Monitoring Corruption and Overcoming the Collective Action Problem: Experimental Evidence from Pakistan (Updated PAP after the January 2019 Pilot)*

Michael Denly† Torben Behmer‡

Mobin Piracha§ Adityamohan Tantravahi¶

January 28, 2020

Abstract

Does monitoring corruption work? If so, which types of monitoring schemes are most effective at reducing corruption? To answer these questions, we distinguish between monitoring schemes that provide: (i) collective benefits that are subject to top-down principal-agent challenges; and (ii) private benefits that are subject to horizontal accountability challenges. Consistent with recent research emphasizing the drawbacks of principal-agent approaches, we model corruption as a multiple-equilibria collective action problem, using stability sets to adjudicate between the likelihood of different equilibria. To test the utility of our model, we undertake a related conjoint experiment on relatively poor Pakistani factory workers, a demographic that is frequently solicited for bribes. Consistent with our model, we expect to find that monitoring schemes with either collective or private benefits increase citizens' willingness to refuse to pay bribes. However, we expect to find that the most effective monitoring schemes in fomenting citizen-level collective action against corruption combine both collective and private benefits. Such results would suggest that monitoring corruption is not only a good use of scarce resources but that success is mainly a matter of policy design.

*For a summary of the design changes in this Pre-Analysis Plan, see Appendix F. For excellent research assistance, we thank Josh Evans and Ethan Tenison. For advice or feedback, we thank Megan Farrell, Mike Findley, Michael Gibbs, Joelean Hall, Juan Sebastian Lozano, Xiaobo Lu, Alex Norris, Nimrah Siddiqui, and Scott Wolford. Michael Denly gratefully acknowledges financial assistance from the University of Texas at Austin's Ward Endowed Fellowship. All errors are those of the authors.

†PhD student and Graduate Research Fellow, Innovations for Peace and Development, University of Texas at Austin. mdenly@utexas.edu

‡PhD student, University of Texas at Austin. tbehmer@gmail.com

§Undergraduate student and Research Affiliate, Innovations for Peace and Development, University of Texas at Austin. mobinapiracha@gmail.com

¶Undergraduate student and Research Affiliate, Innovations for Peace and Development, University of Texas at Austin. amtantravahi@gmail.com

Each year, corruption costs the world economy about USD 3.6 trillion or 5% of world gross domestic product (GDP), and bribes account for circa USD 1 trillion of that prodigious sum of money (United Nations, 2018). Research is also clear that the the most susceptible to paying bribes are the politically powerless and the poor, especially given their high levels of interaction with the state (Justesen and Bjørnskov, 2014; Peiffer and Rose, 2018; Robinson and Seim, 2018). Corruption is thus a costly and regressive phenomenon, afflicting nearly every country and sector of the world economy.

Since at least the 1990s, the development community has invested significantly in anti-corruption efforts, and a large part of those efforts has involved increased monitoring in line with the principal-agent model (Naím, 1995; Marquette, 2007; Hough, 2013; Levy, 2014). At its core, the principal-agent model suggests that it is possible to mitigate corruption through the monitoring and sanctioning of corrupt actors. However, recent research questions principal-agent approaches, suggesting that there is a shortage of people with the willingness to monitor and sanction corruption in corrupt societies (Persson, Rothstein and Teorell, 2013).[1] This shortage of what Peiffer and Alvarez (2016) call "principled principals" is mainly not due to resource constraints but informal institutions, such as norms, expectations, and beliefs. They tend to be immutable, making it difficult for even the best anti-corruption reformers and programs to drastically change trajectories in corrupt societies (Rothstein, 2011*b*; Fisman and Golden, 2017).[2]

In light of the above drawbacks to the principal-agent approach, many influential scholars argue that the collective action model better conceptualizes the challenge of overcoming corruption (e.g. Rothstein, 2011*a*; Mungiu-Pippidi, 2013; Fisman and Golden, 2017).[3] Under the collective action model, contributing to a public good such as reducing corruption is in

[1] For related earlier work, see Andvig and Moene (1990), Aidt (2003), and Kingston (2008). For related empirical work documenting how rule-breaking is contagious and generally higher in corrupt societies, see Gino, Ayal and Ariely (2009), Gächter and Schulz (2016), and Shalvi (2016).

[2] In more technical terms, we are referring to the fact that corruption is a collective action problem of the second order. For more on collective action problems of the second order, see Ostrom (1998), Rothstein (2011*a*,*b*) and Persson, Rothstein and Teorell (2013).

[3] Collier (2000) makes a related but not identical argument.

the best interest of society as a whole. By the same token, contributing to that public good incurs individual-level costs that people want to avoid, and the only way to minimize these costs is by people contributing collectively. Theoretically, any well-designed anti-corruption program can encourage collective action, regardless of the context, so scholars have convincingly argued that corruption is a multiple-equilibria phenomenon (e.g. Persson, Rothstein and Teorell, 2013; Fisman and Golden, 2017).

Given that scholars typically associate monitoring with the principal-agent model, in this study we aim to discern how monitoring corruption integrates with the collective action framework. In doing so, we ask: Does monitoring corruption work? If so, which types of monitoring schemes are most effective at reducing corruption? To answer these questions, we distinguish between monitoring schemes that provide: (i) collective benefits that are subject to top-down principal-agent challenges (i.e. "unprincipled principals"); and (ii) private benefits that do not map well to the principal-agent framework and are subject to horizontal accountability challenges (Fox, 2015, 347).[4] Consistent with the above research emphasizing the drawbacks of principal-agent approaches to corruption, we model corruption as a collective action problem with multiple equilibria. To adjudicate between the likelihood of different equilibria, we employ the stability sets method in line with Medina's (2007) lay person's theory of collective action.

The benefits of our modeling approach for understanding monitoring's contribution toward collective against corruption are manifold. First, we improve real-world correspondence by modeling corruption as a multiple-equilibrium behavior. In turn, our approach allows for variation in the outcome of interest: whether people take (costly) action against corruption. By contrast, a model with a unique equilibrium such as a prisoner's dilemma in the spirit of Olson (1965) cannot account for such outcome variation (Medina, 2007; Persson, Rothstein and Teorell, 2013). Second, our model incorporates Schelling's (1978) fundamental insight from focal points and tipping games (the threshold model) on contingent behavior: that is, a

[4] Horizontal accountability refers to the ability of the bureaucracy or different branches of government to exert checks and balances on itself (O'Donnell, 1998).

person's decision to take action against corruption depends' on his/her belief on whether others will reciprocate (Gneezy, Saccardo and van Veldhuizen, 2019; Sundström, 2019). Third, our model improves upon the tipping games approach by allowing the outcome to change as the environment changes. By extension, our model allows for more empirically grounded predictions that account for individuals' psychological benefits and differential costs of taking action against corruption.[5] Fourth, our incorporation of Medina's (2007) stability sets method allows us to derive comparative statics and, in turn, the likelihood that monitoring corruption will affect collective action choices. Finally, our model not only allows us to make predictions about monitoring but also its mechanisms, which we decompose into: (i) the collective benefits of monitoring, entailing greater top-down supervision of agents; and (ii) the private (individual) of benefits of monitoring, entailing horizontal accountability protection through a government anti-corruption office or a non-state actor. The collective benefits are subject to "unprincipled principals", and the private benefits are contingent upon the effectiveness of the relevant accountability channels. Nevertheless, we show that both types of monitoring interventions foment citizen-level collective action against corruption, and the most effective monitoring interventions combine collective and private benefits.

The model's predictions directly contrast Rothstein's (2011*a*) conjecture that incremental, monitoring-based approaches crowd out collective action against corruption.[6] The principal-agent and collective action models are indeed complementary (Marquette and Peiffer, 2018). However, the benefits of monitoring corruption in fomenting collective action against corruption are best understood as probabilistic, not absolute. In other words, we agree with Persson, Rothstein and Teorell (2013) that, by itself, monitoring generally cannot

[5] By "differential costs of taking action against corryption", we mean that for some individuals it will be more costly to take action against corruption than it will be for others. More specifically, some individuals may receive a higher "duty" payoff or "psychic benefit" from taking that costly action than others, and the relevant psychic benefit likely depends on the outcome associated with taking that action (Medina, 2007, 35). For example, a successful anti-corruption protest that brings about changes may bring more psychological benefits to participants than an unsuccessful anti-corruption that brings about no changes or spurs a government response such as repression.

[6] To be fair, Rothstein (2011*a*) at least partly walks back this conjecture in Persson, Rothstein and Teorell (2012) and other later work (e.g. Persson, Rothstein and Teorell, 2013, 2019).

eliminate high-corruption equilibria in line with a game of harmony.[7] However, monitoring can render these equilibria less likely to occur.

We test our model's predictions using an experimental design. It is ideal because whether monitoring contributes to collective action against corruption is dependent on covariates that are endogenous to corruption, and an experiment helps overcome such challenges. The experimental method that best maps to our theoretical model is the conjoint experiment. In contrast to traditional field, survey, and lab experiments, the conjoint experiment does not estimate an Average Treatment Effect (ATE). Instead, our conjoint experiment allows for estimation of the Average Marginal Component Effect (AMCE) of monitoring: that is, the marginal effect of monitoring averaged over the joint distribution of the remaining covariates (Hainmueller, Hopkins and Yamamoto, 2014, 10). The AMCE, not the ATE, is the primary estimand of interest for a simple reason: it is improbable that citizens' decisions to take costly action against corruption are independent of, for example, bribe amounts and initial beliefs about that action being worthwhile.

We will conduct all of these experiments on factory workers in Pakistan, a country with high levels of corruption according to Transparency International (2018). Although our sample is not randomly selected, most participants in the sample are relatively poor, politically powerless, and less well-educated—exactly the demographic that is most susceptible to corruption (Justesen and Bjørnskov, 2014; Peiffer and Rose, 2018; Robinson and Seim, 2018). Data from our pilot confirm that participants are frequently solicited for bribes in the course of their daily lives, thereby confirming our sample's relevance for answering our question.

This pre-analysis plan proceeds as follows. In Section 1, we define corruption, show how its definition relates to the principal-agent model, and then detail the model's main challenge for explaining successful anti-corruption efforts. Section 2 first explains why the collective action framework is useful for explaining corruption. We then explain how our

[7] Readers who are unfamiliar with basic collective action games, including the game of harmony, may refer to Table 9 in Appendix D.

modeling strategy following Medina's (2007) stability sets and the lay person's theory of collective action improves upon the predictions of the threshold model. Section 3 lays out the game-theoretic model of our study. In doing so, we first present a collective action game without monitoring, and then alter it to allow for collective as well as, separately, individual and collective returns to monitoring, showing how the implications of each model vary even when the general collective action format is maintained (i.e., multiple equilibria continue to exist). Our model thus provides insight into how citizens' calculus changes with monitoring, indicating the conditions under which monitoring will yield more collective action against corruption. Section 4 provides more details on our research design, and the final Section specifies the necessary details of pre-analysis, as suggested by Olken (2015).

1. Corruption as a Principal-Agent Problem

The most common definition of corruption is the "misuse of public office [or entrusted power] for private gain" (e.g. Treisman, 2000; Transparency International, 2009).[8] The principal-agent model, along with most anti-corruption efforts, follow directly from that definition (Ugur and Dasgupta, 2011; Rose-Ackerman and Palifka, 2016, 9).

Under the first manifestation of the principal-agent model, a politician or highly-ranked bureaucrat (i.e., the principal) is entrusted with power to perform certain tasks, assuming that he/she will not misuse his/her power for private gain. Since the principal cannot accomplish all of the tasks by himself or herself, the principal delegates at least some of these tasks to lower-ranking bureaucrats (i.e., the agents). According to the model, these agents are self-interested and must have some informational advantage over principals. Otherwise, the gains from delegating would be minimal to none, and principals would not delegate

[8] As Dixit (2016) explains, this definition of corruption maps well onto bribery, which entails a supply side (i.e. those providing the bribes—the private sector) and a demand side (i.e. those accepting/requesting the bribes—the public sector). In reality, though, corruption entails much more than bribery. For example, corruption entails kickbacks, coercion/extortion, nepotism, cronyism, financial fraud, electoral fraud, collusion, obstruction, and patronage (Søreide, 2014). For more on the defintion of corruption, refer to, for example, Rose-Ackerman and Palifka (2016) and Fisman and Golden (2017).

(Hawkins et al., 2006, 25). By the same token, it is the task of principals to monitor the agents and ensure that the latter do not misuse their monopoly and discretionary power for private gain. Klitgaard's (1988) famous formula summarizes the accountability task at hand for principals:

$$Corruption = Monopoly + Discretion - Accountability \quad (1)$$

Under the second manifestation of the principal-agent model, the politician or high-ranking bureaucrats assume the role of agent, and the public becomes the principal (Vaubel, 2006; Marquette and Peiffer, 2018). In democracies, the public can sanction corrupt agents by voting them out of office or by shaming them through the press, watchdog organizations, etc. In dictatorships, the public might rebel against the corrupt politicians or bureaucrats in whatever way possible.

The main issue with either manifestation of the principal-agent model is that it requires "benevolent" or "principled" principals: that is, bureaucrats, politicians, or a public with the will and capacity to monitor and sanction corrupt actors (Aidt, 2003; Peiffer and Alvarez, 2016). In societies where corruption is the predominant equilibrium behavior, "principled principals" with such qualities are in short supply. In most instances, shared norms, expectations, beliefs, and other informal institutional institutions tend to be stronger than any formal institutions designed to control corruption (Collier, 2000; Fisman and Golden, 2017). That is why, according to an influential article by Persson, Rothstein and Teorell (2013), most monitoring-based anti-corruption reforms and programs fail.

2. Corruption as a Collective Action Problem

A collective action problem "arise[s] when the individual pursuit of self-interest generates socially undesirable outcomes" (Ferguson, 2013, 4). Corruption is a collective action problem because most people in corrupt societies would benefit from having less corruption. By the same token, reducing corruption (i.e. contributing to the public good) is not in most people's individual self-interest. That is not just the case for the select people who financially benefit from corruption but also for its victims. Taking any type of action against corruption, such as refusing to pay a bribe, can carry potential costs such as intimidation, violence, inability to obtain government services, and being put on a blacklist (e.g. Wrong, 2009; Kingston, 2008; Stokes et al., 2013).

The prisoner's dilemma and assurance game provide most compelling macrostructural manifestations of corruption as a collective action problem (Yap, 2013; Dixit, 2018).[9] Although the prisoner's dilemma elucidates how citizens may want to free-ride on others' actions against corruption, the prisoner's dilemma has a fatal flaw: It necessarily entails a dominant strategy and unique equilibrium at defection (see Table 9 in Appendix D). Consequently, the prisoner's dilemma cannot explain variation in the outcome of whether citizens take action against corruption (Medina, 2007; Persson, Rothstein and Teorell, 2013, 2019). By contrast, the assurance game can explain such variation because of its multiple equilibria: one pure strategy Nash equilibrium at both parties taking action, another at doing nothing, as well as a mixed strategy equilibrium (Dixit, Skeath and Reiley, 2014; Persson, Rothstein and Teorell, 2013).[10]

The assurance game, however, does not explain how citizens' calculations can change about whether to take action against corruption. Fisman and Golden (2017, 5) correctly suggest that it relates to Schelling's (1978) fundamental insight from focal points and tipping games (i.e., the threshold model) on contingent behavior: that is, a person's decision to take action against corruption depends' on her belief on whether others will reciprocate (Dong,

[9]Ostensibly, the game of harmony does not describe corruption well in corrupt countries, because incentives to defect and thus free-ride on the contributions of others are still present; otherwise, corruption would not be so prevalent across the world. Deadlock is also inappropriate, since the payoffs of defecting are not that high for all citizens; if so, there would be no reason to mitigate corruption. The game of chicken is similarly uncompelling for describing corruption: taking action is the challenge, and chicken assumes that someone always takes action. Notably, there are two pure strategy Nash equilibria in Chicken at $Cooperate/Defect$ and $Defect/Cooperate$ (see Table 9 in Appendix D).

[10]For more on the origins of the assurance game, see Sen (1967).

Dulleck and Torgler, 2012; Lee and Guven, 2013; Banerjee, 2016; Gneezy, Saccardo and van Veldhuizen, 2019; Sundström, 2019).[11]

A main drawback of focal points and tipping games is that expectations and other determinants of contingent behavior vary with events in the external environment (Medina, 2007, 59-61). For example, even social trust, which numerous authors argue is the most important determinant of corruption (e.g. Rothstein, 2011*b*), changes according to external shocks—a new leader taking power, repression, scandals, etc. By construction, therefore, focal points and tipping games select on the dependent variable through the reliance on successful cases and cannot generate precise predictions when there are multiple equilibria (Medina, 2007, 60; Medina, 2018, 47).

To accomplish our goal of better understanding corruption as a collective action problem as well as whether monitoring can help overcome it, we adopt an alternative paradigm to guide our theory: Medina's (2007) lay person's theory of collective action. It not only incorporates multiple equilibria and contingent behavior but allows for precise calculations of comparative statics and the ability to adjudicate between equilibria. In its most basic form, Medina's (2007, 7) lay person's theory of collective action suggests:

> "When individuals can achieve some beneficial result by coordinating in a group, they are likely to coordinate. As the potential benefits of coordination increase (or the costs decrease), these individuals are more likely to coordinate and, conversely, as the potential benefits decrease (or the costs increase), they are less likely to do so."

[11] For a threshold model of revolution, see Kuran (1989).

3. A Model of Corruption as a Collective Action Problem

To precisely depict how different forms of monitoring probabilistically increase decisions to partake in costly collective action against corruption, we formally model corruption as a collective action problem. For ease of exposition, our game focuses on bureaucratic/petty corruption,[12] but it also applies to other forms of corruption, including grand corruption.[13] As with any collective action problem, all forms of corruption always entail two choices for citizens: (a) contribute to the public good by taking unilaterally costly action to reduce corruption—in the hope that others will follow suit; or (b) defect and pay no individual costs while refraining from contributing to the public good.

Our model entails a one-shot, two-player game between citizen i and all other citizens $-i$. We specifically avoid an iterative, multiplayer structure for reasons pertaining to substance and the literature. Substantively, corruption is a clandestine behavior that provides a very low signaling environment to facilitate citizens' short-term preference falsification.[14] Relatedly, no country has ever drastically and sustainably reduced corruption in the very short run outside of a regime change scenario (see Table 8 in Appendix A).[15] From the perspective of the literature, group size thresholds for overcoming corruption are context-

[12] Petty corruption refers to bribery by public officials when citizens try to "access basic goods or services in places like hospitals, schools, police departments and other agencies" (Transparency International, 2009).

[13] Grand corruption refers "collusion among the highest levels of government that involves major public sector projects, procurement, and large financial benefits among high-level public and private elites" (Bauhr and Charron, 2018). For more on the various forms of corruption, see Søreide (2014) and Rose-Ackerman and Palifka (2016).

[14] Kuran (1991) refers to the quick change between people's public and private preferences as preference falsification, which leads to imperfect observation (measured on a continuum) and, in turn, (political) surprises. For the case of corruption, the surprise would entail either: (a) immediate coordination on the $Cooperate/Cooperate$ equilibrium away from the $Defect/Defect$ equilibrium in an Assurance Game; or (b) an immediate switch to a game of Harmony (see Table 9 in Appendix D). Certainly, citizens observe others' behavior today, and *gradually* update their beliefs for the next period on the basis of past behavior. However, past research is clear the updating happens too gradually, if at all, for an iterated game to be more useful than a one-shot game (see Table 8 in Appendix A). For more work on virtuous circles of overcoming corruption, see Mungiu-Pippidi and Johnston (2017).

[15] Depending on how one defines the "very short run", it may be possible to consider Georgia, Estonia, and Hong Kong as exceptions. Regardless, such cases represent almost impossible rare events throughout the course of history (see Table 8 in Appendix A).

dependent, and relevant collective action scholarship offers no firm guidance as well (Sandler, 2015; Weimann et al., 2019), so we avoid group sizes for sake of generality.

3.1. The Basic Setup: Collective Action without Monitoring

Let Γ^A denote a game with the following basic assurance game payoffs for citizen i:[16]

- w_1, if and only if citizen i refuses to pay the bribe (i.e., she "cooperates"), and others join her in doing so. By refusing to pay the bribe, she (most likely) forgoes the opportunity to obtain the demanded service, but provides a public good. As others match her behavior, the benefits from the public good become more pronounced ($u_{public\ good} = t$). She does, however, face the risk of retaliation from the bureaucrat, for which she pays a cost, r. Beyond refusing to pay the bribe, the citizen could further choose to report the bureaucrat, thus producing a perhaps greater public good ($t' \geq t$), while also incurring larger costs from potential retaliation than refusing to pay a bribe ($r' > r$).[17] Mathematically, her payoff is: $w_1 \equiv u_i(C_i|C_{-i}) = t - r > 0$.

- w_2, if and only if citizen i pays the bribe (i.e., she "defects"), but other citizens opt to cooperate, thus providing for a corruption-fighting public good that she can free-ride on. By not supplying the public good herself, its benefit is reduced ($u_{public\ good} = s, t > s > 0$). By paying the bribe, however, citizen i virtually guarantees receiving the service, which she values with g. The (immediate financial) cost of her bribe is denoted by b. By paying the bribe, she also faces no risk of retaliation by the bureaucrat or anyone else. Hence, her payoff is: $w_2 \equiv u_i(D_i|C_{-i}) = s + g - b > 0$.

- w_3 if and only if citizen i cooperates by refusing to pay the bribe, thus investing into the

[16] Readers unfamiliar with the payoff structure of basic collective action games may refer to Table 9 in Appendix D.

[17] Theoretically, the citizen could refuse to pay a bribe, and separately assess whether or not to report the bureaucrat, or alternatively, pay the bribe but report the bureaucrat regardless. We explore this option in the empirical section of this paper. While we maintain a notation of r yielding retaliatory costs, and t yielding public utility of joint reporting, we should note that any increase in t (r) makes cooperative equilibria more (less) likely, respectively (see Appendix C).

public good, but her efforts are not matched by other citizens. As previously noted, refusal to pay results in foregone reception of services and possible retaliation. By singularly investing in the public good, her benefits from public good are small (s). In total, the action yields: $w_3 \equiv u_i(C_i|D_{-i}) = s - r$.

- w_4, if and only if neither citizen i nor others cooperate, no investment in the public good occurs. More concretely, all citizens receive their services in return for paying the bribe, a status quo outcome that produces no individual net gain or loss for either citizen i or others, making the (centered) payoff: $w_4 \equiv u_i(D_i|D_{-i}) = g - b = 0$.

For Game Γ^A to be an assurance game in line with Persson, Rothstein and Teorell (2013), first, the potential costs of retaliation exceed those of singular cooperation (i.e., $r > s$). Second, despite the threat of retaliation, the societal benefits of successful collective action must eclipse the payoff from singular defection (i.e., $w_1 > w_2$, or $t - r > s + g - b$).[18] Third, Game Γ^A must produce three equilibria: one in which citizen i and others cooperate, one in which all citizens defect, and one Mixed Strategy Nash Equilibria (MSNE). As we show in Appendix B.2, the existence of these pure strategy equilibria are contingent on these aforementioned assumptions. For Game Γ^A, as shown in Appendix B.3, the MSNE entails all citizens randomly choosing between cooperation and defection with (mixing) probability α_i, or α_{-i}, where:

$$\alpha_i^\star = \frac{g - b + r - s}{t - 2s} \tag{2}$$

As Medina (2007, 147) shows, we can represent the MSNE more generally as:

$$\alpha_i^\star = \frac{w_3 - w_4}{w_2 - w_4 + w_3 - w_1} \tag{3}$$

[18] For some, this assumption might be considered a strong one. Indeed, forgone services might be at the forefront of many citizens' thoughts as they contemplate the decision to refuse a bribe, despite the undeniable citizen benefits of systemically eradicating corruption. However, without this assumption, the game defaults to a prisoner's dilemma. As we outlined in Section 2, the prisoner's dilemma does not adequately describe the problem of corruption. See also Persson, Rothstein and Teorell (2013) and Fisman and Golden (2017).

Noting that $g - b$ is a scalar that only occurs in the numerator, and foreshadowing that monitoring does not affect the payoffs of citizens under unanimous defection, we can take advantage of the centering of our game. Mathematically, $w_4 = g - b = 0$, so the MSNE simply becomes:

$$\alpha_i^\star = \frac{r - s}{t - 2s} \tag{4}$$

3.2. Facilitating Cooperation: Collective Action with Monitoring

Building on the basic setup above, we introduce two functionally distinct monitoring mechanisms so as to alter the citizens' decision calculus regarding whether or not to take (costly) action against corruption. To begin, we start by illustrating how equilibrium decision making is affected by the introduction of collective benefits through monitoring, which are subject to diminshing returns due to "unprincipaled principals".[19] Let this game also have assurance game payoffs, and be denoted as Γ^B. Further let c refer to these collective benefits that accrue to all citizens, cooperative or not, as long as any citizen cooperates. Hence, payoffs w_1, w_2, and w_3 are all affected evenly, whereas w_4 remains constant, due to a lack of citizen cooperation. If these collective benefits are sufficiently large, the payoff from cooperating when others defect could theoretically exceed the payoff under mutual defection, which would violate the assumptions of the assurance game. To maintain the assurance game structure, and keeping with the centering around $w_4 = g - b = 0$, $s - r + c < 0$ replaces $s - r < 0$ in the revised game with monitoring.[20] In addition to pure strategy equilibria, we obtain the following MSNE:

$$\alpha_i^{\star'} = \frac{r - s - c}{t - 2s - c}, \quad \text{where } c > 0, \alpha_i^{\star'} < \alpha_i^\star \tag{5}$$

[19] More specifically, monitoring implicitly assumes that there is a higher level bureaucrat or official supervising the lower-level bureaucrat. This dynamic brings about top-down principal-agent challenges that diminished the returns of monitoring when the principals are "unprincipled" (Persson, Rothstein and Teorell, 2013; Peiffer and Alvarez, 2016). To the extent that citizens are weary of this dynamic, without the loss of generality, we can assume that as c approaches zero, and the added benefit is marginal at best.

[20] $s - r + c < 0$ is more stringent than $s - r < 0$ because $c \in (0, \infty)$.

Now, consider the setting in which benefits do not generate collectively, but rather, only accrue privately and are subject to horizontal accountability challenges.[21] Examples of such benefits, which we denote with p, include whistleblower protection, institutional oversight, and other protective services provided by the state or private actors such as NGOs. Let this game that introduces p be called Γ^C. In reference to Γ^A, only w_1 and w_3 increase evenly with parameter p, while all other payoffs remain unchanged. Just as we have shown for Γ^B, the inclusion of private benefits to those cooperating alters the minimal payoff, w_3, while holding w_4 constant. If these benefits are sufficiently large, the assumptions of an assurance game cannot be maintained (i.e., if $w_3 > w_4$). However, the literature is clear that in most where corruption is prevalent, corruption follows an assurance dilemma, not a game of harmony.[22] Keeping with the assurance game (and centering around $w_4 = 0$), we must now assume $s - r + p < 0$. The assumption is also more stringent than $s - r < 0$, since $p \in (0, \infty)$. In equilibrium, we obtain:

$$\alpha_i^{\star''} = \frac{r - s - p}{t - 2s}, \quad \text{where } p > 0, \alpha_i^{\star''} < \alpha_i^{\star} \tag{6}$$

Finally, consider another game, Γ^D, in which both private (p) and collective (c) benefits accrue to citizens in the event that cooperation takes place. Under Γ^D, w_1 and w_3 increase by $c + p$, whereas w_2 increases by c, and w_4 remains unchanged when compared to Γ^A. If $c+p$ is sufficiently large, the game ceases to take assurance game form, but that is generally not the case in a country where corruption is the norm (Persson, Rothstein and Teorell, 2013). Since w_1 increases at $p + c$, and w_2 increases at c, the remaining structure of the game remains in place. The proofs for the pure strategy equilibria for Γ^D are contained in Appendix B.5. Again solving for the MSNE (see Appendix B.6), and assuming a game

[21] Horizontal accountability refers to the ability of the bureaucracy to exert checks and balances on itself (O'Donnell, 1998).

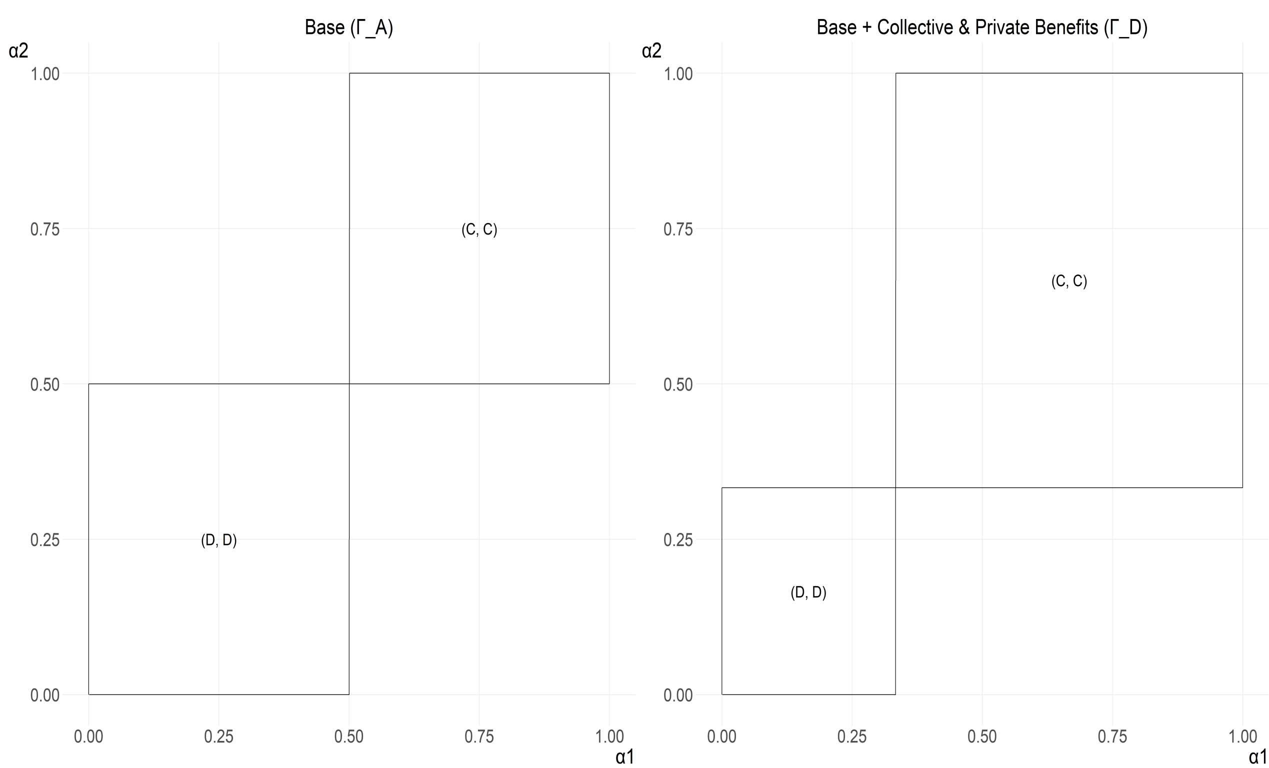
[22] In a game of harmony, cooperation is the only dominant strategy for all citizens. For more on the differences between harmony and assurance, see Table 9 in Appendix D.

centered around $w_4 = g - b = 0$, we obtain:

$$\alpha_i^{\star'''} = \frac{r - s - p - c}{t - 2s - c}, \quad \text{where } \forall\ p, c > 0: \quad \alpha_i^{\star'''} < \alpha_i^{\star'} < \alpha_i^{\star} \quad \text{and} \quad \alpha_i^{\star'''} < \alpha_i^{\star''} < \alpha_i^{\star} \tag{7}$$

In summary, we began by presenting corruption in the form of an assurance game, as the literature suggests (e.g. Persson, Rothstein and Teorell, 2013). Short of abandoning the structure of an assurance game, we show that introducing monitoring can nudge the MSNE closer towards zero. Irrespective of parameter size, this effect is additive—that is, the inclusion of collective *and* private benefits yields a greater move towards zero than the inclusion of any one parameter alone, *ceteris paribus*. This shift in the MSNEs and the resulting cooperative gains are most apparent in the best response plots showcased in Figure 1. These best response plots compare the basic setup of our game to an altered version with collective and private benefits from monitoring, demonstrating that the cooperative space for producing corruption-fighting public goods is much larger for the game with monitoring.

Figure 1: Best Response Plots



3.3. Adjudicating between Equilibria

As outlined above, introducing collective and private benefits to cooperation in the form of monitoring and protection alters citizens' mixing probability under the MSNE—even though it does not transform the game to one of harmony.[23] Next, we investigate which equilibrium is most likely to occur, so as to probe the central claim of this paper: that the combination of monitoring and protection can make taking costly action against corruption more likely. Substantively, it is unrealistic to assume that in any given period, countries randomly choose between the level of corruption, which is why simply proving the existence of these equilibria is not sufficient for the purposes of this study.[24]

To move from proving the existence of our three equilibria (two pure, one mixed) towards making a statement about their likelihood, we use Medina's (2007) stability sets method. To calculate the stability sets, we begin by stating the standard utility function under complete rationality ($\lambda = 1$) for game Γ^0 with multiple equilibria:

$$u_i^{0,\lambda=1}(\alpha_i, \alpha_{-i}) = \alpha_i(\alpha_{-i}w_1 + (1-\alpha_{-i})w_3) + (1-\alpha_i)(\alpha_{-i}w_2 + (1-\alpha_{-i})w_4) \tag{8}$$

Similar to tipping games, players in stability sets draw upon their prior beliefs that others will cooperate with them (denoted by β_{-i}), but stability sets entail even more precision. Rather than assuming complete rationality on behalf of the other citizens ($\lambda = 1$), calculating stability sets involves determining the optimal strategy for citizen i for all values of others' rationality and various levels of citizen beliefs ($\{\lambda, \beta_i\} \in (0,1)$). Consequently, the following utility function must be maximized:

$$\begin{aligned} u_i^{0,\lambda}(\alpha_i, \alpha_{-i}, \beta_i, \beta_{-i}) = {} & \lambda(\alpha_i(\alpha_{-i}w_1 + (1-\alpha_{-i})w_3) + (1-\alpha_i)(\alpha_{-i}w_2 + (1-\alpha_{-i})w_4)) + \\ & (1-\lambda)(\alpha_i(\beta_{-i}w_1 + (1-\beta_{-i})w_3) + (1-\alpha_i)(\beta_{-i}w_2 + (1-\beta_{-i})w_4)) \end{aligned} \tag{9}$$

[23] For more on basic collective action games, see Appendix D.
[24] For proofs regarding the existence of these equilibria, see Appendix B.

A clear benefit of the stability sets method is the researcher's ability to discern which of the pure strategy equilibria appear in the stability set of the game, conditional on initial beliefs, β. An equilibrium appears in the stability set of the game when it presents an optimal strategy for citizen i for all values of λ, not just for $\lambda = 1$. To calculate these stability sets, it is necessary to select initial beliefs about the other citizens' likely course of action (cooperation or defection), β_{-i}. To that end, we define $f(\beta_{-i})$ to be a probability distribution of these initial beliefs, which allows us to subsequently define the probability of the high-corruption equilibrium for any game (Γ^0) as follows:

$$Pr(High\ Corruption) = F_{\beta_{-i}}(\alpha_i^\star) = \int_0^{\alpha_i^\star} f(\beta_{-i})d\beta_{-i} \tag{10}$$

whereas the low-corruption equilibrium occurs in the stability set with the probability:

$$Pr(Low\ Corruption) = 1 - F_{\beta_{-i}}(\alpha_i^\star) = \int_{\alpha_i^\star}^1 f(\beta_{-i})d\beta_{-i} \tag{11}$$

Note that $F_{\beta_{-i}}$ are cumulative density functions that depend on the MSNEs of each game. Accordingly, it becomes evident that regardless of the assumed probability distribution over the initial beliefs, the closer the MSNEs are to zero, the less likely they are to fall within the stability set.[25]

Conditional on a given probability distribution over the above initial beliefs, we can also calculate the change in probability of the low-corruption equilibrium occurring across different games. Moving from our original setup, Γ^A, to a game where citizens receive both

[25] Without loss of generality, one might opt for a uniform distribution of initial beliefs in order to select these initial beliefs. In such a case, citizens are equally likely to believe that their fellow citizens will cooperate with 10% probability as they are to cooperate with 25% or 95%. The formation of these beliefs, as well as the probability distribution from which they are drawn are non-trivial, but go beyond the scope of this paper.

collective and private benefits to citizens in return for cooperation, Γ^D, we obtain:

$$\begin{aligned} Pr(Low\ Corruption\ Game\ D) - Pr(Low\ Corruption\ Game\ A) \\ = (1 - F_{\beta_{-i}}(\alpha_i^{'''\star})) - (1 - F_{\beta_{-i}}(\alpha_i^{'\star})) \\ = \int_{\alpha_i^{'''\star}}^1 f(\beta_{-i})d\beta_{-i} - \int_{\alpha_i^{'\star}}^1 f(\beta_{-i})d\beta_{-i} > 0 \end{aligned} \tag{12}$$

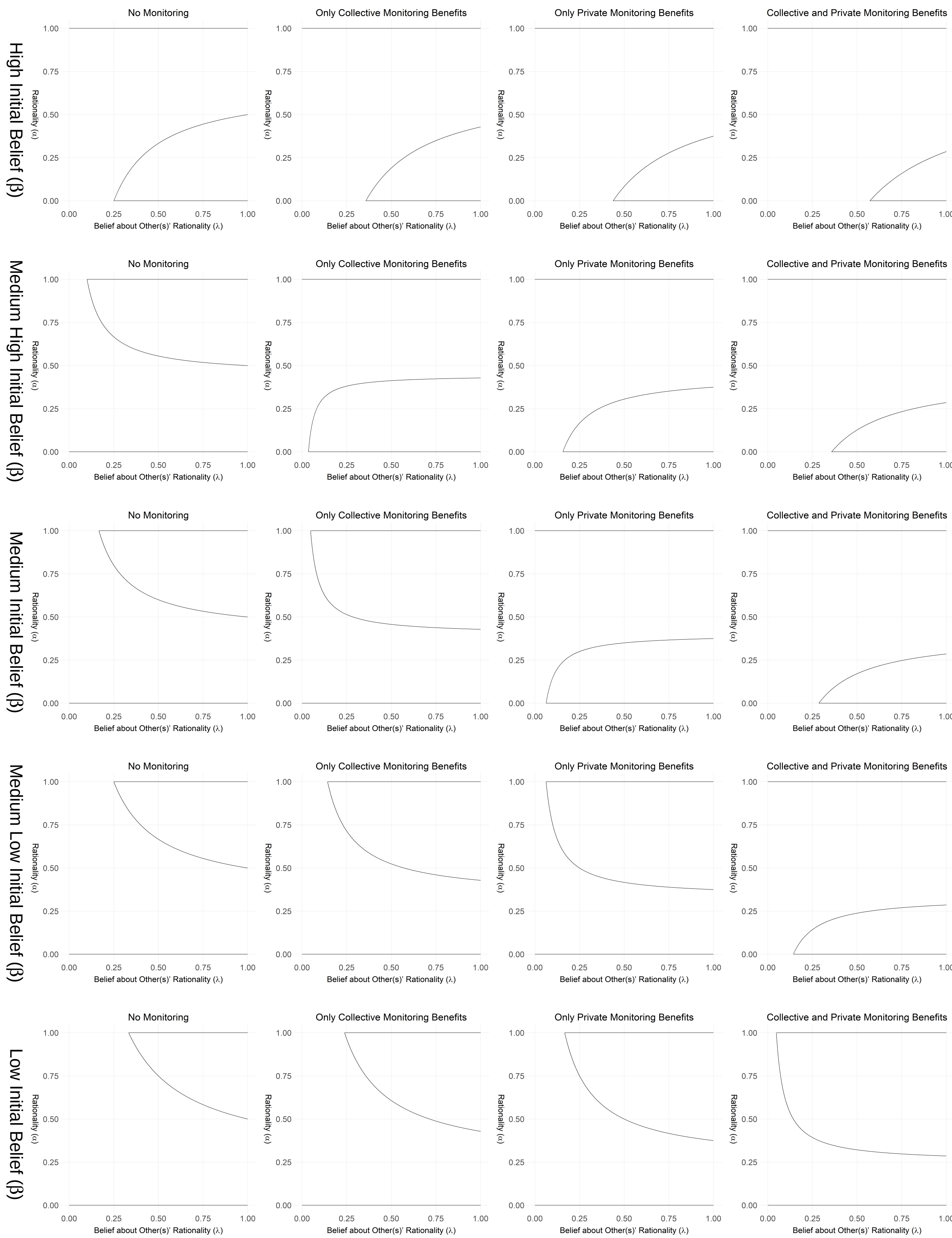
Using Harsanyian trace plots,[26] we graphically show the results of Equation (12) in Figure 2. Proofs for calculating the stability sets as well as slopes for the trace plots can be found in Appendix B.7. Overall, when beliefs over other citizens' behavior are sufficiently high, neither game results in high corruption. In such instances, the only consistently available equilibrium is (C, C), but when the beliefs are sufficiently low, the opposite holds true. By contrast, when the beliefs fall in between the MSNEs for games Γ^A and Γ^D, citizen behavior diverges. More specifically, given these beliefs, the cooperative equilibrium is in the stability set in the presence of monitoring Γ^D, and it is not in its absence Γ^A.

3.4. Summary

This formal section has shown how the introduction of collective and private benefits resulting from monitoring matters in the context of corruption. Even while maintaining the original game structure, and thus the presence of multiple equilibria, the introduction of monitoring (or protection) makes the occurrence of low-corruption equilibria more likely.

[26] For more on Harsayanian trace plots, see Harsanyi (1987).

Figure 2: Harsaniyan Trace Plots



4. Research Design

4.1. Data

To test our hypotheses about the utility of monitoring, we rely on a sample of factory workers from a large city in Pakistan. According to Transparency International (2018), Pakistan is a country with high levels of corruption, ranking 117/180 in the annual Corruption Perception Index. Of the 66 factory workers who participated in our pilot, 45 participants made less than 20,000 Rupees per month (USD 130 equivalent), and 21 participants made 20,000-50,000 Rupees per month (USD 130-320 equivalent). The participants in our sample can thus be classified as poor to middle class, and none are politically powerful, making them the exact type of demographic that is susceptible to bribery (Justesen and Bjørnskov, 2014; Peiffer and Rose, 2018; Robinson and Seim, 2018). In fact, results from pilot testing reveal that our demographic is frequently solicited for bribes in line with petty corruption in their daily lives. Therefore, our sample is appropriate for our research question.

For the final version of our experiment, we will plan on having a sample of approximately 600 different factory workers from those who participated in the pilot. Given that the final version of the experiment will entail a conjoint experiment (see below) in which each person takes the experiment 5 times, we will have a sample size of approximately 3,000 responses. See Section 5.9 for related power analysis. The data collection will begin on January 28, 2020.

4.2. Experiment Details

To test the utility of our model, we use an experimental design. It represents the best possible design choice given that our model requires the manipulation of a number of variables that are endogenous to corruption. Although social scientists have recently expressed a general preference for field experiments (e.g. Levitt and List, 2008), such a design is not ideal for the purposes of our study. First, adequately testing our model requires the randomizing of bribe amounts, something that is extremely unethical in the context of a field experiment. Second, field experiments generally estimate Average Treatment Effects (ATEs), not Average Marginal Component Effects (AMCEs): that is, the marginal effects of each parameter averaged over the joint distribution of the remaining covariates (Hainmueller, Hopkins and Yamamoto, 2014, 10). The AMCE, not the ATE, is our estimand of interest because citizen preferences about whether to take costly action against corruption are *multidimensional*. In line with our model in Section 3, preferences depend on a number of covariates that citizens simultaneously consider, such as levels of risk, bribes, and prior beliefs about the action being worthwhile. To estimate an ATE through a single-attribute design that only compares treatment to control would thus compromise the external validity of the study's findings. As Hainmueller, Hopkins and Yamamoto (2014, 2) summarize, "the conclusions based on such single-attribute designs may lack external validity if respondents' views when focused on a single component differ from those in more realistic scenarios in which they consider various components at once."

Against the above backdrop, we test our hypothesis our about the utility of monitoring and which of its mechanisms are most determinant in fomenting citizen-level collective action against corruption through a conjoint survey experiment. In doing so, we expose the respondents to multiple relevant attributes that can affect the decision of how to respond to a bribe request, thereby increasing realism by allowing for multidimensional preferences. Because conjoint experiments entail iteratively randomizing similar scenarios on the same participants, in some instances our conjoint analysis overcomes the fundamental problem of causal inference and actually provides the recipients with the exact counterfactuals.[27] In the process, we also reduce potential social desirability bias: that is, respondents not providing true responses but ones that are (more) socially acceptable. In contrast to typical survey experiments, conjoint analysis reduce social desirability concerns because it provides

[27] The fundamental problem of causal inference generally refers to the fact that it is impossible to observe something simultaneously in both treatment and control states (see Holland, 1986).

Table 1: Conjoint Attributes for the Driver's License Vignette (in Sequential Order)

Parameter/Introduction	Treatment Description(s)	Control Description
Introduction	You are in the waiting area at the driver's license office, waiting for your number to be called.	
Monitoring (Collective Benefit) equivalent to c parameter	You hear a loudspeaker announcing that **bribery is unlawful, and that the office is being monitored by video camera to prevent unlawful behavior from all government employees. You also see the video cameras.**	You hear a loudspeaker announcing **the weather report**.
Protection (Individual Benefit) **Government/NGO** equivalent to p parameter	The loudspeaker further announces if a government employee asks you for a bribe, **the National Accountability Bureau (NAB- anti-corruption office)/a local NGO** is offering to protect you from bureaucrat retaliation.	The loudspeaker further announces **the weather report**. (Note: Two consecutive controls entails only one mention of the weather report.)
Bribe **High/Low** equivalent to b parameter	Once your number is called, a government employee asks you to pay **2000/1000** rupees for the driver's license. However, you know that the true cost of the license is 500 rupees.	N/A
Initial Belief or Trust **Yes/No** equivalent to β parameter	You believe that **at least another person** in the waiting area is likely to support you if you protest in case the employee threatens you or refuses to provide you with the driver's license	You believe that **no one** in the waiting area is likely to support you if you protest in case the employee threatens you or refuses to provide you with the driver's license.

respondents with multiple potential reasons (cover) to justify sensitive choices (Hainmueller, Hopkins and Yamamoto, 2014, 14).[28]

Building on Bertrand et al. (2007) and Rothstein and Eek (2009), our conjoint experiments randomly assigns participants to one of two vignettes involving services that are often

[28] For other studies justifying conjoint analysis on the basis of its relatively superior ability to deal with potential social desirability bias, see, for example, Hainmueller, Hangartner and Yamamoto (2015), Auerbach and Thachil (2018), and Teele, Kalla and Rosenbluth (2018).

subject to petty corruption (bribery) in Pakistan: getting a driver's license (Table 1), or seeing the doctor in a public hospital for a common cold (Table 2).[29] We chose vignettes as opposed to the more common tabular presentation for our conjoint experiment. As Bansak et al. (2020) highlight, there are external validity concerns with the tabular presentation. For this experiment, it is more realistic to present collective action problems in a vignette than a tabular form. More specifically, it is unrealistic to believe that less-informed citizens, in particular, would approach each collective action problem in the perfectly rational way that the table conveys.

Each vignette informs participants about the true costs of the services and provides very similar language. The slight differences in language are to reflect how bribery actually happens in Pakistan. The driver's license vignette entails a bureaucrat asking participants to pay an inflated service fee, whereas the public hospital vignette entails a bureaucrat asking participants to pay to cut the line. After all, seeing the doctor for a common cold at a public hospital in Pakistan is generally free.

In each vignette, we randomly assign participants a monitoring treatment, a protection treatment, an initial belief about levels of citizen trust, and a high or low bribe amount based on actual levels of bribes for the respective services in Pakistan. We keep the number of attributes low not only to prevent satisficing (Bansak et al., 2018),[30] but also because potential attributes that we omit such as ethnicity and risk are subsumed within the trust and protection treatments. Since Pakistan has lower levels of literacy, totaling circa 60% according to the World Bank (2017), a loudspeaker announces the monitoring and protection

[29] During pilot testing, we also included the following tasks: getting a passport, reporting a crime to the police, and seeing the judge. We eliminated the passport task because most participants in our sample are not wealthy and generally do not need to travel out the country. We eliminated the crime reporting task because, after speaking to numerous participants during piloting, we could not find a crime in which all people would only report under certain circumstances. For example, people have to report a car being stolen for insurance purposes, and the extent to which people will report another person being robbed depends on a number of factors that are hard to measure. We similarly eliminated the seeing the judge task because of the inability to find a unique reason for seeing the judge that would be applicable to all participants.

[30] Satisficing refers to when respondents provide inaccurate responses because the cognitive load associated with too many conjoint attributes is too high (Bansak et al., 2018).

Table 2: Conjoint Attributes for the Public Hospital Vignette (in Sequential Order)

Parameter/Introduction	Treatment Description(s)	Control Description
Introduction	You are in the waiting area at the public hospital, waiting for your number to be called so you may be treated for a common cold.	
Monitoring (Collective Benefit) equivalent to c parameter	You hear a loudspeaker announcing **that bribery is unlawful, and that the office is being monitored by video camera to prevent unlawful behavior from all government employees. You also see the video cameras**.	You hear a loudspeaker announcing **the weather report**.
Protection (Individual Benefit) **Government/NGO** equivalent to p parameter	The loudspeaker further announces if a government employee asks you for a bribe, **the National Accountability Bureau (NAB- anti-corruption office)/a local NGO** is offering to protect you from bureaucrat retaliation.	The loudspeaker further announces **the weather report**. (Note: Two consecutive controls entails only one mention of the weather report.)
Bribe **High/Low** equivalent to b parameter	A hospital employee then approaches you, telling you that the line is very long and that to see the doctor you need to pay **500/250** rupees to cut the line—even though you know that the service is free.	N/A
Initial Belief or Trust **Yes/No** equivalent to β parameter	You believe that **at least another person** in the waiting area is likely to support you if you protest in case the employee threatens you or refuses to let you see the doctor.	You believe that **no one** in the waiting area is likely to support you if you protest in case the employee threatens you or refuses to let you see the doctor.

treatments in vignettes. The monitoring treatment follows Duflo, Hanna and Ryan (2012), using video cameras as disciplining devices for the bureaucrats. Given that citizens tend to trust state- and non-state actors differently (de la Cuesta et al., 2019), notably in Pakistan (Acemoglu et al., 2020), the protection treatments take both state and non-state forms. The non-state treatment entails protection from a local Non-Governmental Organization (NGO), whereas the state treatment entails protection from the National Accountability Bureau (NAB)—Pakistan's primary anti-corruption body, which has a wide purview akin to that of the U.S. Federal Bureau of Investigation (FBI).[31]

Especially since we model corruption as a collective action problem, the initial belief/trust treatment merits particular attention. As Marquette and Peiffer (2019, 812) emphasize, the literature arguing that corruption is (only) a collective action problem and that short-term monitoring is of a little help suggests that levels of citizen trust are the primary determinant of anti-corruption success (e.g. Persson, Rothstein and Teorell, 2013, 2019). We thus present this attribute that is most related to collective action last in the vignette so that participants are least likely to forget it. The trust treatment emphasizes whether the participant belief that at least another person will support him/her in case he/she decides to take any form of costly action against the corrupt bureaucrat. We are deliberately vague about the number of people given that the collective action literature still has not developed a firm conclusion about group sizes (Sandler, 2015). Also, this deliberate vagueness helps correspond with the game described earlier.

5. Pre-Analysis Section

In this section, we follow the guidance of Olken (2015) to specify the necessary requirements for pre-analysis.

5.1. Outcome Variables

The primary outcome variable for each iteration of the game is whether each individual decides to take (costly) unilateral action against the corrupt bureaucrat. Moreover, in the words of collective action theory, we aim to discern what factors make unilateral cooperation

[31] For more on Pakistan's anti-corruption authorities, see Khan (2018).

against corruption more likely despite uncertainty about other players' behavior. Thus, each outcome is binary: either the person can cooperate (i.e. contribute to a public good of attempting to reduce corruption) or defect (i.e. not contribute to a public good).

Since reporting a corrupt bureaucrat is a distinct behavior from simply refusing to pay a bribe, we ask participants separately for both the driver's license and public hospital scenarios: a) Will you pay [the inflated cost of] XXX rupees for the service? b) Will you report the [corrupt] bureaucrat? In cases where respondents are very unsure of their answer, we also provide them with the option of simply saying that they "don't know". Given that reporting may depend on the ability of the individual to report anonymously, we randomize whether the recipient must provide their name to the National Accountability Bureau or a local NGO. To further mitigate potential social desirability bias given that corruption is a sensitive topic, we re-ask each participant the same questions, substituting "you" for "most people you know." In order to mitigate the threat of respondent fatigue, we randomly assign respondents once into either the driver's license or the hospital condition, which then remains constant for all iterations. In addition we maintain constant treatment order for each vignette, such that respondents could learn where their critical attribute ought to occur, and consequently fatigue at a slower rate.

5.2. Treatments

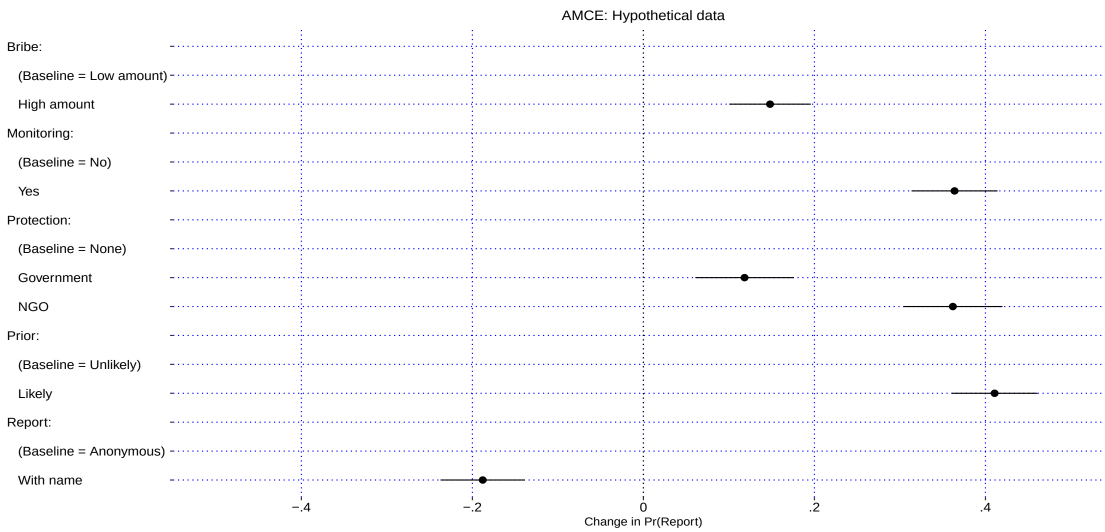
In addition to varying the reporting options in the aforementioned dependent variables, we explore how our key theoretical factors alter the respondents' willingness to cooperate. In line with our theoretical approach, our experiment does not aim to gauge how likely cooperation against corruption is in absolute terms. Rather, we aim to show that the introduction of individual and collective benefits to players make cooperation relatively more likely. We operationalize these individual benefits to cooperation by offering cooperators protection from bureaucrat retaliation. We then further distinguish between government- and NGO-provided protections, as mistrust in the government could render its protections ineffective. In addition, we include references to video cameras aimed at keeping an eye on bureaucrats, and to loudspeakers speaking to the illicit nature of corruption in the vignette so as to yield collective benefits to cooperation.

Moving beyond monitoring and protection, we include two additional attributes in the vignette. Firstly, following Medina's (2007) lay person's theory of collective action, we either include a small, or a substantially larger, bribe amount in the vignette, with the chosen amount having been chosen with the respective price of the public service that is to be rendered in mind. Note that traditional collective action theory would predict that as long as the original game form is maintained, such alterations in the payoffs should have no impact on the likelihood of cooperation (see, for example, Olson, 1965). The inclusion of the bribe amount as a treatment attribute allows us to both gauge to what extent this logic holds true for ordinary citizens, and also allows us to see how important our theoretically central attributes (monitoring and protection) are relative to the bribe amount. In line with conventional wisdom, higher bribe amounts should coincide with greater willingness to cooperate (i.e., refuse to pay the bribe, and/or report the bureaucrat), *ceteris paribus*.

As discussed in the previous section, we probe if anonymous reporting generates more reporting than reporting by name. The idea here is that reporting by name yields greater retaliatory threat potential than when anonymity is preserved. Note that if retaliation against reporting individuals is sufficiently severe, free riding on others' reporting becomes the individually opportunistic choice, as a prisoner's dilemma sets in, and cooperative behavior (i.e., reporting) becomes an off-equilibrium strategy. While we keep this retaliatory threat short of becoming prohibitively large in the theoretical model (so as to maintain the assurance game), naturally, we cannot assume that respondents will uniformly understand this threat as such. In fact, we make no assumptions that respondents understand which type of game they are playing under a given set of treatments.

Lastly, our vignette further contains an attribute informing the respondent about any prior likelihood that the respondent will be joined in their efforts of taking action against corrupt bureaucrats by other people in the waiting area. Since our theoretical section has shown that beliefs about others' willingness to cooperate plays an integral role in determining the equilibrium outcome, we include this variable both to gauge the importance of monitoring and protection, and also with the intent to reduce the need for satisficing on behalf of the respondents.[32] In line with the aforementioned predictions, we compute comparative statics from our model in Appendix C. They allow us to derive hypotheses, which we summarize graphically in Figure 3.

Figure 3: Hypothesized Average Marginal Component Effects on Simulated Data



5.3. Demographics

In this study, we keep track of a number of demographically-focused covariates, including: age, gender, income, education, political leanings, importance of elected government representatives, if they use the internet, position at the factory, and confidence in the courts to provide a fair trial (see Appendix E). Although our survey is limited to Pakistani factory workers, tracking these variables allows us to assess to what extent our sample represents the

[32]Note that unlike monitoring and protection payoffs, beliefs only begin to enter the calculations once the assumption of other players' rationality (λ) is relaxed. Since, MSNEs are computed under complete rationality by all actors, beliefs do not form a part of the MSNE, and thus cannot be regular calculations of comparative statics.

greater local population, and to uncover subgroup effect patterns that can motivate future research.[33] At the end of the experiment, we ask participants what they plan to do with their earnings, yielding a qualitative answer. It helps us uncover whether, potentially, there is a relationship between corruption and modernization; and whether the poor and middle class in Pakistan spend their money in a manner consistent with Banerjee and Duflo (2007, 2011).

5.4. Variable Definitions

Please refer to Tables 1 and 2 for the experimental treatments. The post-experimental questionnaire can be found in Appendix E. We make no prior predictions on these post-experimental variables, as they are not central to this project.

5.5. Inclusion/Exclusion Rules

The Pakistan factory experiment will accept any adult (18+ years old) factory worker who agrees to be part of the experiment.

5.6. Statistical Model Specification

See Hainmueller, Hopkins and Yamamoto (2014) for more on the conjoint model specification.

5.7. Subgroup Analysis

We plan on analyzing participants who are randomly assigned to the driver's license and doctor's office tasks separately. This is necessary because participants will likely value

[33] We purposely do not make multiple subgroup predictions for these demographic variables, because our conjoint involves so many potential permutations (see Table 3 in Section 5.9).

each service differently. As our model predicts, when citizens value a service exceedingly high, the payoff received under unilateral defection (D, C) could increase beyond the payoff received under mutual cooperation (C, C), such that the game could take a prisoner's dilemma format. That is, when the circumstances are such that citizens consider it imperative to receive services during hospital visits, their willingness to respond to our treatments might be markedly more muted than when the service is considered less critical. In addition, even if the assurance game is maintained, our model predicts less cooperation overall (see Appendix C), potentially leading to floor effects.

Greater familiarity with corruption in one of the two settings could further yield setting-specific results. Finally, visiting a driver's license office might trigger greater concern about engaging with the state apparatus than visiting a public hospital. As a consequence, respondents might be overall more fearful of recriminations (that is, larger r) in this setting, which should yield lesser willingness to cooperate overall (see Appendix C). Importantly, these hypotheses are secondary to our central argument. Moreover, since we're interested in the relative effectiveness of monitoring and protection, one would have to hypothesize if this effectiveness is more muted in a given setting or not. Potential floor and ceiling effects on the dependent variables in given settings notwithstanding, we are agnostic as to the settings' impact on the relative effectiveness of our key attributes.

5.8. Interactive Effects

Beyond calculating Average Marginal Component Effects (AMCEs), which lie at the core of our experimental analysis, we further hypothesize that the monitoring and protection treatments should be particularly strong when respondents can expect to be joined by others in cooperating. As we have argued in Section 3, initial beliefs held by the respondents are crucial to facilitating cooperation. Citizens who believe they will not be joined by others face greater threat of retaliation (vis-á-vis the reward), and only substantial changes to the payoff structure could trigger any willingness to cooperate. Including individual and collective benefits through monitoring and protection could alter this payoff structure and yield cooperation by itself. However, if citizens find it very likely that others will join them, beliefs may be high enough to facilitate much more cooperation. As a consequence, we will analyze Average Component Interaction Effects (ACIEs).[34] They help gauge how much more likely cooperation becomes when monitoring (and protection) are met by high initial beliefs about others' cooperation.

In a similar vein, we have argued that the bribe amount could theoretically be interpreted as exorbitantly high (given the value of the service), although we have intentionally chosen bribe amounts that citizens could be expected to pay regularly. In the event that the high bribe amounts are deemed excessive, respondents could see themselves trapped in a prisoner's dilemma, and consequently would never seek to cooperate. In those instances, monitoring and protection treatments should be more muted, although they could alleviate the situation. This outlined scenario is less likely to occur when bribe amounts are small, such that we would expect stronger effects of monitoring and protection under these circumstances. Again, we compute ACIEs to gauge the accuracy of our prediction.

5.9. Power Analysis

When planning the design of an experiment, a power analysis is one of the most fundamental tools to ensure replicability (Perugini, Gallucci and Costantini, 2018). Computing a power analysis when designing a study increases the probability of finding an effect that interests researchers and also increase the chances of obtaining accurate predictions. When researchers discover a low probability of power, it allows prudent researchers the opportunity to consider design changes.

Due to the design of this experiment, we conduct two power analyses, because a change in the dependent variable (report; not report) will increase the number of factors, and as a result, the number of treatments (see Table 3). More specifically, if the respondent chooses

[34] See Hainmueller, Hopkins and Yamamoto (2014)

to report the bureaucrat, we randomize whether they can report anonymously or must report with their name.

Table 3: Levels for the Power Analysis

Treatment/Block	**Levels**
Scenario	Driver's License, Public Hospital
Monitoring (Collective Benefit)	Yes, No
Protection (Individual Benefits)	Government, NGO, No
Bribe	High, Low
Initial Belief/Trust	Yes, No
Reporting	Anonymous, With Name

Conducting a power analysis for a factorial design, of which a conjoint experiment is a variant, is done similarly to a one-way ANOVA with some slight modifications (Perugini, Gallucci and Costantini, 2018). Because of the nature of the larger design, we specify the total number of groups in the design and degrees of freedom to calculate accurate power parameters. In our case, it is $2^5 3$—meaning there are 5 factors with 2 levels and 1 factor with 3 levels. The required N to gain .80 power requires a $k-1=5$ degrees of freedom for the effect, the total number of groups (i.e. $2*2*2*2*2*3=96$), and the effect size, η_p^2. The power of the interaction of the entire design (i.e. $(2-1)*(2-1)*(2-1)*(2-1)*(2-1)*(3-1)=2$) is the numerator for the degrees of freedom.

With regard to effect size, η_p^2 represents the variance explained by the effect as a proportion of the variance not explained by other effects. So, σ_f^2 as population variance explained by the effect and σ^2 as population residual variance gives us

$$\eta_p^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma^2} \tag{13}$$

Similar to a one-way design, the expected effect size is the population effect size, and, as such, we must consider the same empirical estimates. We adjust the sample eta-squared by calculating the partial epsilon-squared:

$$\epsilon_p^2 = 1 - (1 - \eta_p^2) * \frac{N - K + df}{N - K} \tag{14}$$

in which df are the degrees of freedom, and K is the total number of groups in the design. If our $2^5 3$ design has a total sample of 3000, and $\eta_p^2 = .2$, the formula will yield:

$$\epsilon_p^2 = 1 - (1 - .20) * \frac{3000 - 96 + 5}{3000 - 96} = 0.1986226 \tag{15}$$

When an effect size is not available, we can guess the variance explained and the residual variance as a proportion (Perugini, Gallucci and Costantini, 2018). To do this, we also need to guess the variance explained by other factors in the design, because that variance influences the residual variance.

In order to account for the correlation from repeated measures in our sample, we compute the design effect (C) and multiply that by the effect size specified to gain the real effect size due to the repeated measures. We compute design effect by:

$$C = \sqrt{\frac{R}{1 + (R - 1)\rho}} \tag{16}$$

where R is the number of repeated measures (5 in our case), and ρ is the correlation. We take the assumption of .5 as the correlation in line with Kubinec (2019). We then compute the effect size accounting for correlated responses by:

$$f = \frac{\eta_p^2}{1 - \eta_p^2} * \sqrt{\frac{R}{1 + (R - 1)\rho}} \tag{17}$$

which becomes:

$$f = \frac{.1}{1 - .1} * \sqrt{\frac{5}{1 + (5 - 1).5}} = 0.1434438 \tag{18}$$

Using the software G*Power, we input the following parameters into a priori analysis of a repeated measures, between factors ANOVA:

Table 4: G*Power Priori Analysis Inputs

Effect size	Number of Groups	Power	Number of Measurements	Correlation among repeated measures
0.1434438	96	.8	5	.5

We are provided with an output of:

Table 5: G*Power Priori Analysis Outputs

Non-centrality parameter	Critical F	Numerator df	Denominator df	Total Sample Size	Actual Power
42.7983374	1.2638322	95	1152	1248	0.8152058

The following graph shows the critical F point:

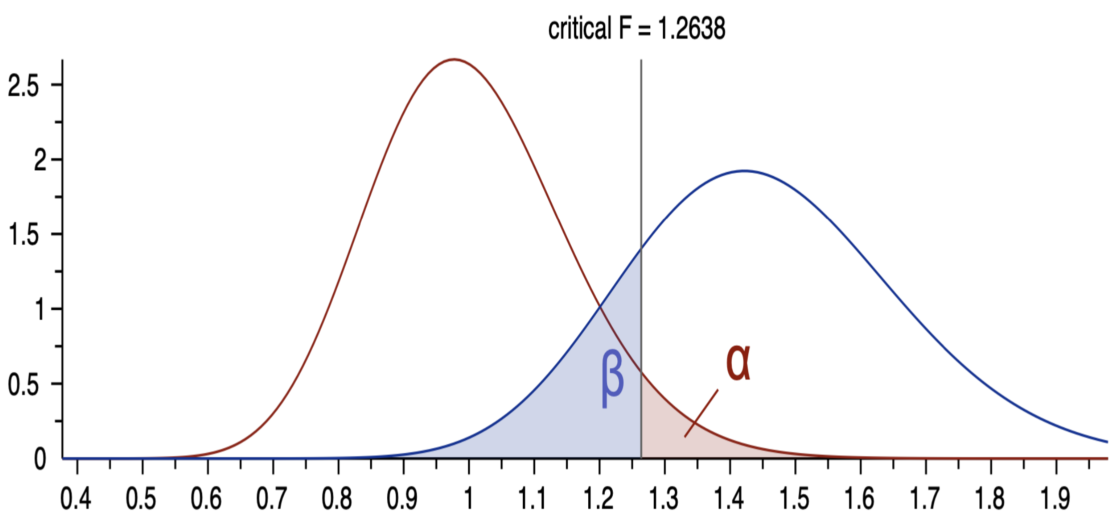
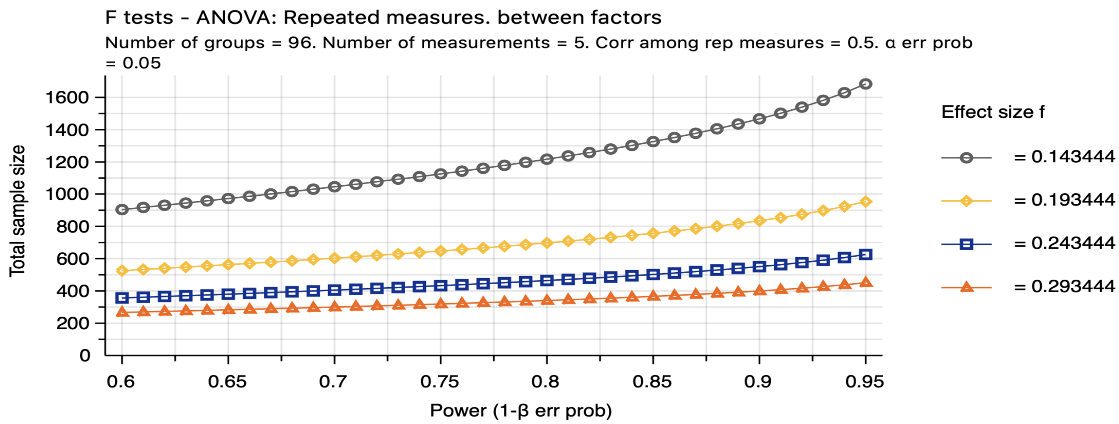


Figure 4: Critical F Point

In order to obtain 80% power with our $2^5 3$ design, we need a total sample size of 1,248. The power plot below shows the power obtained with 4 different effect sizes. You'll notice the larger the effect size, the smaller the sample size needed to obtain the same level of power.

Figure 5: Power Plot



Now we will calculate the number of responses needed to obtain 80% power when the DV is Report/Not Report. The same concepts above follow, except now the design notation of the experiment is $2^4 3$, meaning we have 4 factors with 2 levels and one factor with 3 levels. This design gives us a df of $k-1=4$ and $2*2*2*2*3=48$ treatments instead of 96. The power of the interactions of this design remains at $(2-1)*(2-1)*(2-1)*(2-1)*(3-1)=2$ as the numerator degrees of freedom.

Using G*Power, we input the following peramteres into a priori analysis of an ANOVA:

Table 6: G*Power Priori Analysis Inputs

Effect size	Number of Groups	Power	Number of Measurements	Correlation among repeated measures
0.1434438	48	.8	5	.5

We are provided with an output of:

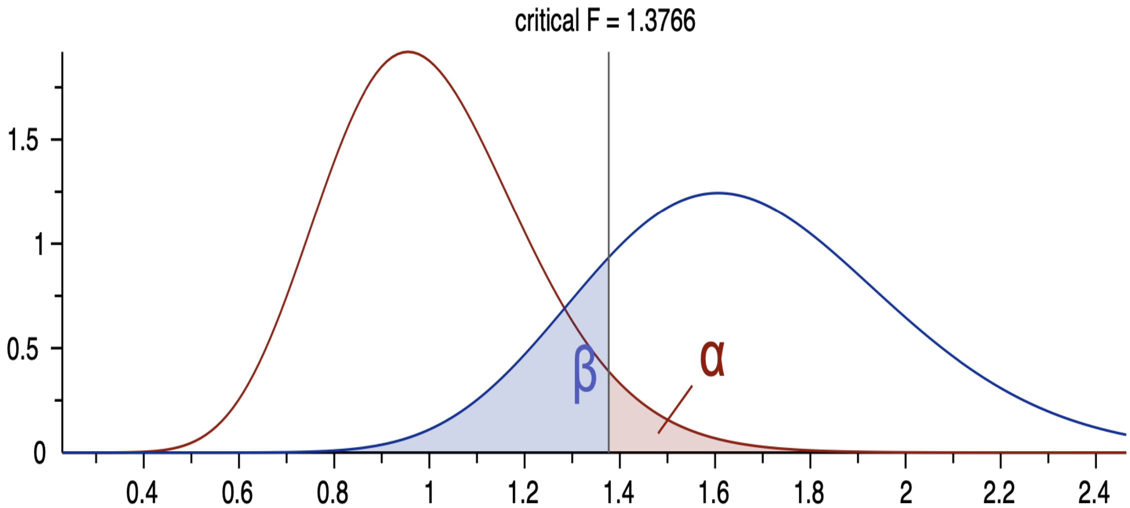
Table 7: G*Power Priori Analysis Outputs

Non-centrality parameter	Critical F	Numerator df	Denominator df	Total Sample Size	Actual Power
31.2757081	1.3766305	47	864	912	0.8133502

The following graph shows the critical F point.

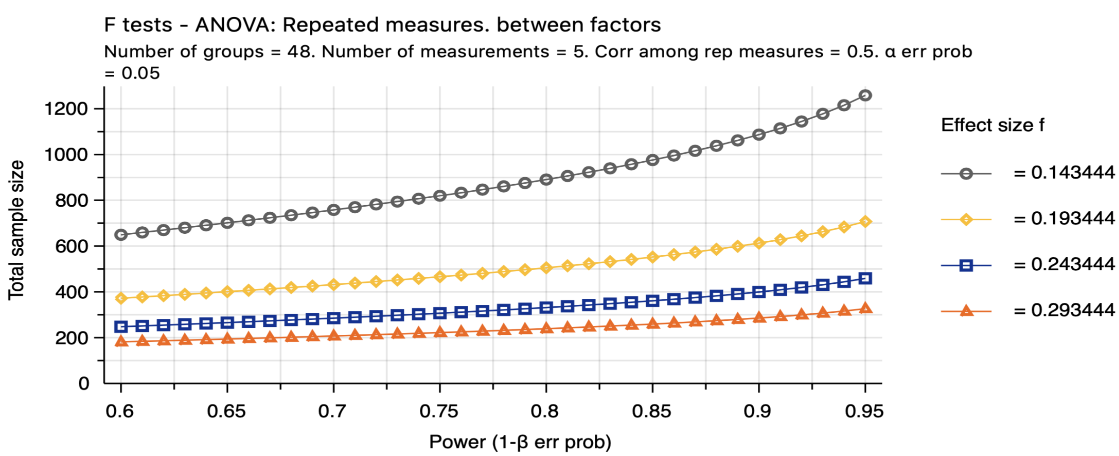
Like the previous analysis, in order to obtain 80% power with our $2^4 3$ design, we need

Figure 6: Critical F Point for 48 treatments



a total sample size of 864. The power plot below shows the power obtained with 4 different effect sizes. Overall, the larger the effect size, the smaller sample size needed to obtain the same level of power.

Figure 7: Power Plot for 48 treatments



A. How Countries Shifted to a Lower Corruption Equilibria

Table 8: Countries that Have Shifted to Lower Corruption Equilibria

Country	Critical Period(s)	How the Country Shifted to a Lower Corruption Equilibrium	Maintained?
Denmark	1658-1665, 1814, 1849	Loss in a wars against Sweden; top-down reform initiated by kings; drafting of a new constitution following demonstrations	Yes
Sweden	1810-1850	Losing the 1808-1809 war against Russia, followed by a series of reforms	Yes
Great Britain	1780-1883	Civil service reform; legislation; a secret ballot; suffrage reform, resulting in the decline of clientelism and more funds for public services	Yes
France	1791-1975	The French Revolution; gradual decline of patronage appointments; construction of impartial institutions	Yes
Germany	1919, 1945	End of World War I and World War II.	Yes
Spain	1975-1978	End of the dictatorship of Francisco Franco.	Yes
Portugal	1974	End of the Estado Novo dictatorship, fueled by the Carnation Revolution	Yes
Ireland	1995-2007, 2012	Privatization, market reforms, corruption law reforms introduced (2012), the rise of the Celtic Tiger ended up in increased corruption, Council of Europe's Greco initiative (anti-corruption program)	Yes
Italy	1992-1996	The Clean Hands scandal, prompted by the arrest of one well-connected individual, who provided information that led to to the arrest of hundreds and changed the party system	No
Estonia	1990-1995	Tax reform; e-governance; procurement reform; privatization	Yes

Continued on next page

Table 8: Countries that Have Shifted to Lower Corruption Equilibria – *continued*

Country	**Critical Period(s)**	**How the Country Shifted to a Lower Corruption Equilibrium**	**Maintained?**
Georgia	2004-2008	The Georgian Transition, followed by a "big bang" approach from President Mikhail Saakashvilli (i.e., large-scale dismissal of civil servants, televised arrests, and e-governance)	Yes, though with creeping authoritarianism and human rights issues
Tunisia	2011-2014	Citizen demonstrations over autocratic rule fueled the Arab Spring	Mostly, though some patronage remains a challenge
Botswana	1966-present	Excellent natural resource management; protection of property rights; transparent policy-making; management of potential ethnic tensions. Strong and independent judiciary, and an exceptional rule of law.	Regular scandals imperil progress. The discovery of rich resource deposits has also led to an increase in corruption.
United States	1870-1920	The regulation of patronage appointments through the Pendleton Act; the press; the Progressivist movement; successful prosecutions.	Yes, though the role of money in politics is significant
Hong Kong	1974-1977	Egregious malfeasance by the head of police, which prompted the creation of an independent anti-corruption agency and many subsequent arrests	Yes
Taiwan	1992-	Civil service reform; high-level corruption initiatives; legislation; party system change	Yes
Singapore	1959-1990	Authoritarian leader Lee Kwan Yew pushed through a series of reforms	Yes
South Korea	1961-2003	Education; import-substitution industrialization that fueled economic growth; market reforms; legislation; protests	Yes
Japan	1945-1993,	Loss of World War 2; MacArthur Plan; resolution of a series of corruption scandals involving kickbacks and vote-buying.	Yes

Continued on next page

Table 8: Countries that Have Shifted to Lower Corruption Equilibria – *continued*

Country	**Critical Period(s)**	**How the Country Shifted to a Lower Corruption Equilibrium**	**Maintained?**
Chile	1984-1990	Economic liberalization; privatization; loss of natural resource rents; democratic and authoritarian legacies from previous periods	Yes
Uruguay	1984	Fiscal/tax system consolidation; privatization; a democratic history; an educated and active citizenry; loss of patronage funds.	Yes
United Arab Emirates	1989-2016	E-Governance; institutional development; avoided the misuse of oil rents by investing in physical capital and institutional fabric, establishing of Prosecution of Public Funds Office to prosecute corruption, adopting a federal penal code anchored on UN Convention Against Corruption (UNCAC)	Yes
Namibia	2000-2010	Effective legislation (Anti-Corruption Act, 2003, under the ambit of the Constitution, Electoral Act, 2014) and enforcement bodies (e.g. the Anti-Corruption Commission)	Yes
Cape Verde	1990-present	A wave of democratization, introduction of parties, e-governance	Yes
Costa Rica	1948-present	civil war of 1948, followed by new regime and electoral institute to deter election fraud; improvement in professionalism of judicial branch staff; audit and constitutional improvements following scandals; signing of OECD anti-bribery measures	Yes

Sources: McCarthy (2003), Therialt (2003), Lizzeri and Persico (2004), Glaeser and Goldin (2007), Baker (2009), Rothstein (2011*b*), Acemoglu and Robinson (2012), Grindle (2012), Weyland (2012), Camp, Dixit and Stokes (2014), Al Athmay (2015), Arroyo Chacón (2015), Mungiu-Pippidi (2015, 2016), Polus, Kopinski and Tycholiz (2015), Rothstein and Teorell (2015), Soto and Haouas (2016), Teorell and Rothstein (2015), Soto and Haouas (2016), Fisman and Golden (2017), Fukuyama (2018), Masoud (2018), Nyblade and Reed (2008), Coppedge et al. (2018), Wilson and Villarreal (2017)

B. Mathematical Proofs for Theoretical Model

B.1. Basic Collective Action Game Γ^A

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/Take Action*	$t-r$, $t-r$	$s-r$, $g-b+s$
	Defect/Do Nothing	$g-b+s$, $s-r$	$g-b$, $g-b$

Where:

$$
\begin{aligned}
W_1 &= t - r \\
W_2 &= g - b + s \\
W_3 &= s - r \\
W_4 &= g - b;
\end{aligned}
\tag{19}
$$

$$
\begin{aligned}
W_1 &> W_4 > W_3, \\
W_2 &> W_4 > W_3,
\end{aligned}
\tag{20}
$$

And ("at least for some players") $W_1 > W_2$, in line with Medina's (2007, 53) assumptions of collective action games.

B.2. Pure Strategy Nash Equilibria for Game Γ^A

$\oplus$ (C, C) is a PSNE if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} u_i(C_i|C_{-i}) &> u_i(D_i|C_{-i}) \\ = t - r &> g - b + s, \\ = t &> s + r \text{ (centering game at zero (g-b=0))} \end{aligned} \tag{21}$$

$\oplus$ (D, D) is a PSNE, if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} u_i(D_i|D_{-i}) &> u_i(C_i|D_{-i}) \\ = g - b &> s - r, \\ = 0 &> s - r \text{ (centering game at zero (g-b=0))} \\ = r &> s \end{aligned} \tag{22}$$

B.3. Mixed Strategy Nash Equilibrium for Game Γ^A

Player i chooses to play C [i.e. report the bureaucrat/refuse to pay the bribe] with probability α_i, such that player j $(i \neq j)$ is indifferent between playing C [reporting/refusing to pay the bribe] and playing D [paying the bribe]

Eq. 2.1

$$\begin{aligned} EU_j(C) &= EU_j(D) \\ \alpha_i(W_1) + (1 - \alpha_i)(W_3) &= \alpha_i(W_2) + (1 - \alpha_i)(W_4) \\ \alpha_i(t - r) + (1 - \alpha_i)(s - r) &= \alpha_i(g - b + s) + (1 - \alpha_i)(g - b) \\ \alpha_i(t - s) + (s - r) &= \alpha_i(s) + (g - b) \\ \alpha_i &= \frac{g - b + r - s}{t - 2s} \end{aligned} \tag{23}$$

Which is a valid mixing probability under the previously made assumptions when players are fully rational. Specifically, centering the game at zero again, $\alpha_i = \frac{r-s}{t-2s}$, where $r - s > 0$ as $r > s$ by assumption (see above); and $t - 2s > 0$, since $t > s + r$, and $r > s$.

In such a case, their best response is to choose α_i as follows:

$$\alpha_i^*(\alpha_j) = \begin{cases} 1 \text{ if } \alpha_j > \frac{g-b+r-s}{t-2s} \\ [0,1] \text{ if } \alpha_j = \frac{g-b+r-s}{t-2s} \\ 0 \text{ if } \alpha_j < \frac{g-b+r-s}{t-2s} \end{cases} \tag{24}$$

B.4. Augmented Collective Action Game Γ^D

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/Take Action*	$t-r+p+c$, $t-r+p+c$	$s-r+p+c$, $g-b+s+c$
	Defect/Do Nothing	$g-b+s+c$, $s-r+p+c$	$g-b$, $g-b$

Where:

$$\begin{aligned} W_1 &= t - r + p + c \\ W_2 &= g - b + s + c \\ W_3 &= s - r + p + c \\ W_4 &= g - b \end{aligned} \tag{25}$$

And still:

$$\begin{aligned} &W_1 > W_4 > W_3, \\ &W_2 > W_4 > W_3 \end{aligned} \tag{26}$$

And $W_1 > W_2$,

and $\{t,\, r,\, g,\, b,\, s,\, p,\, c\} \in (0, \infty)$.

B.5. Pure Strategy Nash Equilibria for Game Γ^D

$\oplus$ (C, C) is a PSNE, because for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(C_i|C_{-i}) > u_i(D_i|C_{-i}) \\ & = t - r + p + c > g - b + s + c \\ & = t - r + p + c > s + c \text{ (applying centering at g-b = 0)} \\ & = t > s + r - p \text{ (True by base game's assumption, as } t > s + r) \end{aligned} \tag{27}$$

$\oplus$ (D, D) is a PSNE, if and only if for each player $i (i \in \{1, 2\})$:

$$\begin{aligned} & u_i(D_i|D_{-i}) > u_i(C_i|D_{-i}) \\ & = g - b > s - r + p + c \\ & = 0 > s - r + p + c \text{ (applying centering at g-b = 0)} \\ & = r - p - c > s \end{aligned} \tag{28}$$

B.6. Mixed Strategy Nash Equilibria for Game Γ^D

Again, player i chooses to play C with probability α_i such that player j is indifferent between playing C and D.

$$
\begin{aligned}
\Lsh EU_j(C) &= EU_j(D) \\
\Leftrightarrow \alpha_i(W_1) + (1-\alpha_i)(W_3) &= \alpha_i(W_2) + (1-\alpha_i)(W_4) \\
\Leftrightarrow \alpha_i(t-r+p+c) + (1-\alpha_i)(s-r+p+c) &= \alpha_i(g-b+s+c) + (1-\alpha_i)(g-b) \\
\Leftrightarrow \alpha_i(t-s) + (s-r+p+c) &= \alpha_i(s+c) + g - b \\
\Leftrightarrow \alpha_i(t-2s-c) &= g-b+r-s-p-c \\
\Leftrightarrow \alpha_i &= \frac{g-b+r-s-p-c}{t-2s-c}
\end{aligned} \tag{29}
$$

Which is a valid mixing probability as long as $t-2s-c>0$, and (as previously assumed) $r>s+p+c$. Again under full rationality player i's best response is:

$$
\alpha_i^*(\alpha_j) = \begin{cases} 1 \text{ if } \alpha_j > \frac{g-b+r-s-p-c}{t-2s-c} \\ [0,1] \text{ if } \alpha_j = \frac{g-b+r-s-p-c}{t-2s-c} \\ 0 \text{ if } \alpha_j < \frac{g-b+r-s-p-c}{t-2s-c} \end{cases} \tag{30}
$$

Generally, if players are not fully rational, their utility functions can be re-written as a weighted average of their payoffs under rationality and their payoffs given their beliefs about the likelihood of cooperation by the other player (B):

Eq 5.1

$$
\begin{aligned}
u_i^\lambda(\alpha_i,\alpha_j,\beta_i,\beta_j) = &\lambda[\alpha_i(\alpha_j(W_1) + (1-\alpha_j)(w_3)) \\
&+ (1-\alpha_i)(\alpha_j(w_2) + (1-\alpha_j)(w_4))] \\
&+ (1-\lambda)[\alpha_i(\beta_j(W_1) + (1-\beta_j)(w_3)) \\
&+ (1-\alpha_i)(\beta_j(w_2) + (1-\beta_j)(w_4))]
\end{aligned} \tag{31}
$$

Eq 5.2

Let $\lambda = 0$:

$$
\begin{aligned}
u_i^{\lambda=0}(\alpha_i, \beta_j) &= [\alpha_i(\beta_j(W_1) + (1-\beta_j)(w_3)) + (1-\alpha_i)(\beta_j(w_2) + (1-\beta_j)(w_4))] \\
&= \alpha_i\beta_j W_1 - \alpha_i\beta_j W_3 - \alpha_i\beta_j W_2 + \alpha_i\beta_j W_4 + \alpha_i W_3 - \alpha_i W_4 + \beta_2 W_2 \\
&= \beta_2 W_4 + W_4
\end{aligned} \tag{32}
$$

B.7. Calculating Stability Sets: A numerical example

Using Eq. 5.2, we can now compute the (single) equilbria for both games (Γ^A and Γ^D), and for different initial belief values.

First, we must choose values for all the parameters in the game, consistent with the assumptions we made (i.e. $W_1 > W_4 > W_3, W_2 > W_4 > W_3,$ and $W_1 > W_2$), so as to identify the numerical equilibria for each game when $\lambda = 1 \Rightarrow$ Let's normalize the game around zero, and let $g - b = 0$

$\Rightarrow$ Since $W_4 > W_3$ and $W_4 \equiv g - b = 0$, $W_3 \equiv s - r < 0$. Let $s = 1$, $r = 2$ to fulfil this

$\Rightarrow$ Given those choices, $W_2 \equiv g - b + s = 1$, and $W_2 > W_4 > W_3$ holds

$\Rightarrow$ Further, let $t = 4$, such that $W_1 \equiv t - r = 2$, and $W_1 > W_4 > W_3$ as well as $W_1 > W_2$ is met

$\Rightarrow$ Finally, let us define monitoring's additional payoffs (p, c) such that $W_3 < W_4$ is maintained. For ease of exposition, let $p = c = 0.25$ and to meet this conditions.

Given these parameter values, Γ^A has the following payoffs: $W_1 = 2, W_2 = 1, W_3 = -1, W_4 =$

0

Meanwhile, Γ^D provides the following payoffs: $W_1 = 2.5, W_2 = 1.25, W_3 = -0.5, W_4 = 0$

The MSNE for Γ^A is at $\alpha_i = \frac{1}{2}$, and at $\alpha_i = \frac{2}{7}$ for Γ^D

For illustrative purposes, we choose initial beliefs that are either (a) both smaller than these numeric equilibria, (b) larger than both, and (c) (separately) falling between these equilibrium values. Starting with Γ^A, we get:

(a) let $\beta_1 = \beta_2 = \frac{1}{4}$:

Eq 5.2 simplifies to:

$$\begin{aligned} u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{1}{4}) &= \frac{2}{4}\alpha_i + \frac{1}{4}\alpha_i - \frac{1}{4}\alpha_i - \alpha_i + \frac{1}{4} \\ &= \frac{1-2\alpha_i}{4} \end{aligned} \tag{33}$$

Which is maximized when $\alpha_i = 0$. Symmetrical initial beliefs of $\beta = \frac{1}{4}$ thus correspond with an optimal strategy of always defecting, (D,D) is the only equilbrium. (b) let $\beta_1 = \beta_2 = \frac{2}{3}$:

Eq 5.2 simplifies to:

$$\begin{aligned} u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{2}{3}) &= \frac{4}{3}\alpha_i + \frac{2}{3}\alpha_i - \frac{2}{3}\alpha_i - \alpha_i + \frac{1}{4} \\ &= \frac{2+\alpha_i}{3} \end{aligned} \tag{34}$$

Which is maximized when $\alpha_i = 1$.

Now, given these (high) initial beliefs, (C,C) becomes the optimal strategy.

(c) let $\beta_1 = \beta_2 = \frac{1}{3}$:

Again, Eq 5.2 simplifies to:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{1}{3}) = \frac{2}{3}\alpha_i + \frac{1}{3}\alpha_i - \frac{1}{3}\alpha_i - \alpha_i + \frac{1}{3} \\ = \frac{1-\alpha_i}{3} \tag{35}$$

Which is maximized when $\alpha_i = 0$.

For Γ^A, medium beliefs ($\beta = \frac{1}{3}$ correspond with an uncooperative equilbirium (D,D) with $\lambda = 0$

Turning to Γ^D, we employ the same three sets of initial beliefs:

(a) let $\beta_1 = \beta_2 = \frac{1}{4}$:

Eq 5.2 simplifies to:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{1}{4}) = \frac{2.5}{4}\alpha_i + \frac{0.5}{4}\alpha_i - \frac{1.25}{4}\alpha_i - 0.5\alpha_i + \frac{1.25}{4} \\ = \frac{1.25 - 0.25\alpha_i}{4} = \frac{5-\alpha_i}{16} \tag{36}$$

Which is maximized when $\alpha_i = 0$.

↳ Given low initial beliefs ($\beta = \frac{1}{3}$, when $\lambda = 0$, games Γ^A and Γ^D both only yield an uncooperative equilbrium (D,D).

(b) let $\beta_1 = \beta_2 = \frac{2}{3}$:

Eq 5.2 simplifies to:

$$u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{2}{3}) = \frac{5}{3}\alpha_i + \frac{1}{3}\alpha_i - \frac{2.5}{3}\alpha_i - 0.5\alpha_i + \frac{2.5}{3} \\ = \frac{3\alpha_i + 2.5}{3} \\ = \frac{6\alpha_i + 5}{3} \tag{37}$$

Which is maximized when $\alpha_i = 1$.

↳ When initial beliefs are sufficiently high ($\beta = \frac{2}{3}$), both games Γ^A and Γ^D yield cooperative equilbria (C,C) under $\lambda = 0$.

(c) let $\beta_1 = \beta_2 = \frac{1}{3}$:

Eq 5.2 simplifies to:

$$\begin{aligned} u_i^{\lambda=0}(\alpha_i, \beta_j = \frac{1}{3}) &= \frac{2.5}{3}\alpha_i + \frac{0.5}{3}\alpha_i - \frac{1.25}{3}\alpha_i - 0.5\alpha_i + \frac{1.25}{3} \\ &= \frac{0.25\alpha_i + 1.25}{3} \\ &= \frac{\alpha_i + 5}{12} \end{aligned} \tag{38}$$

Which is maximized when $\alpha_i = 1$.

↳ When initial beliefs are intermediate ($\beta = \frac{1}{3}$),

Γ^A yields an uncooperative outcome under $\lambda = 0$ (D,D), but

Γ^D yields a cooperative one (C,C) as the unique equilbrium

Having calculated $u_i^{\lambda=0}(\alpha_i, \beta_j)$ for all relevant sets of intial beliefs, and for both games $(\Gamma^A, \Gamma^D)_i$ we can now return to Eq. 5.1 to calculate the tracing path for all values of λ for each game and sets of beliefs

Recall Eq 5.1:

$$\begin{aligned} u_i^{\lambda}(\alpha_i, \alpha_j, \beta_i, \beta_j) = \lambda[\alpha_i(\alpha_j(W_1) + (1-\alpha_j)(w_3)) + (1-\alpha_i)(\alpha_j(w_2) + (1-\alpha_j)(w_4))] \\ + (1-\lambda)[\alpha_i(\beta_j(W_1) + (1-\beta_j)(w_3)) + (1-\alpha_i)(\beta_j(w_2) \\ + (1-\beta_j)(w_4))] \end{aligned} \tag{39}$$

$$\begin{aligned} &= \lambda[\alpha_i\alpha_j W_1 - \alpha_i\alpha_j w_3 + \alpha_i W_3 - \alpha_i\alpha_j w_2 + \alpha_i\alpha_j w_4 \\ &- \alpha_i W_4 + \alpha_j W_2 - \alpha_j W_4 + W_4] + (1-\lambda)[\alpha_i\beta_j W_1 - \alpha_i\beta_j W_3 \\ &+ \alpha_i W_3 - \alpha_i\beta_j W_2 + \alpha_i\beta_j W_4 - \alpha_i W_4 + \beta_j W_2 - \beta_j W_4 + W_4 \end{aligned} \tag{40}$$

Eq 10.1:

$$
\begin{aligned}
&= \lambda\alpha_i\alpha_j(W_1) - \lambda\alpha_i(W_3) + \lambda\alpha_i\alpha_j(W_3) + \lambda\alpha_j(W_2) + \lambda(W_4) \\
&- \lambda\alpha_j(W_4) - \lambda\alpha_i\alpha_j(W_2) - \lambda\alpha_i(W_4) + \lambda\alpha_i\alpha_j(W_4) + \alpha_i\beta_j(W_1) \\
&+ \alpha_i(W_3) - \alpha_i\beta_j(W_3) + \beta_j(W_2) + W_4 - \beta_j(W_4) - \alpha_i\beta_j(W_2) \\
&- \alpha_i(W_4) + \alpha_i\beta_j(W_4) - \lambda\alpha_i\beta_j(W_1) - \lambda\alpha_i(W_3) + \lambda\alpha_i\beta_j(W_3) \\
&- \lambda\beta_j(W_2) - \lambda(W_4) + \lambda\beta_j(W_4) + \lambda\alpha_i\beta_j(W_2) + \lambda\alpha_i(W_4) \\
&- \lambda\alpha_i\beta_j(W_4)
\end{aligned} \tag{41}
$$

Next, we can plug-in our initial priors, and then calculate the slope for the tracing path for games Γ^A and Γ^D

(a) let $\beta_j = \frac{1}{4}$:

$$
\begin{aligned}
u_i^\lambda(...) &= \lambda\alpha_i\alpha_j(W_1) - \lambda\alpha_i\alpha_j(W_3) + \lambda\alpha_j(W_2) \\
&- \lambda\alpha_i\alpha_j(W_2) - \lambda\alpha_j(W_4) + \lambda\alpha_i\alpha_j(W_4) \\
&+ \frac{\alpha_i(W_1)}{4} + \frac{4\alpha_i(W_3)}{4} - \frac{\alpha_i(W_3)}{4} + \frac{W_2}{4} \\
&- \frac{\alpha_i(W_2)}{4} + W_4 - \alpha_i(W_4) - \frac{W_4}{4} + \frac{\alpha_i(W_4)}{4} \\
&- \frac{\lambda\alpha_i(W_1}{4} + \frac{\lambda\alpha_i(W_3)}{4} - \frac{\lambda(W_2)}{4} + \frac{\lambda\alpha_i(W_2)}{4} \\
&+ \frac{\lambda(W_4)}{4} - \frac{\lambda\alpha_i(W_4)}{4}
\end{aligned} \tag{42}
$$

$$
\begin{aligned}
&= \alpha_i \Bigg(\lambda\alpha_j(W_1) + \frac{W_1}{4} - \frac{\lambda(W_1)}{4} - \lambda\alpha_j(W_2) \\
&- \frac{W_2}{4} + \frac{\lambda(W_2)}{4} + \frac{3(W_3)}{4} - \lambda\alpha_j(W_3) \\
&+ \frac{\lambda(W_3)}{4} + \lambda\alpha_j(W_4) - \frac{\lambda(W_4)}{4} - \frac{3(W_4)}{4} \Bigg) \\
&+ \lambda\alpha_j(W_2) - \lambda\alpha_j(W_4) + \frac{W_2}{4} + \frac{3(W_4)}{3} \\
&- \frac{\lambda(W_2)}{4} + \frac{\lambda(W_4)}{4}
\end{aligned} \tag{43}
$$

$$
\begin{aligned}
\frac{\partial u}{\partial \alpha_i} &= \lambda\alpha_j(W_1) + \frac{W_1}{4} - \frac{\lambda(W1)}{4} - \lambda\alpha_j(W2) \\
&- \frac{W_2}{4} + \frac{\lambda(W_2}{4} - \lambda\alpha_j(W_3) + \frac{3(W_1)}{4} \\
&+ \frac{\lambda(W3)}{4} - \lambda\alpha_j(W4) - \frac{3(W_2)}{4} + \frac{\lambda(W_4}{4}
\end{aligned} \tag{44}
$$

$$
\begin{aligned}
\boxed{\text{Plug in A}} : \frac{\partial u}{\partial \alpha_i} &= 2\lambda\alpha_j + \frac{2}{4} - \frac{2\lambda}{4} - \lambda\alpha_j \\
&- \frac{1}{4} + \frac{\lambda}{4} + \lambda\alpha_j + \frac{3}{4} - \frac{\lambda}{4} \\
&= 2\lambda\alpha_j - \frac{\lambda}{2} - \frac{1}{2} \overset{!}{=} 0 \Rightarrow \alpha_j = \frac{\lambda + 1}{4\lambda}
\end{aligned} \tag{45}
$$

$$
\begin{aligned}
\boxed{\text{Plug in B}} : \frac{\partial u}{\partial \alpha_i} &= 2.5\lambda\alpha_j + \frac{2.5}{4} - \frac{2.5\lambda}{4} - 1.25\lambda\alpha_j \\
&- \frac{1.25}{4} + \frac{1.25\lambda}{4} + 0.5\lambda\alpha_j - \frac{1.5}{4} - \frac{0.5\lambda}{4} \\
&= 1.75\lambda\alpha_j - \frac{1.75\lambda}{4} - \frac{0.25}{4} \overset{!}{=} 0 \Rightarrow \alpha_j = \frac{1.75 + 0.25}{7\lambda}
\end{aligned} \tag{46}
$$

(b) let $\beta_j = \frac{2}{3}$:

$$\begin{aligned}
u^{\lambda}(...) = {} & \lambda\alpha_i\alpha_j(W_1) - \lambda\alpha_i\alpha_j(W_3) + \lambda\alpha_j(W_2) \\
& - \lambda\alpha_i\alpha_j(W_2) - \lambda\alpha_j(W_4) + \lambda\alpha_i\alpha_j(W_4) \\
& + \frac{2\alpha_i(W_1)}{3} + \frac{3\alpha_i(W_3)}{3} - \frac{2\alpha_i(W_3)}{3} + \frac{2(W_2)}{3} \\
& - \frac{2\alpha_i(W_2)}{3} + W_4 - \alpha_i(W_4) - \frac{2(W_4)}{3} + \frac{2\alpha_i(W_4)}{3} \\
& - \frac{2\lambda\alpha_i(W_1}{3} + \frac{2\lambda\alpha_i(W_3)}{3} - \frac{2\lambda(W_2)}{3} + \frac{2\lambda\alpha_i(W_2)}{3} \\
& + \frac{2\lambda(W_4)}{3} - \frac{2\lambda\alpha_i(W_4)}{3}
\end{aligned} \tag{47}$$

$$\begin{aligned}
& = \alpha_i\Bigg(\lambda\alpha_j(W_1) + \frac{2(W_1)}{3} - \frac{2\lambda(W_1)}{3} - \lambda\alpha_j(W_2) \\
& - \frac{2(W_2)}{3} + \frac{2\lambda(W_2)}{3} + \frac{W_3}{3} - \lambda\alpha_j(W_3) \\
& + \frac{2\lambda(W_3)}{3} + \lambda\alpha_j(W_4) - \frac{W_4}{3} - \frac{2\lambda(W_4)}{3}\Bigg) \\
& + \lambda\alpha_j(W_2) + \frac{2(W_2)}{3} - \frac{2\lambda(W_2}{3} + \frac{W_4}{3} - \lambda\alpha_j(W_4) + \frac{2\lambda(W_4)}{3}
\end{aligned} \tag{48}$$

$$\begin{aligned}
\Rightarrow \frac{\partial u}{\partial\alpha_i} = {} & \lambda\alpha_j(W_1) + \frac{2(W_1)}{3} - \frac{2\lambda(W1)}{3} - \lambda\alpha_j(W2) \\
& - \frac{2(W_2)}{3} + \frac{2\lambda(W_2}{3} - \lambda\alpha_j(W_3) + \frac{W_3}{3} \\
& + \frac{2\lambda(W3)}{3} + \lambda\alpha_j(W4) - \frac{W_4}{3} + \frac{2\lambda(W_4}{3}
\end{aligned} \tag{49}$$

$$\begin{aligned}
\boxed{\text{Plug in A}} : \frac{\partial u}{\partial\alpha_i} & = 2\lambda\alpha_j + \lambda\alpha_j + \lambda\alpha_j + \frac{4}{3} - \frac{2}{3} - \frac{1}{3} - \frac{4\lambda}{3} \\
& + \frac{2\lambda}{3} - \frac{2\lambda}{3} \\
& = 2\lambda\alpha_j - \frac{4\lambda}{3} + \frac{1}{3} \stackrel{!}{=} 0 \Rightarrow \alpha_j = \frac{4\lambda - 1}{6\lambda}
\end{aligned} \tag{50}$$

$$\boxed{\text{Plug in B}} : \frac{}{\partial\alpha} = 2.5\lambda\alpha_j - 1.25\lambda\alpha_j + 0.5\lambda\alpha_j + \frac{5}{3} - \frac{2.5}{3} - \frac{0.5}{3} \\ - \frac{5\lambda}{3} + \frac{2.5\lambda}{3} - \frac{\lambda}{3} \\ = 1.75\lambda\alpha_j - \frac{3.5\lambda}{3} - \frac{2}{3} \overset{!}{=} 0 \Rightarrow \alpha_j = \frac{3.5\lambda - 2}{5.25\lambda} \tag{51}$$

ⓒ let $\beta_j = \frac{1}{3}$:

$$u^\lambda(...) = \lambda\alpha_i\alpha_j(W_1) - \lambda\alpha_i\alpha_j(W_3) + \lambda\alpha_j(W_2) \\ - \lambda\alpha_i\alpha_j(W_2) - \lambda\alpha_j(W_4) + \lambda\alpha_i\alpha_j(W_4) \\ + \frac{\alpha_i(W_1)}{3} + \frac{3\alpha_i(W_3)}{3} - \frac{\alpha_i(W_3)}{3} + \frac{W_2}{3} \\ - \frac{\alpha_i(W_2)}{3} + W_4 - \alpha_i(W_4) - \frac{(W_4)}{3} + \frac{\alpha_i(W_4)}{3} \\ - \frac{\lambda\alpha_i(W_1}{3} + \frac{\lambda\alpha_i(W_3)}{3} - \frac{\lambda(W_2)}{3} + \frac{\lambda\alpha_i(W_2)}{3} \\ + \frac{\lambda(W_4)}{3} - \frac{\lambda\alpha_i(W_4)}{3} \tag{52}$$

$$= \alpha_i\Bigg(\lambda\alpha_j(W_1) + \frac{(W_1)}{3} - \frac{\lambda(W_1)}{3} - \lambda\alpha_j(W_2) \\ - \frac{(W_2)}{3} + \frac{\lambda(W_2)}{3} + \frac{2W_3}{3} - \lambda\alpha_j(W_3) \\ + \frac{\lambda(W_3)}{3} + \lambda\alpha_j(W_4) - \frac{2(W_4)}{3} - \frac{\lambda(W_4)}{3}\Bigg) \\ + \lambda\alpha_j(W_2) - \lambda\alpha_j(W_4) + \frac{(W_2)}{3} + \frac{2}{4}W_4 - \frac{\lambda(W_2)}{3} + \frac{\lambda(W_4)}{3} \tag{53}$$

$$\Rightarrow \frac{\partial u^\lambda}{\partial\alpha} = \lambda\alpha_j(W_1) + \frac{(W_1)}{3} - \frac{\lambda(W1)}{3} - \lambda\alpha_j(W2) \\ - \frac{(W_2)}{3} + \frac{\lambda(W_2}{3} + \frac{2(W_3}{3} - \lambda\alpha_j(W_3) + \frac{\lambda(W_3)}{3} \\ + \lambda\alpha_j(W4) - \frac{2(W_4)}{3} - \frac{\lambda(W_4)}{3} \tag{54}$$

$$\boxed{\text{Plug in A}} : \frac{\partial u}{\partial \alpha} = 2\lambda\alpha_j - \lambda\alpha_j + \lambda\alpha_j - \frac{2\lambda}{3} + \frac{\lambda}{3} - \frac{\lambda}{3} + \frac{2}{3} - \frac{1}{3} - \frac{2}{3}$$
$$= 2\lambda\alpha_j - \frac{2\lambda}{3} - \frac{1}{3} \stackrel{!}{=} 0 \Rightarrow \alpha_j = \frac{2\lambda + 1}{6\lambda} \quad (55)$$

$$\boxed{\text{Plug in B}} : \frac{\partial u}{\partial \alpha} = 2.5\lambda\alpha_j - 1.25\lambda\alpha_j + 0.5\lambda\alpha_j - \frac{2.5\lambda}{3} + \frac{1.25\lambda}{3} - \frac{0.5\lambda}{3}$$
$$+ \frac{2.5\lambda}{3} - \frac{1.25\lambda}{3} - \frac{1}{3}$$
$$= 1.75\lambda\alpha_j - \frac{1.75\lambda}{3} + \frac{0.25}{3} \stackrel{!}{=} 0 \Rightarrow \alpha_j = \frac{1.75\lambda - 0.25}{5.25\lambda} \quad (56)$$

C. Comparative Statics

Recall from Appendix B.6 that the augmented game's (Γ^D) mixed strategy Nash Equilibrium is given by:

$$\alpha_i^* = \frac{g-b+r-s-c-p}{t-2s-c}$$

Further recall that by virtue of the consistently applied assumption of the assurance game (or Stag Hunt) for games Γ^A through Γ^D form:

$$t-r > g-b+s > g-b > s-r$$

and

$$t-r+c+p > g-b+s+c > g-b > s-r+p+c$$

and all of the above parameters are assumed non-negative.

C.1. Bribe Amount, *b*

$$\frac{\delta\alpha_i^*}{\delta b} = -\frac{1}{t-2s-c} < 0$$

The partial derivative of α_i^* is negative, as $t > 2s + c$, which is necessary for the existence of the MSNE (see Appendix B.6). Substantively, this means that an increase in the demanded bribe amount coincides with a reduction in the mixing probability of the MSNE.

Within the logic we've posited here, this would mean a reduced willingness to pay said bribe, which is to be expected. A note of caution, however. Throughout the analysis, we have centered our analysis around $w_4 = g - b = 0$, such that a partial derivative with respect to b is a bit illogical. Additionally, if this assumption was given up, and we allowed b to vary freely, intuition would have it that any substantial increase in b would potentially lower w_4 beyond $s - r$, which would result in the abandonment of the assurance game format.

C.2. Value of the Good/Service, g

$$\frac{\delta\alpha_i^*}{\delta g} = \frac{1}{t - 2s - c} < 0$$

Following the opposite logic of the bribe amount, if the value of the good increases, citizens become less likely to forgo the good, and thus less likely to cooperate. Here too though, we should mention that throughout we have centered the game around $g - b = 0$, such that we're computing the partial derivative of an assumed constant.

C.3. Collective Benefits from Monitoring, c

$$\frac{\delta\alpha_i^*}{\delta c} = \frac{r - s - c - p}{(t - 2s - c)^2} - \frac{1}{t - 2s - c}$$

$$= \frac{r - s - c - p}{(t - 2s - c)^2} - \frac{t - 2s - c}{(t - 2s - c)^2}$$

$$= \frac{r - s - c - p - t + 2s + c}{(t - 2s - c)^2}$$

$$= \frac{r - t + s - p}{(t - 2s - c)^2} < 0$$

Which is negative, since $t - r + p > s$, as shown in Appendix B.5. Hence, an increase in collective benefits emanating from monitoring coincides with a shrinking MSNE, and thus greater likelihood of cooperation for any probability distribution of prior beliefs.

C.4. Private Benefits from Monitoring, p

$$\frac{\delta \alpha_i^*}{\delta p} = -\frac{1}{t - 2s - c} < 0$$

The partial derivative of α_i^* is negative, as $t > 2s + c$, which necessary for the existence of the MSNE (see Appendix B.6). Hence, an increase in the private benefits emanating from monitoring coincides with a shrinking MSNE, and thus greater likelihood of cooperation for any probability distribution of prior beliefs.

C.5. Bureaucrat Retaliation, r

$$\frac{\delta \alpha_i^*}{\delta r} = \frac{1}{t - 2s - c} > 0$$

Again, the partial derivative of α_i^* is positive, as $t > 2s + c$, which necessary for the existence of the MSNE (see Appendix B.6). The MSNE shifts towards the one, and given a probability distribution of prior beliefs, this means cooperation becomes less likely.

Naturally, we assume that $t - r > g - b + s$, such that the cost from potential retaliation, r, cannot exceed the difference between joint cooperation and singular cooperation, $t - s$. If r did increase beyond that threshold, the game cedes to be an assurance game, and takes on a prisoner's dilemma-type of structure (see Table 9 in Appendix D).

C.6. Benefits from Joint Cooperation, t

$$\frac{\delta \alpha_i^*}{\delta t} = -\frac{r - s - c - p}{(t - 2s - c)^2} < 0$$

Which is negative, because $r > s + p + c$ by assumption, as shown in Appendix B.6.

Consequently, increases in these benefits shift the MSNE towards zero, and under some probability distribution of initial beliefs, this should coincide with more cooperation.

C.7. Benefits from Individual Cooperation, *s*

$$\begin{aligned}\frac{\delta\alpha_i^*}{\delta s} &= \frac{2(r-s-c-p)}{(t-2s-c)^2} - \frac{1}{t-2s-c} \\ &= \frac{2(r-s-c-p)}{(t-2s-c)^2} - \frac{t-2s-c}{(t-2s-c)^2} \\ &= \frac{2r-2s-2c-2p-t+2s+c}{(t-2s-c)^2} \\ &= \frac{2r-c-2p-t}{(t-2s-c)^2}\end{aligned}$$

Which can be either positive (if $t > 2r - c - 2p$), or negative (if not). In other words, if benefits from joint cooperation (t) are sufficiently large, increases to s shift the MSNE towards one, thus rendering cooperation less likely (under a given probability distribution of initial beliefs). This might seem counterintuitive at first, but recall that these benefits accrue to free riding actors as much as they do to individual cooperators.

Meanwhile, since we assume $r > s + p + c$, as well as $t > s + r + c$ to maintain the assurance game, as shown in Appendix B.6, s can only increase to a limited degree, *ceteris paribus*, before game no longer resembles an assurance game.

D. Basic Collective Action Games

Table 9: Types of Collective Action Games

Prisoner's Dilemma

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	2, 2	0, 3
	Defect/ Do Nothing	3, 0	1, 1

Assurance Game/Stag Hunt

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	3, 3	0, 2
	Defect/ Do Nothing	2, 0	1, 1

Deadlock

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	0, 0	1, 2
	Defect/ Do Nothing	2, 1	3, 3

Harmony

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	3, 3	2, 1
	Defect/ Do Nothing	1, 2	0, 0

Chicken

		Citizen 2: *Cooperate/ Take Action*	Citizen 2: *Defect/ Do Nothing*
Citizen 1	*Cooperate/ Take Action*	2, 2	1, 3
	Defect/ Do Nothing	3, 1	0, 0

Note: The numeric payoffs denote the preference orderings in each game, which are not perfectly comparable across games. For more, see Dixit, Skeath and Reiley (2014) and Humphreys (2017).

E. Post-Experiment Script

1. What is your age in years?
    - 18 to 25
    - 26 to 34
    - 35 to 44
    - 45 to 54
    - 55 or older
2. What is your gender?
    - Male
    - Female
3. What is your income?
    - Less than Rs. 20,000
    - Less than Rs. 20,000 - 50,000
    - Less than Rs. 50,000 - 100,000
    - Greater than Rs. 100,000
4. What was the highest class you completed?
    - Less than Primary
    - Primary
    - Middle
    - High
    - Intermediate

- Bachelors
- Graduate

5. What language did you grow up speaking?
   - { insert answers here }
6. How important is it for you to live in a country that is governed by representatives elected by the people through democracy?
   - Not at all important
   - Not important
   - Important
   - Fairly important
   - Very important
7. Where do you fall on the political spectrum?
   - PTI
   - PPP
   - PMLN
   - MQM
   - Jamat-ul-Islami
   - Other:
   - Don't know/prefer not to respond
8. Do you use the internet?
   - Yes
   - No

9. How would you characterize your health?
    - Very Good
    - Good
    - Fair
    - Poor
    - Very poor
    - Prefer not to say
10. What is your position at the factory?
    - { insert answers here }
11. How much confidence do you have in the courts to provide due process and a fair trial?
    - We trust them a lot
    - We trust them
    - We trust them a little bit
    - We don't trust them
    - We don't trust them at all
12. Have you ever been asked for a bribe? If so, can you describe the scenario?
    - Yes
    - No
13. If, hypothetically, someone asked you for a bribe today, and you choose to refuse to pay such a bribe, how else would you spend the money you saved? Please elaborate.
14. Could we kindly ask you not to speak with the other people in the factory about the game until we leave? We ask because we everyone's want your opinion, there is no

correct answers, and your response to the question could affect others as well.

- I acknowledge.

F. Changes from the Previous Pre-Analysis Plan

In the pre-analysis plan submitted to EGAP in January 2019, we had initially planned this project as a lab-in-the-field experiment, involving two-person games. When we piloted the project on 60 university students in the United States (i.e. 30 groups of 2 students; sample size 30), the project went according to plan, and we learned things along the way (e.g. eliminating the police station and judge scenarios). However, when we piloted the project on 66 Pakistani factory workers (34 groups of 2 factory workers; sample size 34) in January 2019, participants had trouble fully understanding the game that we proposed. It was too complicated given levels of education and knowledge. Additionally, after discovering Medina's (2007) lay person's theory of collective action, we recognized that we needed to change our theory accordingly and incorporate stability sets. After doing so, we realized that a conjoint experiment—with its estimation of Average Marginal Component Effects (AMCEs), not Average Treatment Effects (ATEs)—better corresponded with our theoretical logic. Those are the primary reasons accounting for our design changes. As the above sample sizes show, we did not conduct the pilot with any sort of a representative sample size to detect an effect. Now that our design is finalized, we conducted power analysis (see Section 5.9), and do have enough power to detect an effect when we run our conjoint experiment.

References

Acemoglu, Daron, Ali Cheema, Asim Ijaz Khwaja and James A. Robinson. 2020. "Trust in State and Non-State Actors: Evidence from Dispute Resolution in Pakistan." *Journal of Political Economy* .

Acemoglu, Daron and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown Business.

Aidt, Toke S. 2003. "Economic Analysis of Corruption: A Survey." *Economic Journal* 113(491):F632–F652.

Al Athmay, Alaa Aldin Abdul Rahim A. 2015. "Demographic Factors as Determinants of E-Governance Adoption." *Transforming Government: People, Process and Policy* 9(2):159–180.

Andvig, Jens Chr and Karl Moene. 1990. "How Corruption May Corrupt." *Journal of Economic Behavior and Organization* 13(1):63–76.

Arroyo Chacón, Jennifer Isabel. 2015. "Primeros frutos en la lucha contra la corrupción en Costa Rica: condenatoria en sede penal por delitos tipificados por la Ley contra la corrupción." *Derecho Penal y Criminología* 37(101):51–86.

Auerbach, Adam Michael and Tariq Thachil. 2018. "How Clients Select Brokers: Competition and Choice in India's Slums." *American Political Science Review* 112(4):775–791.

Baker, Bruce. 2009. "Cape Verde: Marketing Good Governance." *Africa Spectrum* 44(2):135–147.

Banerjee, Abhijit V. and Esther Duflo. 2007. "The Economic Lives of the Poor." *Journal of Economic Perspectives* 21(1):141–168.

Banerjee, Abhijit V. and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: PublicAffairs.

Banerjee, Ritwik. 2016. "Corruption, Norm Violation and Decay in Social Capital." *Journal of Public Economics* 137:14–27.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2018. "The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments." *Political Analysis* 26(1):112–119.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2020. Conjoint Survey Experiments. In *Cambridge Handbook of Advances in Experimental Political Science*, ed. James N. Druckman and Donald P. Green. New York: Cambridge University Press.

Bauhr, Monika and Nicholas Charron. 2018. "Insider or Outsider? Grand Corruption and Electoral Accountability." *Comparative Political Studies* 51(4):415–446.

Bertrand, Marianne, Simeon Djankov, Rema Hanna and Sendhil Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122(4):1640–1675.

Camp, Edwin, Avinash Dixit and Susan C. Stokes. 2014. "Catalyst or Cause? Legislation and the Demise of Machine Politics in Britain and the United States." *Legislative Studies Quarterly* 39(4):559–592.

Collier, Paul. 2000. "How to Reduce Corruption." *African Development Review* 12(2):191–205.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Haakon Gjerløw, Adam N. Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Moa Olin, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig and Daniel Ziblatt. 2018. Varieties of Democracy

Codebook (V8). Technical report University of Gothenburg, V-Dem Institute Gothenburg, Sweden: .

de la Cuesta, Brandon, Helen V. Milner, Daniel L. Nielson and Stephen Knack. 2019. "Oil and Aid Revenue Produce Equal Demands for Accountability as Taxes in Ghana and Uganda." *Proceedings of the National Academy of Sciences* 116(36):17717–17722.

Dixit, Avinash. 2016. Corruption: Supply-Side and Demand-Side Solutions. In *Development in India: Micro and Macro Perspectives*, ed. S. Mahendra Dev and P. G. Babu. New Dehli, India: Springer chapter 4, pp. 57–69.

Dixit, Avinash. 2018. Anti-Corruption Institutions: Some History and Theory. In *Institutions, Governance, and the Control of Corruption*, ed. Kaushik Basu and Tito Cordella. Cham, Switzerland: Palgrave Macmillan chapter 2, pp. 15–50.

Dixit, Avinash, Susan Skeath and David Reiley. 2014. *Games of Strategy*. New York: W.W. Norton and Company.

Dong, Bin, Uwe Dulleck and Benno Torgler. 2012. "Conditional Corruption." *Journal of Economic Psychology* 33(3):609–627.

Duflo, Esther, Rema Hanna and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102(4):1241–1278.

Ferguson, William. 2013. *Collective Action and Exchange: A Game-Theoretic Approach to Contemporary Political Economy*. Stanford, California: Stanford University Press.

Fisman, Raymond and Miriam A. Golden. 2017. *Corruption: What Everyone Needs to Know*. Oxford: Oxford University Press.

Fox, Jonathan A. 2015. "Social Accountability: What Does the Evidence Really Say?" *World Development* 72:346–361.

Fukuyama, Francis. 2018. Corruption as a Political Phenomenon. In *Institutions, Governance, and the Control of Corruption*, ed. Kaushik Basu and Tito Cordella. Cham, Switzerland: Palgrave Macmillan chapter 3, pp. 51–74.

Gächter, Simon and Jonathan F. Schulz. 2016. "Intrinsic Honesty and the Prevalence of Rule Violations across Societies." *Nature* 531(7595):496–499.

Gino, Francesca, Shahar Ayal and Dan Ariely. 2009. "Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel." *Psychological Science* 20(3):393–398.

Glaeser, Edward L. and Claudia Goldin, eds. 2007. *Corruption and Reform: Lessons from America's Economic History*. Chicago: University of Chicago Press.

Gneezy, Uri, Silvia Saccardo and Roel van Veldhuizen. 2019. "Bribery: Behavioral Drivers of Distorted Decisions." *Journal of the European Economic Association* 17(3):917–946.

Grindle, Merilee S. 2012. *Jobs for the Boys: Patronage and the State in Comparative Perspective*. Cambridge, Massachusetts: Harvard University Press.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.

Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112(8):2395–2400.

Harsanyi, John C. 1987. "The Tracing Procedure: A Self-Correcting Reasoning Procedure." *Theory and Decision* 23:7–23.

Hawkins, Darren, David A. Lake, Daniel L. Nielson and Michael J. Tierney. 2006. Delegation under Anarchy: States, International Organizations, and Principal-Agent Theory. In

Delegation and Agency in International Organizations. New York: Cambridge University Press chapter 1, pp. 3–38.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Hough, Dan. 2013. *Corruption, Anti-Corruption, and Governance*. New York: Palgrave Macmillan.

Humphreys, Macartan. 2017. *Political Games: Mathematical Insights on Fighting, Voting, Lying and Other Affairs of State*. New York: W.W. Norton and Company.

Justesen, Mogens K. and Christian Bjørnskov. 2014. "Exploiting the Poor: Bureaucratic Corruption and Poverty in Africa." *World Development* 58:106–115.

Khan, Feisel. 2018. Pakistan: A Study in Corruption. In *Handbook on the Geographies of Corruption*, ed. Barney Warf. Cheltenham, United Kingdom: Edward Elgar Publishing chapter 17, pp. 274–299.

Kingston, Christopher. 2008. "Social Structure and Cultures of Corruption." *Journal of Economic Behavior and Organization* 67(1):90–102.

Klitgaard, Robert. 1988. *Controlling Corruption*. Berkeley, California: University of California Press.

Kubinec, Robert. 2019. "Politically-Connected Firms and the Military-Clientelist Complex in North Africa.".

Kuran, Timur. 1989. "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice* 61(1):41–74.

Kuran, Timur. 1991. "The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(01):7–48.

Lee, Wang-Sheng and Cahit Guven. 2013. "Engaging in Corruption: The Influence of Cultural Values and Contagion Effects at the Microlevel." *Journal of Economic Psychology* 39:287–300.

Levitt, Steven and John A. List. 2008. "Homo Economicus Evolves." *Science* 319(5865):909–910.

Levy, Brian. 2014. *Working with the Grain: Integrating Governance and Growth in Development Strategies*. New York: Oxford University Press.

Lizzeri, Alessandro and Nicola Persico. 2004. "Why Did the Elites Extend the Suffrage? Democracy and the Scope of Government, with an Application to Britain' s Age of Reform." *Quarterly Journal of Economics* 119(2):707–765.

Marquette, Heather. 2007. "The World Bank's Fight against Corruption." *Brown Journal of World Affairs* 13(2):27–39.

Marquette, Heather and Caryn Peiffer. 2018. "Grappling with the Real Politics of Systemic Corruption: Theoretical Debates Versus Real-World Functions." *Governance* 31(3):499–514.

Marquette, Heather and Caryn Peiffer. 2019. "Thinking Politically about Corruption as Problem-Solving: A Reply to Persson, Rothstein, and Teorell." *Governance* 32(4):811–820.

Masoud, Tarek. 2018. "Why Tunisia?" *Journal of Democracy* 29(4):166–175.

McCarthy, Colm. 2003. Corruption in Public Office in Ireland: Policy Design as a Countermeasure. In *Tribunals of Inquiry*. Dublin: pp. 1–15.

Medina, Luis Fernando. 2007. *A Unified Theory of Collective Action and Social Change*. Ann Arbor, Michigan: University of Michigan Press.

Medina, Luis Fernando. 2018. *Beyond the Turnout Paradox: The Political Economy of Electoral Participation*. Cham, Switzerland: Springer.

Mungiu-Pippidi, Alina. 2013. "Controlling Corruption through Collective Action." *Journal of Democracy* 24(1):102–115.

Mungiu-Pippidi, Alina. 2015. *The Quest for Good Governance: How Societies Develop Control of Corruption*. New York: Cambridge University Press.

Mungiu-Pippidi, Alina. 2016. "Learning from Virtuous Circles." *Journal of Democracy* 27(1):95–109.

Mungiu-Pippidi, Alina and Michael Johnston, eds. 2017. *Transitions to Good Governance: Creating Virtuous Circles of Anti-corruption*. Cheltenham, United Kingdom: Elgar.

Naím, Moises. 1995. "The Corruption Eruption." *Brown Journal of World Affairs* 2(2):245–262.

Nyblade, Benjamin and Steven R Reed. 2008. "Who Cheats? Who Loots? Political Competition and Corruption in Japan, 1947-1993." *American Journal of Political Science* 52(4):926–941.

O'Donnell, Guillermo. 1998. "Horizontal Accountability in New Democracies." *Journal of Democracy* 9(3):1112–1126.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29(3):61–80.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Massachusetts: Harvard University Press.

Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address." *American Political Science Review* 92(1):1–22.

Peiffer, Caryn and Linda Alvarez. 2016. "Who Will Be the Principled-Principals? Perceptions of Corruption and Willingness to Engage in Anticorruption Activism." *Governance* 29(3):351–369.

Peiffer, Caryn and Richard Rose. 2018. "Why Are the Poor More Vulnerable to Bribery in Africa? The Institutional Effects of Services." *Journal of Development Studies* 54(1):18–29.

Persson, Anna, Bo Rothstein and Jan Teorell. 2012. Rethinking the Nature of the Grabbing Hand. In *Good Government*, ed. So¨ren Holmberg and Bo Rothstein. Cheltenham, United Kingdom: Edward Elgar Publishing pp. 251–273.

Persson, Anna, Bo Rothstein and Jan Teorell. 2013. "Why Anticorruption Reforms Fail–Systemic Corruption as a Collective Action Problem." *Governance* 26(3):449–471.

Persson, Anna, Bo Rothstein and Jan Teorell. 2019. "Getting the Basic Nature of Systemic Corruption Right: A Reply to Marquette and Peiffer." *Governance* 32(4):799–810.

Perugini, Marco, Marcello Gallucci and Giulio Costantini. 2018. "A Practical Primer To Power Analysis for Simple Experimental Designs." *International Review of Social Psychology* 31(1):1–23.

Polus, Andrzej, Dominik Kopinski and Wojciech Tycholiz. 2015. "Ready or Not: Namibia as a Potentially Successful Oil Producer." *Africa Spectrum* 50(2):31–55.

Robinson, Amanda Lea and Brigitte Seim. 2018. "Who is Targeted in Corruption? Disentangling the Effects of Wealth and Power on Exposure to Bribery." *Quarterly Journal of Political Science* 13:313–331.

Rose-Ackerman, Susan and Bonnie Palifka. 2016. *Corruption and Government: Causes, Consequences, and Reform.* Second ed. New York: Cambridge University Press.

Rothstein, Bo. 2011*a*. "Anti-Corruption: The Indirect Big Bang Approach." *Review of International Political Economy* 18(2):228–250.

Rothstein, Bo. 2011*b*. *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective.* Chicago: University of Chicago Press.

Rothstein, Bo and Daniel Eek. 2009. "Political Corruption and Social Trust: An Experimental Approach." *Rationality and Society* 21(1):81–112.

Rothstein, Bo and Jan Teorell. 2015. "Getting to Sweden, Part II: Breaking with Corruption in the Nineteenth Century." *Scandinavian Political Studies* 38(3):238–254.

Sandler, Todd. 2015. "Collective Action: Fifty Years Later." *Public Choice* 164(3-4):195–216.

Schelling, Thomas. 1978. *Micromotives and Macrobehavior.* New York: W.W. Norton and Company.

Sen, Amartya. 1967. "Isolation, Assurance, and the Social Rate of Discount." *Quarterly Journal of Economics* 81(1):112–124.

Shalvi, Shaul. 2016. "Corruption Corrupts." *Nature* 531(7595):456–457.

Søreide, Tina. 2014. *Drivers of Corruption: A Brief Review.* Washington, DC: World Bank.

Soto, Raimundo and Ilham Haouas. 2016. Has the UAE Escaped the Oil Curse? In *Understanding and Avoiding the Oil Curse in Resource-Rich Arab Economies*, ed. Ibrahim Elbadawi and Hoda Selim. Cambridge: Cambridge University Press chapter 12, pp. 373–420.

Stokes, Susan C., Thad Dunning, Marcelo Nazareno and Valeria Brusco. 2013. *Brokers, Voters, and Clientelism: The Puzzle of Distributive Politics.* New York: Cambridge University Press.

Sundström, Aksel. 2019. "Why Do People Pay Bribes? A Survey Experiment with Resource Users." *Social Science Quarterly* 100(3):725–735.

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. "The Ties that Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112(3):525–541.

Teorell, Jan and Bo Rothstein. 2015. "Getting to Sweden, Part I: War and Malfeasance, 1720-1850." *Scandinavian Political Studies* 38(3):217–237.

Therialt, Sean M. 2003. "Patronage, the Pendleton Act, and Power of the People." *Journal of Politics* 65(1):485–486.

Transparency International. 2009. The Anti-Corruption Plain Language Guide. Technical report Transparency International Berlin: .

Transparency International. 2018. "Corruption Perceptions Index 2018.".
URL: *www.transparency.org/cpi2018*

Treisman, Daniel. 2000. "The Causes of Corruption: A Cross-National Study." *Journal of Public Economics* 76(3):399–457.

Ugur, Mehmet and Nandini Dasgupta. 2011. Evidence on the Economic Growth Impacts of Corruption in Low-Income Countries and Beyond. Technical Report August EPPI-Centre, Social Science Research Unit, Institute of Education, University of London London: .

United Nations. 2018. "Global Cost of Corruption at Least 5 Per Cent of World Gross Domestic Product, Secretary-General Tells Security Council, Citing World Economic Forum Data.".
URL: *https://www.un.org/press/en/2018/sc13493.doc.htm*

Vaubel, Roland. 2006. "Principal-Agent Problems in International Organizations." *Review of International Organizations* 1(2):125–138.

Weimann, Joachim, Jeannette Brosig-Koch, Timo Heinrich, Heike Hennig-Schmidt and Claudia Keser. 2019. "Public Good Provision by Large Groups – The Logic of Collective Action Revisited." *European Economic Review* 118:348–363.

Weyland, Kurt. 2012. "The Arab Spring: Why the Surprising Similarities with the Revolutionary Wave of 1848?" *Perspectives on Politics* 10(4):917–934.

Wilson, Bruce and Evelyn Villarreal. 2017. Costa Rica: Tipping Points and An Incomplete Journey. In *Transitions to Good Governance: Creating Virtuous Circles of Anti-corruption*, ed. Alina Mungiu-Pippidi and Michael Johnston. Cheltenham, United Kingdom: Elgar chapter 8, pp. 184–212.

World Bank. 2017. "World Development Indicators.".
URL: *http://data.worldbank.org/data-catalog/world-development-indicators*

Wrong, Michaela. 2009. *It's Our Turn to Eat: A Story of a Kenyan Whistleblower*. London: Fourth Estate.

Yap, O. Fiona. 2013. "When do Citizens Demand Punishment of Corruption?" *Australian Journal of Political Science* 48(1):57–70.