persuasion

Our second set of pre-registered heterogeneous effect hypotheses drew on an emerging literature about persuasive technologies. Kaptein et al. (2012) demonstrate using the susceptibility to persuasion strategies (STPS) scale that applying specific influence strategies to individuals most susceptible to a given message has proven highly effective in health interventions (e.g., promoting exercise). We accordingly hypothesized that treatment effects would be larger for subjects with higher STPS scores (S3-H4.1). Second, we followed Kaptein (2012, 23–24) in hypothesizing that individuals with a high Need for Cognition (NfC), who exhibit a stronger need to analyze and elaborate on information received, are more likely to scrutinize arguments with which they are presented — and consequently are less susceptible to social influence strategies (S3-H4.2). Third, although research on the impact of personality on susceptibility to persuasion offers contradictory findings(see Alkış and Temizel 2015; Gerber et al. 2013; Oyibo and Vassileva 2019), we focused on two propositions best supported by the existing, albeit limited, evidence: That individuals who have high agreeableness scores care about fitting in, which may make them more susceptible to messaging that emphasizes social norms (S3-H4.3), and individuals who have low openness scores care more about social conventions and consequently may be more susceptible to such messaging (S3-H4.4). Table D-1 shows that we find some evidence supporting hypotheses H4.2, H4.3, and H4.4, with the most robust support for H4.3.

Table D-1: Evaluating Pre-Registered Heterogeneous Effects Hypotheses

Hypothesis	*D*	*I*	*J*	*D*-*J*	*I*-*J*
Prior Beliefs About Corruption					
For subjects who believe corruption levels are high (low), effects for *D* and *D*-*J* (*I* and *I*-*J*) will be larger (S3-H5.1)	0	0		✔	0
For subjects who believe corruption is increasing (decreasing), effects for *D* and *D*-*J* (*I* and *I*-*J*) will be larger (S3-H5.2)	*x*	✔		***X***	***X***
Effects will be larger for subjects with less prior exposure to anti-corruption messaging (S3-H5.3)	***X***	***X***	***X***	***X***	***X***
Psychological Traits					
Effects will be larger for subjects with higher STPS scores (S3-H4.1)	✓	0	***X***	***X***	***X***
Effects will be larger for subjects with lower NtC scores (S3-H4.2)	✓	✓	✔	0	✓
Effects will be larger for subjects with higher Big 5 agreeableness scores (S3-H4.3)	*x*	✓	✔	✔	✔
Effects will be larger for subjects with lower Big 5 openness scores (S3-H4.4)	✓	✓	✔	✓	***X***
Demographics					
Effects will be larger for subjects 30 years of age or younger (S3-H6)	✓	✓	✔	✔	✔

Key: ✔ = finding consistent with prediction and statistically significant; ✓ = finding consistent with prediction but not statistically significant; 0 = no evidence of heterogeneous effects; *x* = finding contradicts prediction but is not statistically significant; ***X*** = finding contradicts prediction and is statistically significant.

Notes: Enumeration of hypotheses follows our Study 3 pre-analysis plan. For S3-H5.1 and S3-H5.2 we did not pre-register a prediction for the treatment *J* arm (see cells with shaded background), but did find evidence of larger and statistically significant effects among subjects believing corruption is widespread (S3-H5.1) or decreasing (S3-H5.2). For H6, the evidence strongly supports the prediction that treatment effects will be larger among younger citizens, but 30 years of age does not appear to be a critical threshold as we had hypothesized (see discussion in subsection D.3).

D.3 Demographics

Our final heterogeneous effects hypotheses focused on age and predicted that younger subjects would be more responsive to anti-corruption messaging. This, in part, reflected expectations that younger citizens' attitudes about corruption might be more malleable, a belief that has guided prominent anti-corruption reformers in the development of educational campaigns (Klitgaard 1988, 116–117). Given Ukraine's communist history, we also considered the effects of spending formative years under socialism, which may have made attitudes less malleable. Following the literature on Soviet and post-Soviet socialization (e.g., Mishler and Rose 2007; Pop-Eleches and Tucker 2017), we distinguished between citizens born before and after the Soviet collapse and hypothesized that we would observe larger effects among subjects 30 years of age or younger relative to older counterparts (S3-H6). As shown in Figure 4 in the main article, evidence strongly supports the prediction of larger treatment effects among younger subjects. We do not, however, find noteworthy differences between under-30 and over-30-year-olds; rather, the most significant distinction appears to be between under-50 and over-50-year-olds.[3] A detailed explanation for these heterogeneous effects by age is beyond this article's scope, but considering that the cohort of 50-somethings were entering their twenties as the Soviet Union collapsed, we may conjecture that the relevant socialization processes occur in early adulthood rather than childhood.

3. Diagnostics developed by Hainmueller et al. (2019) show no evidence of non-linear interaction effects.

E Supplementary Analyses of Potential Backfire Effects

As discussed in the article's "Heterogeneous Effects" section, we not only find no evidence of "backfire" effects in our primary results but also no evidence of such effects among subsets of subjects. This contrasts with Cheeseman and Peiffer's (2022) finding that citizens who believe corruption is widespread — what they refer to as "pessimistic perceivers" — are particularly susceptible to backfire effects.

Figure E-1 presents treatment effects conditional on pre-treatment beliefs about the percent of Ukrainians who gave a bribe in the last 12 months.[4] For Study 3, there is no evidence of heterogeneous effects of any type for three of the five treatment arms (D, I, and I-J). Heterogeneous effects are apparent for Treatments J and D-J, but there is no indication that any subset of subjects became more willing to engage in corruption. Moreover, these two treatments induced the largest *decrease* in bribe intent specifically among respondents with prior beliefs about corruption's prevalence — that is, specifically among Cheeseman and Peiffer's (2022) pessimistic perceivers. The findings from Study 1 exhibit a similar pattern of subgroup effects. To the extent that any subset of subjects appears prone to backfire effects in Study 1, it was those with prior beliefs about low levels of corruption (i.e., the opposite of pessimistic perceivers).[5] However, we interpret these Study 1 subgroup effects with caution. They are not statistically significant, our data is thin on the lower end of the support (i.e., relatively few subjects believed that corruption levels are low), and this suggestive finding failed to replicate in Study 3.

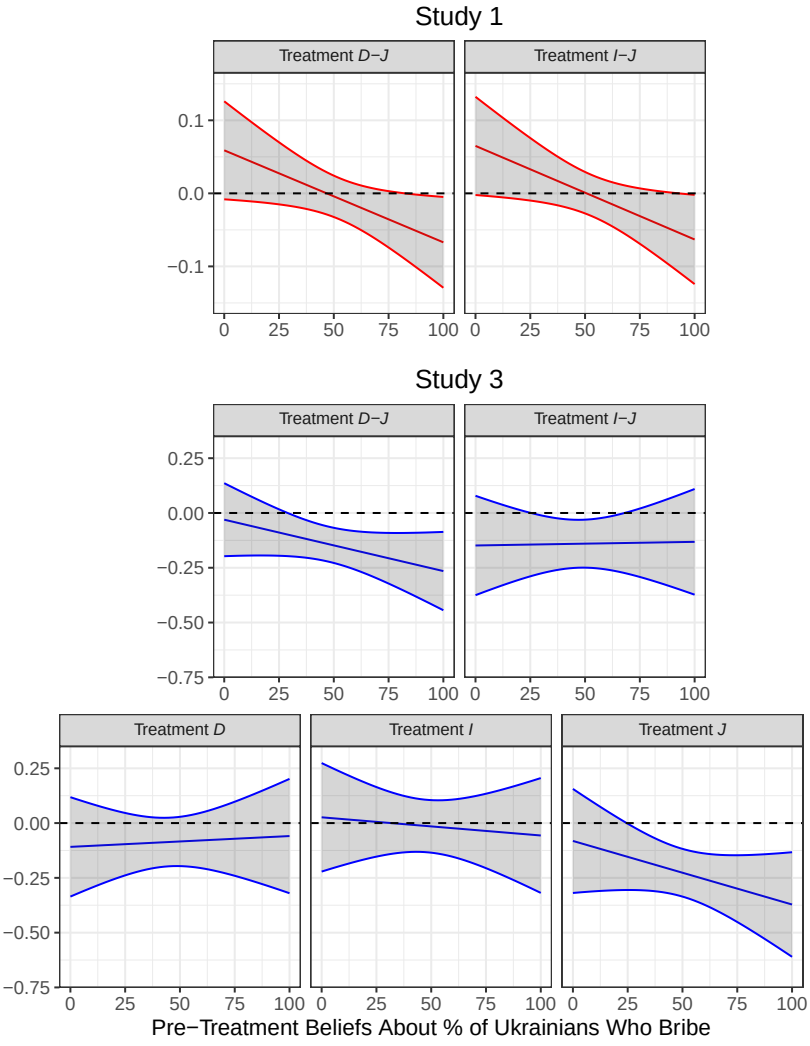
How concerned should anti-corruption practitioners be about backfire effects? Even though we found no backfire effects from the treatments most likely to increase willingness to engage in corruption (Treatment I in particular), we believe there is sufficient evidence overall to warrant avoiding messaging that emphasizes descriptive norms about high corruption levels. Indeed, it was the treatment message about widespread corruption that produced the largest backfire effects in Cheeseman and Peiffer's (2022) study. At a minimum, messaging should counteract potential backfire effects by combining such descriptive norms with injunctive norms.

By drawing attention to the risk of backfire effects, Cheeseman and Peiffer (2022) have made a valuable contribution to scholars' and practitioners' knowledge about anti-corruption messaging's effectiveness. But a critical question for future research will be to establish whether potential backfire effects are likely to result from well-known concerns about inadvertently counterproductive social norms messages, or whether practitioners should be wary of employing *any* anti-corruption messaging, as implied by Cheeseman and Peiffer's (2022) priming theory.

4. To make the survey question more accessible, we asked subjects their opinion about how many people out of a group of 100 typical Ukrainians gave a bribe in the last 12 months, rather than asking about percentages. To mitigate the risk of priming respondents in ways that could affect experimental results, we included several long batteries of questions unrelated to corruption between the corruption beliefs section and the experimental section of the survey.

5. As with our analyses of treatment effects by age, diagnostics proposed by Hainmueller et al. (2019) provide no evidence of non-linearities in these prior belief interaction effects.

Figure E-1: Marginal Effects Conditional on Pre-Treatment Beliefs About Corruption's Prevalence



Notes: The figures above plot treatment effects by beliefs about corruption's prevalence. All subgroup analyses control for gender, education level, language preference (Ukrainian vs. Russian), and region.

F Robustness Checks

F.1 Randomization Checks

Our analyses indicate that randomization for all three studies was successful.[6] For each of the studies, we report the p-values associated with the χ^2 statistic from a likelihood ratio test comparing regressions of the treatment indicators on a vector of covariates and models with no covariates. For all studies, we fail to reject the null hypothesis that the model with covariates does not explain treatment assignment (Study 1: $p = .89$; Study 2: $p = .57$; Study 3: $p = .63$).

F.2 Non-Response Rates by Treatment Arm for Study 1 & Study 3

For the survey item in Study 1 measuring willingness to bribe, enumerators presented respondents with three options: "yes", "no", and "prefer not to answer." Approximately 18% choice the non-response option,[7] but as shown in Table F-1, non-response rates did not vary by treatment arm. Study 2 was a lab study with near perfect compliance; accordingly, data for Study 2 is not included in Table F-1. Table F-1 also shows clearly that for Study 3, there is no evidence of heterogeneous non-response by treatment arm for Wave 1 or heterogeneous attrition between Wave 1 and Wave 2.[8]

Table F-1: Non-Response Rates by Treatment Arm

	Study 1			**Study 3 (W1)**			**Study 3 (W2)**		
Treatment	**Missing %**	**Missing *N***	**Total *N***	**Missing %**	**Missing *N***	**Total *N***	**Missing %**	**Missing *N***	**Total *N***
Control	18.1%	433	2,397	0.7%	12	1,688	60.3%	815	1,351
Decreasing Injunctive (*D*-*J*)	17.0%	372	2,191	1.1%	34	3,076			
Increasing Injunctive (*I*-*J*)	19.3%	451	2,338	1.3%	10	771			
Decreasing (*D*)				0.4%	3	794			
Increasing (*I*)				0.9%	7	776			
Injunctive (*J*)				0.8%	6	796			
t1: D-J / t2: C							61.6%	770	1,251
t1: D-J / t2: D-J							61.8%	763	1,234
Total	-	1,256	6,926	-	72	7,901	-	2,348	3,836

6. Influential methodologists have questioned the merits of covariate balance tests when treatment is randomized (Mutz et al. 2019), but we include them here for the sake of completeness.

7. Results are robust if we code non-responses as willingness to bribe. Indeed, this alternative coding produces larger treatment effects.

8. The difference between the Total N for Waves 1 and 2 results from a small percentage of respondents not clicking the final submit button at Wave 1's end, which was required to trigger the automated Wave 2 invitations. We again find no heterogeneity by treatment arm with respect to clicking the submit button.

F.3 Crosswise Model Used in Study 1

READ THE FOLLOWING TWO STATEMENTS TO YOURSELF BUT DO NOT ANSWER ALOUD

i. My mother was born in October, November, or December
ii. In order to receive my drivers license more quickly or easily, I would be willing to pay a bribe

NOW, THINKING ABOUT THESE TWO STATEMENTS, CHOOSE (A) OR (B):

A. one of the two statements is true
B. both statements are true OR neither statement is true

Remember, your mother's birthdate is unknown to any researchers involved in this survey. Your confidentiality is therefore guaranteed.

To examine the sensitivity of questions about bribery in the Ukrainian context, we employed a crosswise model in Study 1. The crosswise model elicits more honest answers by assuring respondents that researchers cannot observe individual responses to sensitive questions (see Tan et al. 2009). The crosswise model combines questions of a sensitive and non-sensitive nature, as shown in the box above. The respondent chooses option (a) or (b), neither of which requires the respondent to directly admit willingness to bribe. But as long as the proportion of the population answering positively to question (i) is known to the researchers, and as long as respondents' answers to statements (i) and (ii) are statistically independent, then it is feasible to estimate the proportion of respondents replying affirmatively to statement (ii).[9] The crosswise model, however, is not without drawbacks: It produces relatively imprecise estimates and requires complicated instructions.

Among respondents assigned to the study's control (C) arm, our crosswise model and direct question produced relatively similar and statistically indistinguishable results: 21.1% (SE 1.9%) for the crosswise model and 16.5% (SE 0.8%) for the direct question. (We present estimates from the control group so as to focus on social desirability bias among those who could not have been affected by the information treatments.)

Yet despite a dedicated training session for the survey firms' enumerators, diagnostics raise concerns about respondents' compliance with the crosswise model instructions. Assessments of social desirability bias usually assume one-sided lying. Individuals willing to bribe may answer dishonestly when asked directly, but individuals who are *un*willing to bribe should answer truthfully in response to indirect questions in the crosswise model format. In our data, however, comparison of the crosswise model and direct question results suggest that a non-trivial proportion of respondents who answered affirmatively about willingness to bribe when asked directly then answered in the negative to the same question posed in a format ensuring anonymity.

The most likely explanation for this result is that enumerators failed to adequately convey the logic of the crosswise model to respondents. Accordingly, to evaluate the robustness of our crosswise model estimates, we conduct a sensitivity analysis developed by Atsusaka and Stevenson (2023) to account for inattentive respondents. Accounting for inattentiveness further mitigates concerns about social desirability bias: The analysis shows that the

9. Formally, let π represent the proportion of respondents willing to bribe; $\overline{B}$, the proportion of respondents choosing 'B'; and p, the probability (i) is true. Then $\overline{B} = p\pi + (1-p)(1-\pi)$. Rearranging produces the estimate: $\hat{\pi} = \frac{\overline{B}+p-1}{2p-1}$. For an estimate of p, Ukrainian State Statistical Service birthrate data indicates that 24.86 percent of Ukrainians were born in October, November, or December.

estimate of 21.1% noted above, which assumes full comprehension and compliance with instructions, is an upper bound. The estimate falls as the percent of respondents assumed to be inattentive rises, converging with the direct question when 16% of subjects are assumed to be inattentive and falling below the lower bound of the direct question's confidence interval when approximately 20% or more respondents are assumed to be inattentive.

In summary, the crosswise model suggests that direct questions about bribery in Study 1 produced reliable estimates. However, for Study 3, a self-administered online survey, we expected similar or more severe implementation problems. We therefore adopted list experiments in place of crosswise models.

F.4 List Experiments Used in Study 3

To further examine bribery questions' sensitivity in the Ukrainian context, we ran two list experiments corresponding to two of the five bribery scenarios on which our Study 3 dependent variable is based: the scenarios about driver's licenses and medical services. Based on pilot surveys, we expected willingness to bribe for expediting a license to be among the lowest of the five scenarios, and willingness to bribe for expediting medical treatment to be among the highest.

The non-sensitive items were designed to be negatively correlated with each other to mitigate floor and ceiling effects (Glynn 2013). We randomized both the order of the two list experiments and the order in which respondents saw the list experiments and the direct questions about willingness to bribe, facilitating diagnostic tests (discussed below) of list experiments' identifying assumptions (Aronow et al. 2015; Blair and Imai 2012).

For each list experiment, respondents in the control group were presented with four items and respondents in the treatment group with five items, the additional item referring to the potentially sensitive topic of bribery:

`How many of the following activities would you be willing to do? You do not need to indicate WHICH activities, just HOW MANY.`

1. `Make a small donation to a charity that helps orphaned children`
2. `Give a hairdresser a tip worth two times the cost of the haircut`
3. `Give a small amount of money to a homeless person begging on the street`
4. `Pay higher taxes in order to improve the state healthcare system`

Sensitive item for first list experiment:

- `Give a bribe to avoid a long wait to receive a driver's license`

Sensitive item for second list experiment:

- `Give a bribe to avoid a long wait to receive medical treatment for a serious, but not life threatening, ailment`

Focusing on control group respondents in order to evaluate potential concerns about social desirability bias without conflating the effects of information treatments, we find that the list experiment and direct question estimates for the license scenario are similar and statistically indistinguishable: 21.9% (SE 5.0%) for the former and 14.9% (SE 0.8%) for the latter. For the healthcare scenario, however, the list experiment produces a far *lower* estimate than the direct question: 23.5% (SE 5.1%) compared to 41.3% (SE 1.2%).

Neither list experiments provides evidence of social desirability bias. However, a potential concern is that the low list experiment estimate for the healthcare scenario may indicate respondents' lack of compliance with the list experiment instructions. Indeed, although list experiments arguably are more straightforward to implement than crosswise models, recent research has found that the "no liars" assumption (i.e., the assumption of one-sided lying discussed in the preceding sub-section) is frequently violated when implementing list experiments (Li 2019) and that non-strategic respondent errors may bias list experiment estimates (Ahlquist 2018).

Diagnostics developed by Aronow et al. (2015) confirm that for both list experiments, but particularly for the healthcare scenario, the proportion of respondents indicating willingness to bribe in response to the direct question but not in response to the list experiment is suggestive of violations of the "no liars" assumption. We accordingly employ Li's (2019) method for bounding list experiment estimates while relaxing the assumption of one-sided lying. This method requires robustness to the order in which respondents are presented with the list experiments and direct questions, and unfortunately Aronow et al.'s (2015) diagnostic tests indicate that responses to the license scenario do not meet this condition.[10] However, for the healthcare scenario, Li's (2019) method produces prevalence bounds between 21% and 48%. Our direct question bribery estimate for the healthcare scenario is close to this upper bound, confirming that the list experiment produces no evidence of social desirability bias. In short, despite multiple robustness checks of results from two distinct sensitive survey techniques across multiple studies (Study 1 and Study 3), we find no evidence that questions about bribery in the Ukrainian context produce unreliable answers.

F.5 Regressions with Covariate Adjustment & Weights

Table F-2 shows that all findings presented in Table 4 of the article are robust to the inclusion of pre-treatment covariates (columns 1 through 4). For Study 1, we also show in column 5 that the effect of Treatment *D-J* is unchanged, albeit less precisely estimated, when applying survey sample weights. Columns 6 and 7 show that results when weighting the Study 3 sample to correspond with Ukrainian census data are consistent with our main findings but less robust. However, this decline in robustness results directly from our key finding about larger treatment effects among younger citizens combined with our Study 3 research design, which intentionally oversampled younger subjects in order to evaluate subgroup hypotheses about age effects (see Online Appendix D). Columns 8 through 11 of Table F-2 show that when considering under-50-year-olds, results with weighted or unweighted data are nearly identical.[11] In summary, these analyses indicate that our main effects findings generalize well to the broader Ukrainian population 50 years of age or younger, but should not be interpreted as representative of effects among the overall adult Ukrainian population.

10. It deserves emphasis that the sensitive survey technique implementation problems we encountered are widespread, especially for studies utilizing list experiments (see, e.g., Blair et al. 2020; Li 2019). Unfortunately, they are underdiagnosed and underreported.

11. Our Study 3 pre-analysis plan focused on distinctions between under- and over-30-year-olds. Weighted regressions on the subsample of subjects age 30 and younger produce similar results to those in Table F-2. However, we focus here on under-50-year-olds to show our findings' relatively broad generalizability and because 50 appears to be a more substantively important age threshold than 30 (see subsection D.3).

Table F-2: Main Effects with Covariate Adjustments and Weights

	With Covariates				With Weights			Under 50 (no Wgts)		Under 50 (w/ Wgts)	
	S1	S2	S3-W1	S3-W2	S1	S3-W1	S3-W2	S3-W1	S3-W2	S3-W1	S3-W2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Intercept	0.331***	0.577*	0.780***	0.729***	0.193***	1.200***	1.006***	1.135***	1.034***	1.214***	1.124***
	(0.027)	(0.262)	(0.068)	(0.137)	(0.020)	(0.058)	(0.077)	(0.039)	(0.064)	(0.052)	(0.093)
Decreasing (D)			−0.088			−0.032		−0.142*		−0.136	
			(0.059)			(0.099)		(0.066)		(0.088)	
Increasing (I)			−0.032			0.000		−0.079		−0.101	
			(0.062)			(0.092)		(0.069)		(0.089)	
Injunctive (J)			−0.227***			−0.140		−0.313***		−0.362***	
			(0.056)			(0.111)		(0.061)		(0.080)	
Decreasing-	−0.020+	−0.069*	−0.152***		−0.020	−0.084	0.019	−0.190***		−0.183**	
Injunctive (D-J)	(0.011)	(0.035)	(0.041)		(0.030)	(0.068)	(0.099)	(0.047)		(0.064)	
Increasing-	−0.013		−0.137*		−0.004	−0.032		−0.202**		−0.183*	
Injunctive (I-J)	(0.011)		(0.057)		(0.029)	(0.099)		(0.065)		(0.088)	
t1: D-J/				−0.072					−0.093		−0.182
t2: C				(0.078)					(0.092)		(0.122)
t1: D-J/				−0.074					−0.124		−0.128
t2: D-J				(0.080)					(0.090)		(0.128)
Age	−0.003***	−0.012	0.004***	−0.001							
	(0.000)	(0.013)	(0.001)	(0.002)							
Woman	−0.070***	−0.110**	0.001	0.041							
	(0.011)	(0.037)	(0.032)	(0.066)							
Higher Ed	0.014		0.136***	0.250***							
	(0.011)		(0.033)	(0.073)							
Survey in Rus.	−0.013		0.214***	0.341**							
	(0.018)		(0.053)	(0.119)							
Home Lang: Ukr.	-	-	-	-							
Home Lang: Rus.	0.081***	0.033	0.277***	0.227							
	(0.018)	(0.051)	(0.078)	(0.173)							
Home Lang: Both	0.022		0.148**	0.259**							
	(0.017)		(0.048)	(0.099)							
Region: Central	-	-	-	-							
Region: East	−0.050**	0.129+	0.055	−0.089							
	(0.015)	(0.068)	(0.048)	(0.103)							
Region: South	0.039*	0.063	0.031	−0.108							
	(0.016)	(0.049)	(0.058)	(0.126)							
Region: West	0.013	−0.037	0.021	−0.086							
	(0.016)	(0.048)	(0.035)	(0.072)							
N	5610	678	7286	1445	5670	7307	1446	5644	1134	5588	1129

Outcome variable for Study 3 (S3) is an index measuring the mean number of bribe scenarios to which subjects responded affirmatively, on a scale of 0 to 5. Outcome variables for Study 1 (S1) and Study 2 (S2) are dichotomous and measure the proportion of subjects reporting intention to bribe (Study 1) or engaging in a bribe transaction (in the Study 2 bribery game). Robust standard errors in parentheses. Multiple significance thresholds shown, where $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, but .05 is the pre-registered hypothesis test level.

G Supplementary Information About Research Instruments

G.1 Ukrainian & Russian Versions of Information Treatments

Ukrainian and Russian-language information treatments can be viewed here:
https://osf.io/gv4bm?view_only=bd9b3e96e7f34501a3d554f50441184d

G.2 Scripts for Bribery Game for Study 2

Scripts for the Study 2 bribery game can be viewed here:
https://osf.io/hyc5b?view_only=b88eede3d9cc440b90ee49fbb59d0780

G.3 Social Psychology Indices Used in Study 3

Susceptibility to Persuasion Strategies (STPS) scale

For each, respondents were asked to rate their level of agreement on a 1 to 5 scale. For more information, see Kaptein et al. (2012).

1. When I am in a new situation I look at others to see what I should do.
2. I often rely on other people to know what I should do.
3. It is important to me to fit in.
4. I will do something as long as I know there are others doing it too.

Need for Cognition (NfC)

For each, respondents were asked to rate their level of agreement on a 1 to 5 scale. For more information, see Lins de Holanda Coelho et al. (2020).

1. I would prefer complex to simple problems.
2. I like to have the responsibility of handling a situation that requires a lot of thinking.
3. Thinking is not my idea of fun.
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
5. I really enjoy a task that involves coming up with new solutions to problems.
6. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

Big 5 Personality Scale

For each, respondents were asked to rate the extent to which the pair of traits applies to them on a 1 to 5 scale. For more information see Gosling et al. (2003).

1. Extraverted, enthusiastic.
2. Critical, quarrelsome.
3. Dependable, self-disciplined.
4. Anxious, easily upset.
5. Open to new experiences, complex.
6. Reserved, quiet.
7. Sympathetic, warm.
8. Disorganized, careless.
9. Calm, emotionally stable.
10. Conventional, uncreative.

H Mapping to Pre-Analysis Plans

Pre-analysis plans were registered with the Open Science Framework (OSF) prior to beginning data collection for all three studies. Links to anonymized OSF registration pages are provided below. However, for Study 1 and Study 3, most details of our PAPs were included in documents we uploaded to the OSF registry that include authors' names. Accordingly, for peer review, we also provide separate links to anonymized versions of these documents. Other than the removal of authors' names, these anonymized versions are identical to those archived in the OSF registry.

Study 3 OSF anonymized link (note: PAP includes identifying information):
https://osf.io/bfypk/?view_only=f67d739f2f634e048276886a0b12bcd2
Registered: June 10, 2021

Study 3 anonymized PAP: https://osf.io/e32kz?view_only=be73650e70e24155853be5ef7a009de7

Study 2 OSF link (no identifying information):
https://osf.io/ry457/?view_only=e4803dd04af34b0bb34390a6520149bb
Registered: October 25, 2017

Study 1 OSF anonymized link (note: PAP includes identifying information):
https://osf.io/4eptm/?view_only=50ab7d8c7b984276b95bd47b4fa743c6
Registered: July 11, 2017[12]

Study 1 anonymized PAP: https://osf.io/mqyzn?view_only=b9bd4fad641b49a7897af66fc550eb06

Data collection and estimation strategies for all studies adhered to the procedures outlined in our PAPs. Below we map the four main findings discussed in the article text to our pre-registered hypotheses. We emphasize that (1) with the exception of our analyses conditional on treatments' perceived credibility, which we report as exploratory, all findings are based on pre-registered predictions, and (2) the findings discussed cover all primary hypotheses that we pre-registered (i.e., the article does not selectively report a subset of our predictions).

Considering each of our four main findings in turn:

1. *All treatments emphasizing injunctive norms (J, D-J, I-J) produced a modest yet temporary decrease in willingness to bribe.* See rows 1, 2, and 5 of the Main Predictions panel in Table H-1 and rows 1 and 2 of the Duration Effects panel.
2. *No treatments, including those emphasizing descriptive norms about rising corruption (I-J, I), produced "backfire" effects (i.e., increased willingness to bribe).* See rows 2 and 4 in the Main Predictions panel of Table H-1.
3. *Treatments combining descriptive norms about falling corruption with injunctive norms (D-J) produced larger decreases in willingness to bribe among subjects who perceived messages as credible.* As discussed in the main article text, we consider this conditional finding exploratory as it was not pre-registered.
4. *Treatment effects were larger among younger subjects.* See row 4 of the Heterogeneous Effects panel in Table H-1.

12. The PAP for Study 1 was initially filed with the EGAP registry and was imported into the OSF registry on February 3, 2020 after the EGAP and OSF registries were merged.

Table H-1: Overview of Pre-Registered Hypotheses

Hypothesis	Study 3 (S3)	Study 2 (S2)	Study 1 (S1)	Notes
Main Predictions				
(1) *D-J* will decrease bribe intent	**H1.1**	**H1**	*H1.1*	Evidence in favor of H1.1, but exploratory analyses in S3 show largest effects limited to subjects who perceive message as credible
(2) *I-J* will have no effect	H1.2	–	H1.2	Contrary to predictions, *I-J decreased* bribe intent in S1 and S3
(3) *D* will decrease bribe intent	H2.1	–	–	No evidence in favor of H2.1
(4) *I* will increase bribe intent	H2.2	–	–	No evidence in favor of H2.2
(5) *J* will decrease bribe intent	**H2.3**	–	–	Robust evidence in favor of H2.3
Secondary Predictions				
(1) *D-J* will cause larger decrease than *D*	H2.4	–	–	In line with prediction, *D-J* produced larger effect than *D*, but difference not statistically significant
(2) *D-J* will cause larger decrease than *J*	H2.5	–	–	Contrary to prediction, *J* produced *larger* effect than *D-J*
(3) *I-J* will cause larger increase than *I*	H2.6	–	–	H2.6 rendered moot by finding that *I-J* decreased bribe intent
Duration Effects				
(1) *D-J*/*C* will have no effect in *W*2	H3.1	–	–	In line with H3.1, there was no effect
(2) *D-J*/*D-J* will decrease bribe intent in *W*2	H3.2	–	–	No evidence in favor of H3.2 in main effects, but exploratory analyses found large effect among subjects who perceive message as credible
Heterogeneous Effects				
(1) Effects will be larger for subjects with higher STPS, lower NtC, higher agreeableness, lower openness	H4.1, *H4.2*, *H4.3*, *H4.4*	–	–	See Online Appendix Section D.2
(2) Effects will be larger for subjects whose priors about bribery levels/trends contrast with info in messages	*H5.1*, *H5.2*	–	*H4.1*, *H4.2*	See Online Appendix Section D.1
(3) Effects will be larger for subjects with less exposure to anti-corruption info	H5.3	–	–	See Online Appendix Section D.1
(4) Effects will be larger for subjects under 30	**H6**	–	–	See Online Appendix Section D.3

Notes: Bold font indicates that treatment effects were in the predicted direction and significant at $p < .05$ using two-tailed hypothesis tests. Italic font indicates partial confirmation of prediction or a p value less than .10 but greater than .05.

Table H-1 additionally provides details about all secondary hypotheses pre-registered for Study 3, some of which are not discussed in the main article text due to space constraints. For nearly all of these, the findings were not substantively significant or findings from primary hypotheses rendered analyses of the secondary hypotheses moot (see, e.g., notes about S3-2.6 in Table H-1). For additional discussion of hypotheses concerning heterogeneous effects, see Online Appendix Section D.[13]

13. No heterogeneous effects for Study 2 were pre-registered, and analyses of Study 1's pre-registered hypotheses about heterogeneous effects produce findings that largely overlap with those of Study 3 reported in Online Appendix D. All three PAPs included a section on mechanisms focused on messaging's effects on broader attitudes toward corruption. Analyses of these secondary hypotheses produced few robust findings and due to space constraints are not discussed here.

References

Ahlquist, John S. 2018. "List experiment design, non-strategic respondent error, and item count technique estimators." *Political Analysis* 26, no. 1 (January): 34–53.

Alkış, Nurcan, and Tuğba Taşkaya Temizel. 2015. "The impact of individual differences on influence strategies." *Personality and Individual Differences* 87:147–152.

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. "Combining list experiment and direct question estimates of sensitive behavior prevalence." *Journal of Survey Statistics and Methodology* 3, no. 1 (March): 43–66.

Atsusaka, Yuki, and Randolph T. Stevenson. 2023. "A bias-corrected estimator for the crosswise model with inattentive respondents." *Political Analysis* 31, no. 1 (January): 134–148.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments." *American Political Science Review* 114, no. 4 (November): 1297–1315.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical analysis of list experiments." *Political Analysis* 20 (1): 47–77.

Cheeseman, Nic, and Caryn Peiffer. 2022. "The curse of good intentions: Why anticorruption messaging can encourage bribery." *American Political Science Review* 116 (3): 1081–1095.

Druckman, James N., and Thomas J. Leeper. 2012. "Learning more from political communication experiments: Pretreatment and its effects." *American Journal of Political Science* 56 (4): 875–896.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, and Costas Panagopoulos. 2013. "Big five personality traits and responses to persuasive appeals: Results from voter turnout experiments." *Political Behavior* 35 (4): 687–728.

Glynn, Adam N. 2013. "What can we learn with statistical truth serum? Design and analysis of the list experiment." *Public Opinion Quarterly* 77 (S1): 159–172.

Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. 2003. "A very brief measure of the Big-Five personality domains." *Journal of Research in Personality* 37 (6): 504–528.

Hainmueller, Jens. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* 20 (1): 25–46.

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.

Kane, John V., and Jason Barabas. 2019. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." *American Journal of Political Science* 63 (1): 234–249.

Kaptein, Maurits. 2012. "Personalized persuasion in ambient intelligence," Eindhoven University of Technology.

Kaptein, Maurits, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. 2012. "Adaptive persuasive systems: A study of tailored persuasive text messages to reduce snacking." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2 (2): 1–25.

Klitgaard, Robert. 1988. *Controlling corruption.* University of California Press.

Lenz, Gabriel S. 2009. "Learning and opinion change, not priming: Reconsidering the priming hypothesis." *American Journal of Political Science* 53 (4): 821–837.

Li, Yimeng. 2019. "Relaxing the no liars assumption in list experiment analyses." *Political Analysis* 27, no. 4 (October): 540–555.

Lins de Holanda Coelho, Gabriel, Paul Hanel, and Lukas Wolf. 2020. "The very efficient assessment of need for cognition: Developing a six-item version." *Assessment* 27 (8): 1870–1885.

Mishler, William, and Richard Rose. 2007. "Generation, age, and time: The dynamics of political learning during Russia's transformation." *American Journal of Political Science* 51 (4): 822–834.

Mutz, Diana C., Robin Pemantle, and Philip Pham. 2019. "The perils of balance testing in experimental design: Messy analyses of clean data." *The American Statistician* 73 (1): 32–42.

Oyibo, Kiemute, and Julita Vassileva. 2019. "The relationship between personality traits and susceptibility to social influence." *Computers in Human Behavior* 98:174–188.

Pop-Eleches, Grigore, and Joshua Tucker. 2017. *Communism's shadow: Historical legacies and contemporary political attitudes.* Princeton University Press.

Tan, Ming T., Guo-Liang Tian, and Man-Lai Tang. 2009. "Sample surveys with sensitive questions: A nonrandomized response approach." *The American Statistician* 63 (1): 9–16.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate imputation by chained equations in R." *Journal of Statistical Software* 45 (3).