University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 260

Corruption, Norm Enforcement and Cooperation

Justin Buffat and Julien Senn

Revised version, July 2018

Corruption, Norm enforcement and Cooperation*

Justin Buffat Julien Senn

This version: July 18, 2018

Abstract

In many societies, the power to punish is granted to a centralized authority. While the punishment of free-riders has been shown to play an important role in the provision of public goods, corruption might strongly disrupt the ability of a centralized authority to foster cooperation. In this paper, we show that cooperation is reduced by 30% if the punishment authority can be bribed. Two concurrent channels lead to this result. First, low contributors use bribery as a way to tame the punishment authority. The punishment authority tends to reciprocate these bribes by assigning fewer punishment points. These low levels of punishment do not suffice to discipline the free-riders, who never raise their contributions. Second, bribery has negative spillovers on high contributors, who get discouraged and gradually decrease their contributions down to the level of low contributors. Overall, our paper highlights a potential peril of centralization: the sensitivity of the punishment authority to bribery.

JEL CODES: C91, D73, K42
Key words: corruption, bribery, cooperation, public good, institutions

*Buffat: University of Lausanne (justin.buffat@unil.ch). Senn: University of Zurich (julien.senn@econ.uzh.ch). Buffat gratefully acknowledges financial support from the Center for Economic and Social Behavior (C-SEB) of the University of Cologne.

1 Introduction

Why do humans cooperate when they have private incentives to free-ride on others? How can social norms be enforced and sustained? A large body of research has documented that *altruistic punishment*, i.e. the punishment of free-riders by cooperators, plays an important role in the provision of public goods (see e.g., Fehr and Gächter, 2000, 2002). In such settings, the power to punish is decentralized to the individuals themselves, i.e. the duty to punish and enforce social norms is borne by all the members of a group. While peer sanctions can be used in broad range of situations in which norms are transgressed, there are also many instances in which punishment is institutionalized. For example, most countries delegate law enforcement to the judicial system and to the police force. In smaller-scale societies, villages or tribes have chiefs who ensure that rules and norms are followed by the community.

Recent experimental evidence indicates that centralized institutions are successful in sustaining cooperation in controlled settings (see e.g. Kosfeld et al., 2009; Andreoni and Gee, 2012; Nicklisch et al., 2016; Fehr and Williams, 2018) and that centralization emerges endogenously when individuals are given the choice between different institutional regimes (Fehr and Williams, 2018). Several reasons make centralization appealing. First, centralization does not suffer from second-order free-riding. Under peer punishment, all the members of a group have private incentives *not* to punish since punishment comes at a cost for the punisher (see e.g., Fehr, 2004; Panchanathan and Boyd, 2004). If too little or no punishment is assigned to the free-riders, cooperation will not be sustained in the long run. Second, there is no retaliation problem when punishment is centralized, as opposed to decentralized settings which have been shown to generate a significant amount of counter punishment (see e.g., Nikiforakis, 2008). But are there also perils to centralization?

In this paper, we investigate whether corruption can dampen the disciplining power of a centralized authority. Our intuition is the following: if the punishment rights are held by a single individual, some group members might be tempted to influence it by engaging in bribery. In particular, individuals who break the norms (e.g. the free-riders) might benefit from corrupting the punishment authority if the returns to bribery, through lower punishment, are higher than its cost. Examples of such attempts to corrupt the authority are manifold. A classic example is the traffic offender who slips a bank note in his driver's license after being caught by the police. In this situation, the briber tries to corrupt a public official whose role is to enforce the law. In other cases, influence is more subtle. For example, a pharmaceutical company might send gifts to a hospital's procurement officer. Typically, procurement officers *ought to behave* in a certain way:

they are expected to make purchases in the company's best interest. But they also have some leeway in assessing which providers makes the best offer. By buying the product of the company that sent a gift, the procurement officer does not necessarily break the law (even if the quality-price ratio of that company is low compared to competitors).

Empirical evidence suggests that bribery is widespread. In 2016 alone, it was estimated that more than 1 trillion of USD was spent on bribery worldwide (World Bank Group).[1] The extent of corruption and its potential implications for norm enforcement make it an important topic to investigate. Can a centralized institution cope with corruption and foster cooperation? Will corruption completely annihilate the authority's ability to discipline free-riders? To a large extent, these remain empirical questions.

We study these questions using a public good game with bribery. Public good games are well suited to study social dilemmas in which individual incentives are not aligned with social efficiency. In our experiment, participants are either in the role of a *monitor* or of a *citizen*. We investigate the effects of corruption using two treatments in a between-subjects design. Our key variation is whether bribing the monitor is possible or not. In the NO BRIBE treatment, citizens first decide on their contribution to the public good and the monitor subsequently allocates punishment points. In the BRIBE treatment, citizens have the possibility to send a bribe to the monitor before the monitor allocates the punishment points. These two treatments allow us to identify the causal effects of corruption on punishment and on cooperation rates.

Our experimental approach has three main advantages. First, the exogenous manipulation of the institutional environment allows us to cleanly identify the effects of corruption on cooperation. Such an exogenous variation of corruption would be virtually impossible to either observe or achieve in a more natural setting. While instrumental variables can be used to circumvent endogeneity problems and reverse causation, it is not clear that an indisputably good instrument for corruption, i.e. a variable that is correlated with corruption but not with cooperation, exists. In such circumstances, laboratory experiments provide a unique way to gather causal evidence on the effects of corruption on cooperation. Second, our experimental approach delivers unambiguous measures of cooperation (average contribution to the public good) and of corruption (bribes sent to the monitor). Because of its illicit nature, quantifying bribery in a more natural context would be very difficult. Third, and perhaps most importantly, our experimental design allows to isolate the different channels through which bribery affects cooperation. This is important because only a better understanding of how bribery affects cooperation can help the society to come up with more effective ways to tackle it.

[1] Retrieved from `http://www.worldbank.org/en/topic/governance/brief/anti-corruption` on February 22, 2017.

Several important findings emerge from our data. First, we show that overall contributions to the public good are reduced by 33% when participants have the possibility to bribe the monitor. Interestingly, the average initial contribution to the public good in BRIBE is not significantly different than in NO BRIBE. This finding suggests that institutional differences take time before significantly affecting behavior. It is only after the fourth period that the average contribution in NO BRIBE significantly exceeds the average contribution in BRIBE. Interestingly, the prevalence of corruption does not lead to a complete collapse of cooperation. Indeed, citizens contribute on average 50% of their endowment to the public good in the BRIBE condition.[2]

While documenting such a large treatment effect is interesting in itself, our paper also investigates how it originates. We provide evidence that two concurrent channels explain this finding. First, we show that bribery largely softens the punishment behavior of the monitor. On average, BRIBE reduces punishment both at the intensive and at the extensive margin (despite the fact that the average level of cooperation is lower in BRIBE). In particular, monitors in BRIBE are much more lenient with free-riders. For example, a citizen that contributes between 12 and 20 points less than the other citizens of his group (the maximum contribution is 20 points) receives about 5 times more punishment points in NO BRIBE than in BRIBE (7.1 points versus 1.5 point). This difference in the severity of punishment has important implications for cooperation: while harsh sanctions discipline low contributors in NO BRIBE, the soft punishments inflicted in BRIBE are not sufficient to enforce higher levels of cooperation in the subsequent periods.

Second, we show that the prevalence of bribery has a large discouraging effect on initially high contributors, i.e. citizens that contributed more than the average of the other members of their group in period 1. Quite astonishingly, the average contribution in period 1 of initially high contributors is not affected by the treatment. Neither is the average contribution in period 1 of initially low contributors. However, the monitors assign significantly more punishment points to initially low contributors in period 1 in NO BRIBE than in BRIBE. Over time, initially low contributors increase their contributions up to the level of initially high contributors in NO BRIBE. In contrast, initially high contributors decrease their contribution down to the level of initially low contributors in BRIBE. Ultimately, all the citizens end up contributing a relatively high proportion of their endowment (approximately 75%) in NO BRIBE, whereas all the citizen in BRIBE end up contributing a relatively low proportion of their endowment (less than 50%).

[2]Using a n-person Prisoner's Dilemma framework in which players can bribe the punishment authority, Kosfeld (1997) provides theoretical support for this finding. He shows that there exists a maximal number of corrupting and defecting agents under which the public good is formed and the other members of the group cooperate.

We speculate that the main reason for which initially high contributors reduce their contributions so dramatically in BRIBE is that they get discouraged to see that initially low contributors receive few punishment points.

Our paper relates to several strands of the literature. First, our paper is linked to the literature studying cooperation in environments in which punishment is centralized. Previous work has studied how sanctions that are automatically executed (formal sanctions) affect cooperation (Andreoni and Gee, 2012; Markussen et al., 2014; Kosfeld et al., 2009). More recently, several studies have investigated how centralized authorities that have discretionary power over punishment affect cooperation (Baldassarri and Grossman, 2011; Nicklisch et al., 2016; Fehr and Williams, 2018). Finally, Muthukrishna et al. (2017) investigate the comparative statics effects of structural factors and anti-corruption strategies. Overall, while most of these papers highlight the benefits of centralized sanctioning, our results indicate that the ability of a centralized authority to foster cooperation is largely undermined by bribery.

Second, our paper is related to the experimental literature investigating how gifts and bribes distort judgment. Gneezy et al. (2016) investigate the behavioral drivers of bribery and provide both lab and field evidence that bribes distort the judgment of a referee whose role is to assign a prize to one of two contestants. In a similar vein, Malmendier and Schmidt (2017) show that gifts strongly influence decision makers at the expenses of a third party. We contribute to this literature by showing that bribery distorts the way a centralized authority punishes free-riders in social dilemmas.

More broadly, our paper is linked to the literature studying corruption using tailor-made laboratory experiments. Pioneering experimental work on bribery, Abbink et al. (2002) show that reciprocity plays a key role in bribery relationships and discuss the role of externalities and sanctions on corruption. Their paper has paved the way for numerous experiments that have shed light on the effects of anti-corruption policies (see e.g., Abbink, 2005; Van Veldhuizen, 2013) and culture (Barr and Serra, 2010; Cameron et al., 2009) on corruption, among others.[3] While most of this literature has focused on bilateral relationships between the briber and the bribee, our paper extending the study of bribery to situations of social dilemmas.

The paper is organized as follows. In Section 2, we describe the experimental design. In Section 3 we outline our main behavioral predictions. In Section 4, we present our findings and discuss their implications. Finally, Section 5 concludes the paper.

[3]See Abbink and Serra (2012) for a recent overview of existing laboratory studies on bribery.

2 Experimental design

Our experiment consists of a modified public goods game with two treatments. The participants are matched in groups of 4 that remain constant for the entire duration of the experiment. Within a group, one participant is randomly selected to be the monitor. The monitor acts as a centralized sanctioning authority; It is the only player who has the power to assign punishment points to other participants. The other three participants are assigned the role of citizens. Their main activity is to decide how much to contribute to a public good.

The two treatments we implement differ in only one dimension: the possibility to transfer tokens to the monitor, i.e., to bribe her.[4] A period consists of 3 stages: 1) the contribution stage, 2) the bribery stage and 3) the punishment stage. In the BRIBE treatment, the citizens decide how many tokens to transfer to the monitor during stage 2. In the NO BRIBE treatment, the citizens do not have the possibility to transfer tokens to the monitor.

The experiment is repeated over 20 periods. In order to avoid reputation effects, the citizens' identity is reshuffled at every period.[5] This is a realistic assumption: While the citizens of a country generally know whether the authority is corrupted (e.g. whether the police tends to accept bribes), the authority does not necessarily know what to expect from each particular individual. Therefore, it is reasonable to focus on situations in which the monitor does not know the identity, respectively the past behavior, of its citizens.

Let π_i^t be the profit of citizen i at the end of stage $t = \{1, 2, 3\}$, and π_m^t denote the profit of the monitor. At the beginning of a period, each participant (i.e. the monitor and the citizens) receives an endowment of 20 tokens. In the contribution stage, each citizen has to (individually) decide how much of his endowment to invest in the public good. The difference between the endowment and the contribution to the public good is directly transferred to the citizen's private account. The monitor keeps its endowment and cannot contribute to the public good. Each token contributed to the public good benefits all the group members equally, i.e. the citizens and the monitor all receive the same monetary payoff out of the public good. The revenue from the public good consists of the sum of contributions multiplied by 0.4, the marginal per capita return to the public good. At the end of the contribution stage, the revenue generated by the public good and by the private account is privately disclosed to each citizen. Citizens are also made aware of their stage 1 profit $\pi_i^1 = 20 - c_i + 0.4 \sum_{j=1}^{3} c_j$, where c_i is the contribution of citizen

[4] In our experiment, we purposefully did not use of the word "bribe" in order to avoid a framing that is too negatively connoted. Nevertheless, survey data collected at the end of our experiment suggests that virtually all the participants did perceive the transfers of tokens as bribes.

[5] For example, citizen 1 in period 1 is not necessarily called citizen 1 in period 2.

i to the public good.

At the beginning of the bribery stage, each citizen receives an additional endowment of 5 tokens. In the NO BRIBE treatment, neither the citizens nor the monitor are asked to take a decision. In the BRIBE treatment, the citizens must decide how many of these additional 5 tokens to *transfer* to the monitor. Following recent experiments on bribery, we do not allow the monitor to refuse bribes (see e.g. Gneezy et al., 2016; Malmendier and Schmidt, 2017).[6]

The profits at the end of the bribery stage are

$$\pi_i^2 = \pi_i^1 + 5 - b_i \cdot \mathbb{1}(\text{BRIBE}) \quad , \text{ for } i = 1, 2, 3$$
$$\pi_m^2 = 20 + 0.4 \sum_{j=1}^{3} c_j + \sum_{j=1}^{3} b_j \cdot \mathbb{1}(\text{BRIBE})$$

where $b_i \in [0, 5]$ denotes the bribe sent to the monitor by citizen i and $\mathbb{1}(\text{BRIBE})$ is the indicator function taking the value 1 if the participant is in the BRIBE condition and zero otherwise.

The third stage corresponds to the punishment stage. The monitor starts by receiving an additional endowment of 15 tokens (out of which the punishment points can be allocated) and to learn each citizen's contribution. In BRIBE, the monitor also learns whether the citizens bribed her and by how much. The monitor is then asked to decide how many punishment points R_i to allocate to each citizen i. One punishment point reduces the profit of a citizen by 3 tokens. The monitor cannot allocate more than 10 punishment points to one citizen, i.e. she cannot reduce its profit by more than 30 tokens, and cannot allocate more than 15 punishment points per period ($\sum_i R_i \leq 15$).[7] The final profits are

$$\pi_i^3 = \pi_i^2 - 3R_i \quad , \text{ for } i = 1, 2, 3$$
$$\pi_m^3 = \pi_m^2 + 15 - \sum_{i=1}^{3} R_i.$$

At the end of a period, full feedback regarding the actions taken by all the citizens and the monitor is publicly disclosed to all the members of the group. That is, each citizen is provided with a summary of his own actions, the number of punishment

[6] Previous research has shown that people are unlikely to reject small bribes in similar setting (Malmendier and Schmidt, 2017; Gneezy et al., 2016). In Gneezy et al. (2016), most treatments do not incorporate the possibility to decline bribes. They report evidence from two additional treatments in which declining bribes is possible and show that only a very small fraction of the participants (approximately 10%) do so. Moreover, the behavior of their participants is orthogonal to the possibility to decline bribes.

[7] See Figure S1 in Appendix for a screenshot of the monitor's decision screen.

points he received and his final profit in that particular period. Each citizen is also informed about the contributions and (if applicable) the bribery decisions as well as the punishment points received by the two other citizens of his group.[8]

Three important features of our experimental design are worth being discussed. To begin with, the choice to offer complete feedback to all the citizens at the end of a period might seem – at first – unrealistic. Indeed, people do not usually observe the bribes sent by others. However, people do generally have a fairly good idea of the average degree of corruption that prevails in the society they are living in. Their perception of corruption has generally been constructed over time, through experience (see e.g., Olken, 2009). By choosing to fully disclose the actions of all the members of the group at the end of a period we shorten the social learning process, i.e. we reduce the time that would have been needed by the participants to learn what strategy is most commonly being played by the other participants.

Second, note that a bribe sent to the monitor can only make up to 25% of the monitor's endowment. Hence, the bribes that we consider are relatively small.[9]

We purposefully restrict our attention to small bribes because of the variety of corruption situations that they encompass: a traffic offender slips a bank note in his driver's license, a real estate agent sends a box of wine to the person in charge of deciding on which piece of land the construction of a new building will be allowed, etc. In all these examples, the briber sends a small gift to someone in the hope of getting a favor in return. The decision to give every citizen a "bribery endowment"and to force them to only bribe out of this additional endowment of 5 tokens is purely practical: it ensures that their bribery decision is unconstrained by their profit in stage 1, thereby improving comparability between the citizens.

Third, note that the monitor has a selfish interest in being part of a group in which citizens highly contribute to the public good. She is therefore expected to take action in order to foster high levels of cooperation in her group, i.e. punish non-contributors. This modeling choice, although highly stylized, is an accurate representation of many real life situations involving a centralized sanctioning authority. For example, the leader of a tribe – or the mayor of a small town – has an interest in promoting and sustaining high levels of cooperation because both herself and her community reap the benefits of living in a well-functioning society.

[8]See Figure S2 in Appendix for a screenshot of the end-of-period screen.

[9]Note that Gneezy et al. (2016) and Malmendier and Schmidt (2017) also focus on bribes that are "small gifts".

Instructions and framing

At the beginning of the experiment, participants received a set of instructions describing the decision stages and the payoffs at stake for the citizens and the monitor.[10] The instructions were neutrally framed. A citizen was referred to as a participant A, the monitor was referred to as participant B and a bribe was referred to as a "*transfer to participant B*". Note that, within a treatment, both citizens and monitors received the same set of instructions.

3 Behavioral predictions

The experimental design outlined above leads us to formulate the following behavioral predictions.

Prediction 1. *Cooperation in BRIBE is lower than in NO BRIBE.*

While we expect the monitor in the NO BRIBE condition to be able to sustain high levels of cooperation,[11] the possibility to bribe the monitor (BRIBE condition) is predicted to dampen cooperation. We expect two channels to simultaneously undermine cooperations rates in BRIBE: lower punishment of the free-riders (prediction 2) and a discouragement effect (prediction 3).

Prediction 2. *Bribery softens punishment, which reduces the monitor's ability to discipline free-riders and to foster cooperation.*

Following the literature on reciprocity, we expect bribed monitors to assign weaker sanctions to bribers.[12] In turn, we predict that these weaker sanctions will not discipline low contributors as much as tougher sanctions would (see e.g. Fehr and Gächter, 2000), i.e. we expect the disciplining power of punishment to be lower in BRIBE. Over time, weaker sanctions might increase the tendency of citizens to free-ride in the contribution stage. *Ceteris paribus*, if a low contributor (i.e., a free-rider) can avoid a severe punishment by bribing the monitor, then free riding on the contribution of the other members of the group and bribing the monitor might become an attractive strategy.[13]

Prediction 3. *Bribery discourages initially high contributors.*

[10] See the Appendix for a translated version of the instructions.

[11] Indeed, previous literature has documented that a centralized sanctioning authority can foster cooperation (see for example Baldassarri and Grossman, 2011).

[12] For key contributions on reciprocity, see Berg et al. (1995) and Fehr et al. (1997), among others.

[13] To illustrate how bribing might be profitable and part of an equilibrium, consider two strategies available to a citizen. He can either (i) contribute $c_i^0 \geq 0$ to the public good and transfer nothing ($b_i^0 = 0$) to the monitor or (ii) fully free-ride (i.e., $c_i = 0$) and bribe the monitor an amount $b_i > 0$. Furthermore, suppose that he expects to receive R_i^0 punishment points if he does not bribe and $R_i^b < R_i^0$ if he bribes

Finally, we expect the prevalence of bribery to discourage initially high contributors. Because citizens have perfect information about the contributions, the bribes and the punishment points received by the other members of their group, observing that free-riders can get away with low punishment might ultimately discourage them from contributing high amounts in the subsequent rounds.

4 Results

We conducted 7 experimental sessions at the University of Cologne in July 2016. A total of 224 subjects participated in our experiment which was fully computerized using z-Tree (Fischbacher, 2007). Subjects were recruited using ORSEE (Greiner, 2015) and were allowed to take part in the experiment only once. Payoffs were converted in Euros (EUR) at an exchange rate of 100 tokens = EUR 1.20. On average, a session lasted 75 minutes and participants earned EUR 13.80, including a show-up fee of EUR 4. Treatments were randomized within a session: half of the participants were assigned to the BRIBE treatment while the other half was assigned to the NO BRIBE treatment. Tables S1 and S2 in Appendix compare participants' observable characteristics across conditions. While some slight differences exist, these tables provide evidence that both monitors and citizens were well balanced across treatments. Overall, our data comprises 56 independent groups of 4 subjects (28 groups per treatment).

Result 1. *On average, contributions in BRIBE are 33% lower than in NO BRIBE.*

Figure 1 depicts the average contribution in the BRIBE and the NO BRIBE treatments. Overall, subjects in BRIBE contribute an average of 4.14 tokens less – i.e. approximately 33% less – than subjects in NO BRIBE (Wald test, $p < 0.01$, OLS regressions column 1 of Table S3 in the Appendix). This result is robust to an OLS estimation which controls for period fixed effects and individual covariates (Wald test, $p < 0.01$, see column 2 of Table S3 in the Appendix). Interestingly, the average initial contribution is approximately equal to 10, i.e. 50% of the endowment, in both treatments. This suggests that the treatments do not affect the behavior of the citizens from the outset of the experiment. While the average contribution significantly increases by 50% over time in NO BRIBE (from 9.56 tokens in period 1 to 14.3 tokens in period 19, Wald

---

the monitor. Given these beliefs and the profit function defined above, bribing is profitable if and only if

$$20 + 5 - b_i + 0.4 \sum_{j \neq i} c_j - 3R_i^b > 20 + 5 - c_i^0 + 0.4c_i^0 + 0.4 \sum_{j \neq i} c_j - 3R_i^0 \Leftrightarrow b_i < 0.6c_i^0 + 3(R_i^0 - R_i^b),$$

i.e. if the bribe is lower than the cost of contributing, $0.6c_i^0$, and the additional punishment that can be expected from not bribing $3(R_i^0 - R_i^b)$.

test, $p < 0.01$), it slightly decreases by 10% in BRIBE, although not significantly (from 10 tokens in period 1 to 9.05 tokens in period 19, Wald test, $p = 0.43$).[14,15] These dynamics strike in very early in the experiment: In period 4 already, the average contribution in NO BRIBE exceeds the average contribution in BRIBE (Wald test $p = 0.08$). This difference in contribution increases both in size and in significance over time ($p < 0.05$ between periods 6 and 10, $p < 0.01$ after period 10).

Two important findings are worth being highlighted. First, contributions continuously increase over time in NO BRIBE. This finding is in line with Baldassarri and Grossman (2011) who show that the presence of a monitor (either democratically elected or randomly chosen) can sustain high levels of cooperation and significantly increases cooperation compared to a baseline condition with no monitor. Second, cooperation does not collapse in BRIBE. This result is in line with casual observation: corrupted countries generally still provide some positive level of public good to their citizens.[16]

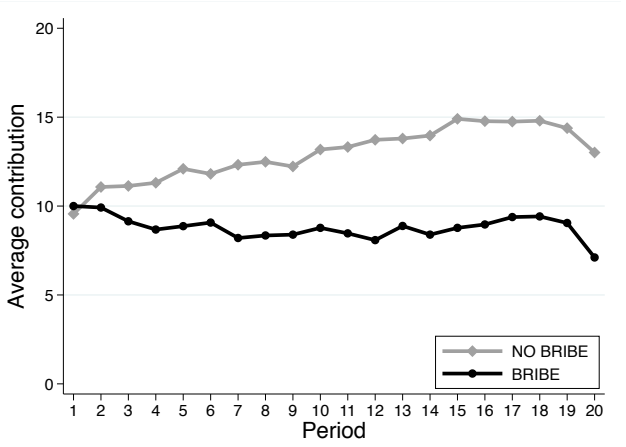


Figure 1: Average contributions over time (by treatment)

In what follows, we investigate the causes of the different dynamics of cooperation observed across the two treatments. We start by investigating whether citizens' response to punishment is weaker in BRIBE, and whether this is due to less severe punishment.

Result 2. *In BRIBE, punishment is too soft to discipline citizens.*

[14]Note that we do not consider the last period in order to account for the often-discussed end-game effects.

[15]OLS regressions of contributions on a linear time trend indicate a positive and significant trend in NO BRIBE (Wald test, $p < 0.01$) and no trend in BRIBE (Wald test, $p = 0.48$).

[16]As noted by Johnston (1998), "the most serious cases of corruption are entrenched political and bureaucratic corruption and such systems are tightly organized, internally stable and do not automatically result in collapse."

Figure 2a displays the average change in contribution from period $t-1$ to period t, conditional on having been punished in $t-1$ or not. The left-hand side of the figure indicates that subjects who did not receive punishment points in $t-1$ significantly reduce their contributions in t in both treatments (-0.68 in NO BRIBE and -0.7 in BRIBE, Wald tests, both $p < 0.001$). These decreases in contribution are not significantly different from each other (Wald test, $p = 0.91$), indicating that the dynamics of contributions of citizens that do not receive punishment points are similar in the two treatments (OLS regressions confirm this finding, see Table S4 columns 1-2 in Appendix).

The right-hand side of the figure depicts the extent to which contributions increase after having received punishment points in the previous period. After having been assigned punishment points in $t-1$, subjects significantly increase their contribution in t in the two treatments (+2.19, $p < 0.001$ in NO BRIBE and +1.15, $p < 0.001$ in BRIBE). However, the increase in contribution is twice as large in NO BRIBE than in BRIBE (Wald test, $p < 0.01$).[17] These results are robust to OLS regressions including a variety of control variables such as group average contribution in the previous period, period fixed effects and demographics (see Table S4, columns 3-4 in Appendix).

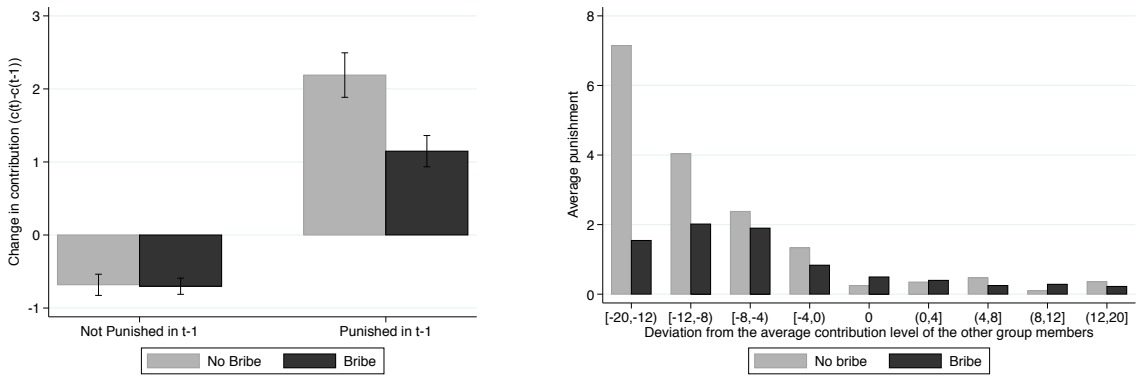
What can explain that contributions do not raise as much after punishment in BRIBE as compared to NO BRIBE? Is it because punishment points are lower in BRIBE? In Figure 2b, we plot the average punishment points inflicted to a citizen as a function of his deviation (positive or negative) from the average contribution of the other members of his group. For example, a citizen that contributed between 12 and 20 tokens less than the average contribution of the other members of his group received an average of 7.15 punishment points in the NO BRIBE condition. Consistent with previous findings (see e.g. Fehr and Gächter, 2000), punishment is mostly targeted at below-average contributors in both treatments. Notice also that larger negative deviations from the group average result in larger punishments and that a zero deviation is virtually not sanctioned.[18]

This figure reveals a clear difference between the treatments: While individuals that contribute the group average or more than the group average are not punished significantly more heavily in one treatment than in another (Mann-Whitney U tests, $p = 0.37$ for those contributing the group average and $p = 0.93$ for those contributing more than the group average), the punishment of below-average contributors is much softer in BRIBE than in NO BRIBE (Mann-Whitney U test, $p < 0.01$).[19] For example, a negative

[17]Note that Baldassarri and Grossman (2011) document an average response to punishment in their randomly elected monitor condition that is very similar to ours. While in our NO BRIBE condition being punished in $t-1$ translates into an approximate 15% increase in contribution in t, Baldassari and Grossman report that punishment leads to an average increase in contributions of 12%.

[18]Note that the data reveal instances of anti-social punishment (0.34 punishment points of above-average deviations in treatment NO BRIBE, Wald test $p < 0.01$).

[19]These results are supported by a two-sample Kolmogorov-Smirnov test of identical distributions



(a) (b)

Figure 2: Panel a) Average change in contribution ($c_t - c_{t-1}$) by treatment, with clustered standard errors at the group level. "Not punished in $t-1$" indicates citizens that did not receive punishment points in $t-1$. "Punished in $t-1$" indicates citizens that received any positive punishment point in $t-1$. Panel b) Punishment points received as a function of the deviation from the average contribution of the other group members.

deviation of 8 to 12 tokens from the group average is sanctioned by an average of 4.04 punishment points in NO BRIBE while it is sanctioned by an average of only 2.02 points in BRIBE (Wald test, $p = 0.01$). Note that the difference between the average sanction inflicted to citizens in NO BRIBE and in BRIBE increases in the size of the (negative) deviation from the average contribution of the other members of the group, i.e., it is mostly large free-riders that benefit – on average – from an environment in which bribery is possible. These result are broadly supported by OLS regressions, as documented in Table 1. In particular, below-average contributors are sanctioned significantly less harshly (about 1.3 token less) in BRIBE compared to NO BRIBE (column 1, $p < 0.01$). This result holds also after controlling for the average contribution of the other group members, period fixed effects and the individual characteristics of the monitor (column 2, $p < 0.05$).

Very similar findings are reported at the extensive margin. On average, a below-average contributor is 23.7 to 28.2 percentage-points less likely to be punished in BRIBE than in NO BRIBE ($p < 0.01$, see columns 3-4 in Table 1).

($D = 0.33$, $p = 0.01$).

Table 1: Effects of BRIBE on punishment (intensive and extensive margin)

	Severity of punishment		Probability of punishment	
	(1)	(2)	(3)	(4)
Treatment Bribe	0.103	0.082	0.055	0.027
	(0.114)	(0.151)	(0.045)	(0.046)
Below-average contribution	2.199***	2.158***	0.541***	0.511***
	(0.455)	(0.472)	(0.067)	(0.064)
Below-average contribution × Bribe	-1.328***	-1.270**	-0.282***	-0.237***
	(0.493)	(0.514)	(0.084)	(0.079)
Others' average contribution ($\bar{C}_{-i,t}$)		-0.029**		-0.011***
		(0.013)		(0.003)
Constant	0.292***	0.333	0.134***	0.128
	(0.072)	(0.535)	(0.029)	(0.110)
Periods fixed effects	No	Yes	No	Yes
Controls	No	Yes	No	Yes
R^2	0.191	0.222	0.185	0.250
# Clusters	56	56	56	56
Observations	3360	3360	3360	3360

Notes: OLS estimations. Standard errors (clustered at the group level) are displayed in parentheses. In columns 1-2, the dependent variable is the number of punishment points assigned to citizens (0 to 10) in t, in columns 3-4 the dependent variable is 1 if the citizen received punishment points, and 0 otherwise. Below-average contribution is a dummy variable. Controls include monitor-specific dummies for gender, German mother tongue and economics major. $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

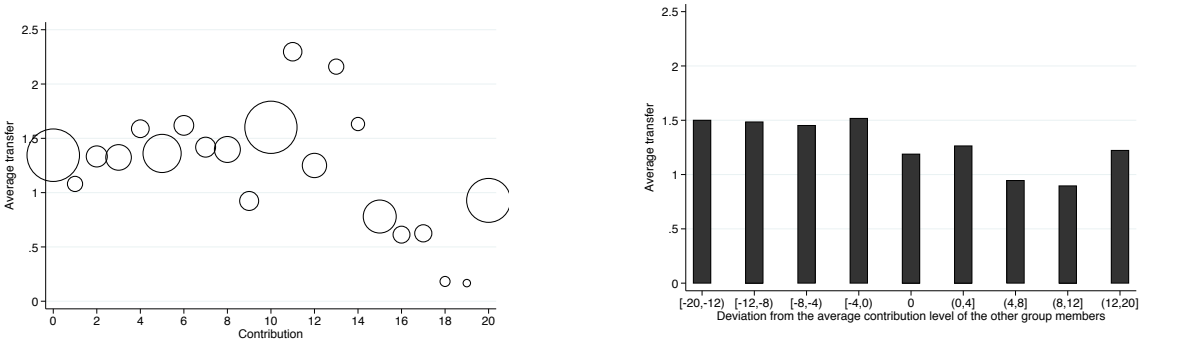
Overall, these findings provide compelling evidence that the monitors' behavior is largely affected by the treatment. While they tend to sanction the free-riders often and harshly in NO BRIBE, they are much more lenient in BRIBE. These finding are particularly remarkable when considered in light of Result 1. Because the average contribution is much lower in BRIBE than in NO BRIBE, one could have expected the punishment inflicted to below-average contributors in BRIBE to be tougher than in NO BRIBE. The fact that we observe a significant reduction in the punishment of below-average contributors between the two treatments despite lower contributions in BRIBE suggests that the punishment motives are indeed different in the two treatments. In what follows, we show that the monitors tend to reciprocate bribes received by below-average contributors by assigning them fewer deduction points than to above-average contributors.

Result 3. *In BRIBE, monitors reciprocate to bribery by assigning fewer deduction points to below-average contributors.*

Throughout the experiment, citizens made a large use of the bribes. They transferred a positive amount of tokens to the monitor in almost 60 % of the cases and

the maximum amount (5 tokens) in about 7 % of the cases (see Figure S3 in Appendix). The average transfer is 1.30 (i.e., 26% of the transfer endowment) and is constant over the 20 periods (OLS trend test, $p = 0.85$). These findings are broadly consistent with the bribery rates documented in previous studies (albeit involving very different experimental designs).[20]

While Figure 3a suggests that there is no strong substitution between the average contribution to the public good and the average transfer sent to the monitor (Wald test, $p = 0.11$),[21] Figure 3b clearly indicates that the transfers depend on the deviation from the average contribution of the group: the average transfer sent to the monitor by below-average contributors is on average 30% higher than the transfer sent by above-average contributors ($t - test = 2.91$, $p < 0.01$).[22]



(a) Average transfers and contribution levels

(b) Average transfers and contribution deviations

Figure 3: Average transfers and contributions (panel a) and average transfer over contribution deviations (panel b). The size of the dots in panel a indicate the relative frequency of observations for each of the different contributions levels.

Is there a reason why below-average contributors bribe more? Do they benefit from doing so? Figure 4 depicts the average punishment received by citizens in BRIBE, conditional on the deviation from the average cooperation level of the other group members and on whether they bribed ($T > 0$) or did not bribe the monitor ($T = 0$).[23] From

[20]For example, Gneezy et al. (2016) report bribery rates ranging from 44% to 74%. In Abbink et al. (2002), subjects enter a bribery relationship in 22% to 25% of the cases and bribe the full amount in 12% of the cases.

[21]While Figure 3a might suggest that the relationship between average transfers and contribution is zero up to contribution level 10 and slightly negative from 10 to 20, the negative slope is only very marginally significant (Wald test, $p = 0.1$). While perhaps surprising, this finding is in line with previous studies (see e.g. Gneezy et al., 2016).

[22]The average transfer of an above-average contributor is 1.15 tokens whereas the average transfer of a below-average contributor is 1.5.

[23]Since we only consider the observations in BRIBE, we have fewer observations. For that reason, we grouped the deviations from the average cooperation into wider bins.

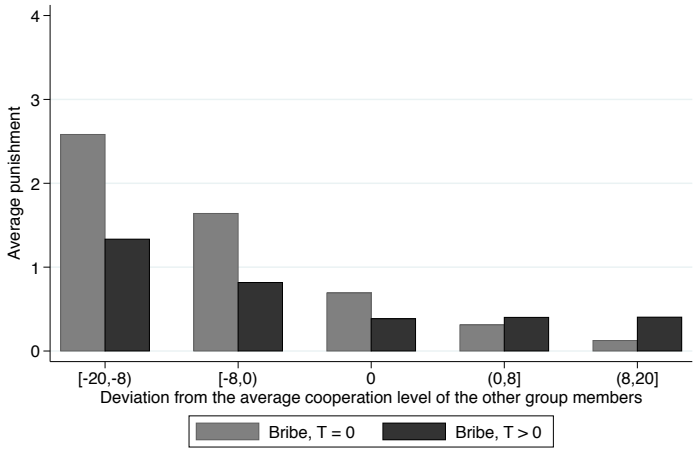


Figure 4: Punishment points received as a function of the deviation from the average contribution of the other group members in BRIBE, by transfer level (no transfer, $T = 0$, or positive transfer, $T > 0$).

this picture, it is clear that below average contributors benefit from bribing the monitor: bribes allow them to dramatically reduce the number of punishment points received ($p < 0.05$). For example, a citizen that contributed between 8 and 20 points less than the average is assigned 2.5 punishment points when she does not bribe and only 1.2 points when she bribes. Hence, bribing the monitor allows such a citizen to reduce punishment by more than 50%.

In Table 2, we regress the punishment points received by the citizens in BRIBE on their contributions and their bribes (columns 1 to 6). We run separate regressions for citizens that contribute less than the average, exactly the average or more than the average of their group. Columns 1 and 2 document that below-average contributors benefit the most from bribing the monitor: An additional token transferred to the monitor reduces the number of punishment points they receive by 0.17 ($p = 0.06$). In contrast, citizens that contribute the average (columns 3-4) or more (columns 5-6) do not receive significantly fewer punishment points if they increase their transfers to the monitor. On the other hand, higher contributions to the public good would allow each type of citizen to reduce the number of punishment points received.

Table 2: The effect of transfers and contributions on punishment points

	$C_{i,t} < \bar{C}_{-i,t}$		$C_{i,t} = \bar{C}_{-i,t}$		$C_{i,t} > \bar{C}_{-i,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Transfer	-0.175*	-0.173*	-0.101	-0.075	0.029	0.038
	(0.088)	(0.088)	(0.105)	(0.082)	(0.031)	(0.031)
Contribution	-0.091**	-0.116***	-0.069***	-0.092***	-0.019*	-0.027**
	(0.033)	(0.036)	(0.024)	(0.025)	(0.010)	(0.012)
Constant	2.002***	1.536	1.335***	2.449*	0.524***	0.444*
	(0.456)	(0.919)	(0.394)	(1.432)	(0.182)	(0.227)
Periods fixed effects	No	Yes	No	Yes	No	Yes
Controls	No	Yes	No	Yes	No	Yes
R^2	0.080	0.133	0.141	0.351	0.019	0.070
# Clusters	28	28	28	28	28	28
Observations	659	659	372	372	649	649

Notes: OLS estimations. The dependent variable is punishment points received (columns 1 to 6). Standard errors (clustered at the group level) are displayed in parentheses. Controls include citizen-specific dummies for gender, German mother tongue and economics major. Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

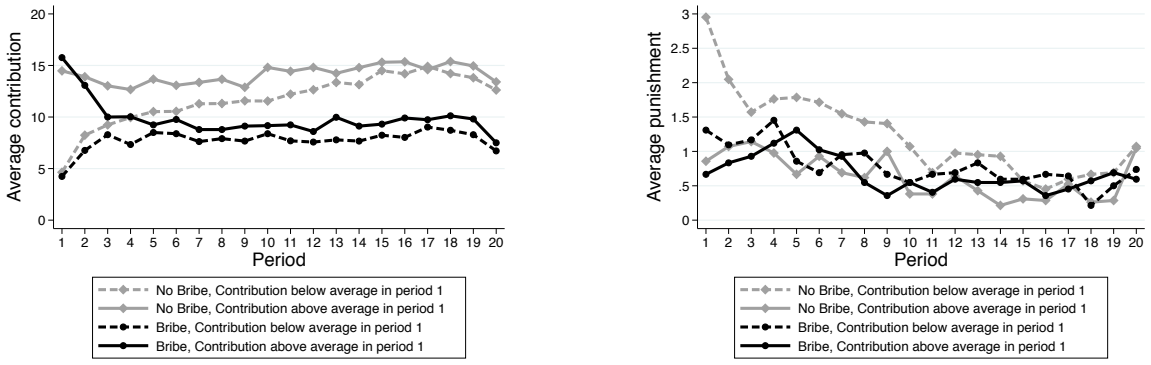
Finally, note that questionnaire measures elicited at the end of the experiment suggest that citizens purposefully use bribes to reduce the punishment points they receive. In particular, citizens expect low contributors to use bribes to avoid punishment and also expect monitors to be more lenient with bribers. Moreover, monitors find it normal to reciprocate bribes despite the fact that they understand that citizens use them to avoid punishment (See Figure S4 in Appendix).[24]

Up to this point, we have analyzed how bribery affects the monitors, and how it is used by below-average contributors to attenuate punishment. But can bribery also have negative spillover effects on "*good citizens*," i.e. citizens that contribute more than the average? In what follows, we show this is indeed the case. One important channel through which bribery undermines cooperation is through a discouragement of the "*good citizens*."

Result 4. *Citizens who contribute more than the average in period 1 (initially high contributors) dramatically reduce their contribution in BRIBE but not in NO BRIBE, suggesting that they get discouraged in BRIBE.*

[24]Roughly 70% of citizens find it normal that a monitor assigns fewer deduction points after a transfer (S4a), while 60% think a transfer is made in the hope of receiving fewer deduction points (S4b). Monitors have similar yet stronger beliefs: 85% of monitors find normal to assign fewer deduction points following a transfer (S4c), and 60% understand citizens make transfers in the hope of being punished less (S4d).

In order to study whether bribery has negative spillover effects on "*good citizens*," we divide our sample into two categories: citizens that contributed more than the average contribution of the other members of their group in period 1 (*initially high contributors*) and citizens that contributed less (*initially low contributors*). We then plot the evolution of the average contribution of these two types of individuals as a function of the treatment. In addition, we also plot the average punishment received by these two types of citizens in the two different treatments. The results are depicted in Figure 5.[25]



(a) Contributions (b) Punishment points

Figure 5: Evolution of (a) contributions and (b) punishment points over the 20 periods for *initially* low or high contributors in BRIBE and NO BRIBE. An initially low (high) contributor is defined as a citizen contributing strictly less (weakly more) than the average contribution of the other group members in period 1.

Two important findings emerge from these panels. First, note that the average contribution in period 1 of initially high contributors (solid lines) is equal to approximately 75% of their endowment, both in BRIBE and in NO BRIBE (Figure 5a). Similarly, the average contribution in period 1 of initially low contributors (dashed lines) corresponds to approximately 25% of their endowment in both treatments. Hence, the treatment does not affect the behavior in period 1 of these two types of citizens. However, the treatment immediately affects the behavior of the monitors, as indicated in Figure 5b. Indeed, the average punishment received by an initially low contributor in period 1 in the NO BRIBE condition is much higher than the average punishment received by an initially low contributor in BRIBE. In period 1, initially low contributors in NO BRIBE receive a punishment that is on average of 2.25 times higher than in BRIBE (Wald test,

[25] As a robustness check, note that we also divided the sample depending on how many times a citizen contributed more than the average of the other group members over either the first 2 or the first 3 periods (see Figures S5a and S5b in Appendix). We also performed the analysis by defining a high contributor as a citizen whose contribution in period 1 is higher than the average contribution at the *treatment* level (see Figure S5c) or higher than the group median contribution in period 1 (Figure S5d). All these robustness checks yield very similar results.

$p < 0.01$), despite the fact that their average contribution to the public good are virtually identical.[26] As for initially high contributors, they are not sanctioned differently in the two treatments.[27]

Second, initially low contributors largely increase their contribution to the public good in NO BRIBE, whereas they keep on contributing only very little in BRIBE. Indeed, striking is that, in NO BRIBE, initially low contributors rise their contribution from 5 in period 1 to 12.62 in period 20. As of period 8, the average contribution of initially low and initially high contributors are statistically indistinguishable from each other in NO BRIBE (Wald test, $p = 0.16$). The picture is very different in BRIBE as it is initially high contributors who strongly decrease their contributions down to the level of initially low contributors. We speculate that this drop in contribution is the consequence of a discouragement effect: initially high contributors in BRIBE get discouraged by the fact that low contributors can get away from harsh sanctions and do not raise their contributions. This discouragement quickly leads them to reduce their contribution dramatically. Indeed, between period 1 and period 3 the average contribution of an initially high contributor in BRIBE drops from 15.76 to 10 (Wald test, $p < 0.001$). In this treatment, the average contribution of initially high contributors is statistically indistinguishable from the average contribution of initially low contributors as of period 3 already.

Quite astonishingly, the contributions of all the citizens in BRIBE converge towards an equilibrium involving low levels of cooperation, whereas all the citizens in NO BRIBE converge towards an equilibrium in which cooperation is high. Interestingly, it is the initially low contributors that catch up with the initially high contributors in NO BRIBE, whereas it is the initially high contributors that decrease their contributions down to the levels of initially low contributors in BRIBE.

As initially low contributors increase their contributions in NO BRIBE, the punishment points they receive gradually drop (Figure 5b). The picture is less clear in BRIBE. While initially high contributors get more heavily punished as their contributions drop from period 1 to period 4, the average deduction they receive decreases again after period 4 despite the fact that they do not increase their contribution.

Before concluding the paper, we investigate how bribery affects profits. In particular, we answer the following questions: Does bribery allow below-average contributors to increase their profit? Are the citizens better off in BRIBE or in NO BRIBE? Which

[26] Initially low contributors in the NO BRIBE condition receive on average 2.95 punishment points in NO BRIBE in period 1, while they receive an average of 1.31 punishment points in BRIBE.

[27] Initially high contributors receive on average 0.86 punishment points in the NO BRIBE condition while they receive an average of 0.66 punishment points in BRIBE ($p = 0.65$).

situation is better for the monitors?

Result 5. *Below-average contributors can decrease punishment through bribery, but this strategy does not increase their profit. In the long run, citizens are better off in NO BRIBE. Monitors' profit are unaffected by the treatment. Overall efficiency is higher in NO BRIBE.*

In Table 3, we regress the profit (columns 1 to 6) of the citizens in BRIBE on their contributions and their bribes. While we previously showed that bribes allow below-average contributors to reduce the number of punishment points received (Table 2), Table 3 shows that they do not significantly affect their profits (columns 1-2). Finally, note that bribes decrease the profit of citizen that contribute at least the average of their group, as indicated by columns 3 to 6 ($p < 0.05$ and $p < 0.01$).

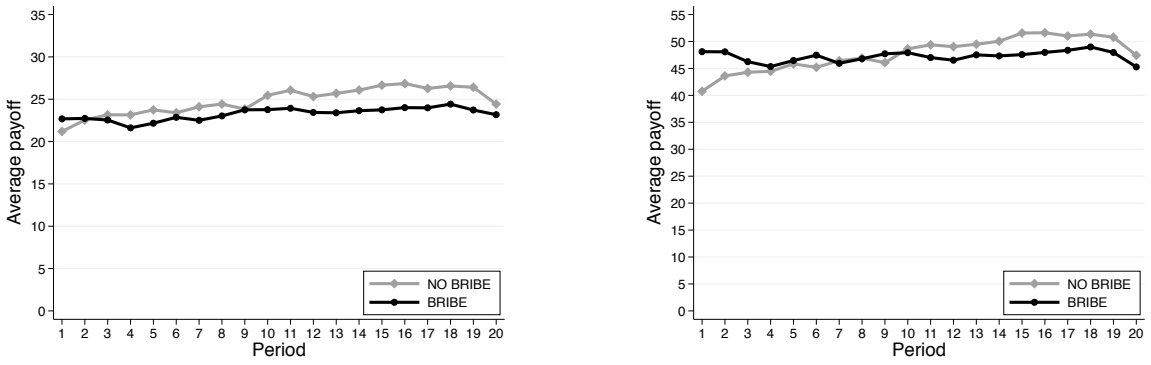
Table 3: The effect of transfers and contributions on payoffs

	$C_{i,t} < \bar{C}_{-i,t}$		$C_{i,t} = \bar{C}_{-i,t}$		$C_{i,t} > \bar{C}_{-i,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Transfer	-0.387	-0.411	-0.697**	-0.786***	-1.008***	-0.952***
	(0.233)	(0.257)	(0.315)	(0.252)	(0.210)	(0.187)
Contribution	0.183*	0.216**	0.408***	0.469***	-0.117**	-0.078
	(0.095)	(0.093)	(0.072)	(0.074)	(0.053)	(0.060)
Constant	24.049***	25.826***	20.995***	17.090***	23.928***	19.506***
	(1.326)	(3.568)	(1.183)	(4.189)	(0.508)	(1.459)
Periods fixed effects	No	Yes	No	Yes	No	Yes
Controls	No	Yes	No	Yes	No	Yes
R^2	0.027	0.129	0.484	0.613	0.139	0.276
# Clusters	28	28	28	28	28	28
Observations	659	659	372	372	649	649

Notes: OLS estimations. The dependent variable is the period-profit. Standard errors (clustered at the group level) are displayed in parentheses. Controls include citizen-specific and monitor-specific dummies for gender, German mother tongue and economics major. Levels of significance: $*p < 0.1, **p < 0.05, ***p < 0.01$.

Figure 6 depicts the evolution of the average payoff of citizens (panel a) and monitors (panel b) by treatment. Over the first 10 periods, the average profit of citizens in BRIBE and in NO BRIBE are not significantly different from each other ($p = 0.40$). However, in the long run, citizens are better off in NO BRIBE. Indeed, over periods 11-20 a citizen earns about 8% more in NO BRIBE than in BRIBE (32.06 vs 29.6, $p < 0.01$). Monitors' payoff remain unaffected by the treatment, as indicated by panel b. The overall efficiency (see Figure S6 in Appendix), measured as the average payoff of all the members

of a group, is significantly higher in NO BRIBE than in BRIBE (taking periods 11 to 20, $p = 0.01$).



(a) Citizens (b) Monitors

Figure 6: Evolution of payoffs by treatment for (a) citizens and (b) monitors.

5 Conclusion

Humans enforce rules and norms through sanctions. Historically, the experimental literature has focused on peer to peer punishment. However, many societies choose to delegate punishment to a centralized authority. For example, law enforcement is delegated to the judicial system and to the police in most countries. More traditional societies have chiefs that enforce social norms. While several studies have documented that centralized punishment successfully fosters cooperation, only very little attention has been devoted to studying whether there are also perils to centralization. In this paper, we study whether corruption affects norm enforcement. Given the prevalence of bribery worldwide[28] and its potential implications for norm enforcement, this is an important question to investigate.

Until today, a large empirical literature has been devoted at measuring corruption and its effects. While some have argued that corruption might increase efficiency (Huntington, 1968), the lion's share of the empirical evidence suggests that it imposes large costs on the society (see e.g. Mauro, 1995; Méon and Sekkat, 2005; Bertrand et al., 2007). However, studying corruption in natural settings generates an important challenge: How to measure bribery in the field? Due to its very nature, bribery is secretive and hence hard to observe. We circumvent this issue by leveraging experimental methods. In our design, bribery is easy to measure and its causal effects on norm enforcement and cooperation can be unambiguously established. Our results indicate that bribery

[28]In 2016, it was estimated that more than 1 trillion of USD was spent on bribery worldwide (World Bank Group, retrieved from `http://www.worldbank.org/en/topic/governance/brief/anti-corruption` on February 22, 2017.)

has two important negative effects. First, we show that citizens (in particular the free-riders) purposefully use bribes in order to reduce the amount of punishment points they receive. The monitors tend to reciprocate these bribes by allocating fewer punishment points to bribers. Second, our results also indicate that bribery has negative spillovers on "*good citizens*." In our setting, initially high contributors in the BRIBE treatment get discouraged. Ultimately, these two forces prevent cooperation rates to increase over time, as opposed to what is observed in a control condition in which bribery is ruled out by design.

Overall, our paper provides clean causal evidence that corruption undermines norm enforcement. Our results speak to the literature on public goods, and more generally to the literature on centralized punishment institutions (Kosfeld et al., 2009; Andreoni and Gee, 2012; Nicklisch et al., 2016; Fehr and Williams, 2018). While many studies have discussed the benefits of centralized institutions, our results clearly highlights a potential weakness of centralization: corruption.

In our opinion, these findings open several interesting avenues for future research. For example, it would be very interesting to investigate whether democratic elections between monitors (or other public policies) can mitigate bribery and its effects on cooperation. It would also be very instructive to uncover whether different monitors react differently to bribery. Previous studies have shown that different leaders manage their public goods very differently (Kosfeld and Rustagi, 2015). This might also apply to bribery. In particular, there might be a large heterogeneity in the way monitors are influenced by bribes. Our design could easily be extended and applied to the study of these questions.

References

Abbink, Klaus, "Fair Salaries and the Moral Costs of Corruption," *Advances in Cognitive Economics, Sofia: NBU Press*, 2005.

_ **and Danila Serra**, "Anticorruption policies: Lessons from the lab," *New Advances in Experimental Research on Corruption*, 2012, *15*.

_ **, Bernd Irlenbusch, and Elke Renner**, "An experimental bribery game," *Journal of Law, Economics, and Organization*, 2002, *18* (2), 428–454.

Andreoni, James and Laura K Gee, "Gun for hire: delegated enforcement and peer punishment in public goods provision," *Journal of Public Economics*, 2012, *96* (11-12), 1036–1046.

Baldassarri, Delia and Guy Grossman, "Centralized sanctioning and legitimate authority promote cooperation in humans," *Proceedings of the National Academy of Sciences*, 2011, *108* (27), 11023–11027.

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

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, reciprocity, and social history," *Games and Economic Behavior*, 1995, *10* (1), 122–142.

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

Cameron, Lisa, Ananish Chaudhuri, Nisvan Erkal, and Lata Gangadharan, "Propensities to engage in and punish corrupt behavior: Experimental evidence from Australia, India, Indonesia and Singapore," *Journal of Public Economics*, 2009, *93* (7), 843–851.

Fehr, Ernst, "Human behaviour: don't lose your reputation," *Nature*, 2004, *432* (7016), 449.

_ **and Simon Gächter**, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 2000, *90* (4), 980–994.

_ **and** _ , "Altruistic punishment in humans," *Nature*, 2002, *415* (6868), 137–140.

_ **and Tony Williams**, "Social Norms, Endogenous Sorting and the Culture of Cooperation," 2018.

_ , **Simon Gächter, and Georg Kirchsteiger**, "Reciprocity as a contract enforcement device: Experimental evidence," *Econometrica*, 1997, pp. 833–860.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 2007, *10* (2), 171–178.

Gneezy, Uri, Silvia Saccardo, and Roel Van Veldhuizen, "Bribery: Behavioral drivers of distorted decisions," *Journal of the European Economic Association (forthcoming)*, 2016.

Greiner, Ben, "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 2015, *1* (1), 114–125.

Huntington, Samuel P, *Political order in changing societies*, Yale University Press New Haven, CT, 1968.

Johnston, Michael, "What can be done about entrenched corruption?," in "Annual world bank conference on development economics 1997" World Bank Washington DC 1998, pp. 69–90.

Kosfeld, Michael, "Corruption within a Cooperative Society," Economics Series 48, Institute for Advanced Studies July 1997.

_ , **Akira Okada, and Arno Riedl**, "Institution formation in public goods games," *American Economic Review*, 2009, *99* (4), 1335–55.

_ **and Devesh Rustagi**, "Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia," *American Economic Review*, 2015, *105* (2), 747–783.

Malmendier, Ulrike and Klaus M Schmidt, "You owe me," *American Economic Review*, 2017, *107* (2), 493–526.

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran, "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes," *Review of Economic Studies*, 2014, *81* (1), 301–324.

Mauro, Paolo, "Corruption and growth," *Quarterly Journal of Economics*, 1995, pp. 681–712.

Méon, Pierre-Guillaume and Khalid Sekkat, "Does corruption grease or sand the wheels of growth?," *Public Choice*, 2005, *122* (1-2), 69–97.

Muthukrishna, Michael, Patrick Francois, Shayan Pourahmadi, and Joseph Henrich, "Corrupting cooperation and how anti-corruption strategies may backfire," *Nature Human Behaviour*, 2017.

Nicklisch, Andreas, Kristoffel Grechenig, and Christian Thöni, "Information-sensitive Leviathans," *Journal of Public Economics*, 2016, *144*, 1–13.

Nikiforakis, Nikos, "Punishment and counter-punishment in public good games: Can we really govern ourselves?," *Journal of Public Economics*, 2008, *92* (1-2), 91–112.

Olken, Benjamin A, "Corruption perceptions vs. corruption reality," *Journal of Public Economics*, 2009, *93* (7), 950–964.

Panchanathan, Karthik and Robert Boyd, "Indirect reciprocity can stabilize cooperation without the second-order free rider problem," *Nature*, 2004, *432* (7016), 499.

Veldhuizen, Roel Van, "The influence of wages on public officials' corruptibility: A laboratory investigation," *Journal of Economic Psychology*, 2013, *39*, 341–356.

Appendix

Additional Tables and Figures

Table S1: Check of randomization of treatments to citizens

	BRIBE		NO BRIBE		Test	
	Mean	S.D.	Mean	S.D.	F stat	P-value
Male (= 1)	0.333	(0.474)	0.452	(0.501)	2.503	0.116
German mother tongue (= 1)	0.833	(0.375)	0.893	(0.311)	1.254	0.264
Economics (= 1)	0.452	(0.501)	0.488	(0.503)	0.213	0.645
Observations	84		84		168	

Notes: Variables include a dummy for gender, a dummy for German mother tongue and a dummy for economics major.

Table S2: Check of randomization of treatments to monitors

	BRIBE		NO BRIBE		Test	
	Mean	S.D.	Mean	S.D.	F stat	P-value
Male (= 1)	0.321	(0.476)	0.500	(0.509)	1.839	0.181
German mother tongue (= 1)	0.893	(0.315)	1.000	(0.000)	3.240	0.077
Economics (= 1)	0.321	(0.476)	0.429	(0.504)	0.669	0.417
Observations	28		28		56	

Notes: Variables include a dummy for gender, a dummy for German mother tongue and a dummy for economics major.

Table S3: Contribution in t ($C_{i,t}$), Periods 1-20

	Periods 1-20	
	(1)	(2)
Treatment Bribe	-4.136***	-4.006***
	(1.291)	(1.313)
Constant	11.848***	10.849***
	(0.952)	(1.486)
Periods fixed effects	Yes	Yes
Controls	No	Yes
R^2	0.092	0.096
# Clusters	56	56
Observations	3360	3360

Notes: OLS estimation. Standard errors (clustered at the group level) are displayed in parentheses. The dependent variable is the contribution level in period t. Controls include citizen-specific dummies for gender, German mother tongue and economics major. Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table S4: Change in contribution from $t-1$ to t ($C_{i,t} - C_{i,t-1}$).

	Not punished at $t-1$		Punished at $t-1$	
	(1)	(2)	(3)	(4)
Treatment Bribe	-0.020	0.275	-1.042***	-0.848**
	(0.183)	(0.186)	(0.372)	(0.384)
Others' average contribution ($\bar{C}_{-i,t-1}$)		0.068***		0.040
		(0.016)		(0.043)
Constant	-0.682***	-2.911***	2.190***	1.204
	(0.145)	(0.641)	(0.305)	(1.046)
Periods fixed effects	No	Yes	No	Yes
Controls	No	Yes	No	Yes
R^2	0.000	0.035	0.010	0.039
# Clusters	56	56	53	53
Observations	2238	2238	954	954

Notes: OLS estimation. Standard errors (clustered at the group level) are displayed in parentheses. The dependent variable is the change in contribution from period $t-1$ to period t. Controls include citizen-specific dummies for gender, German mother tongue and economics major. Levels of significance: $^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Figure S1: Screenshot of the monitors' decision screen in the BRIBE treatment. Note that in the NO BRIBE treatment, the line "transfer to the monitor" did not appear.

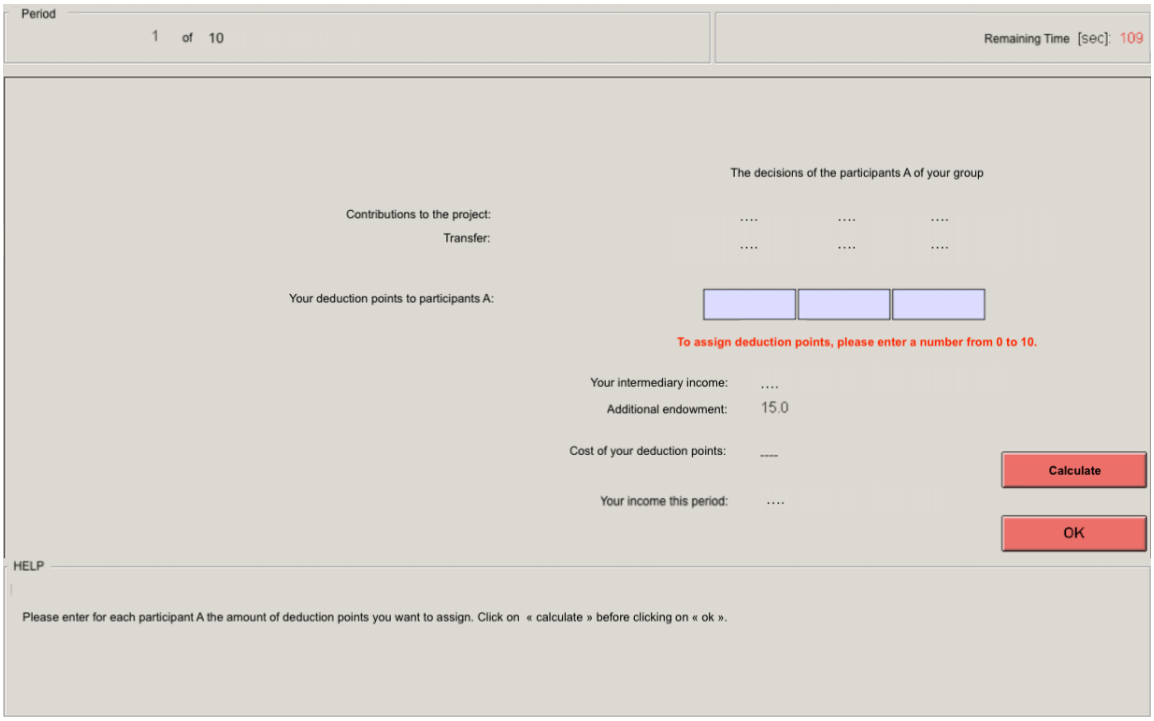


Figure S2: Screenshot of the citizens' final screen in the BRIBE treatment. Note that in the NO BRIBE treatment, the line "transfer to the monitor" did not appear.

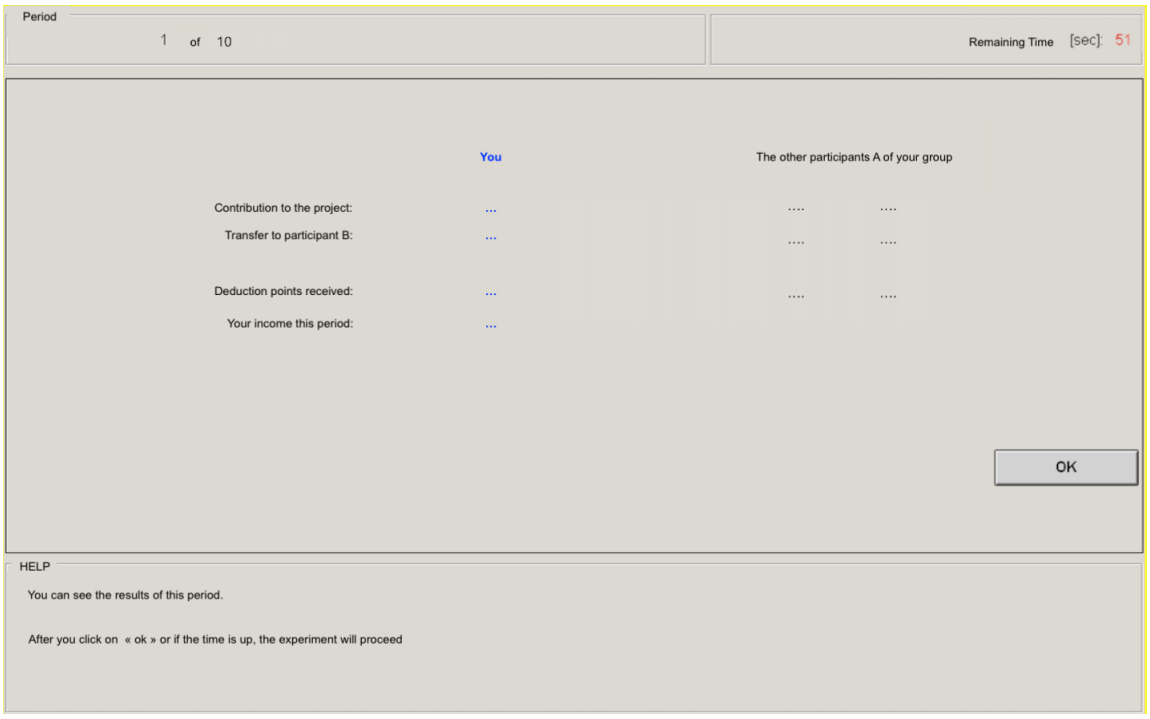
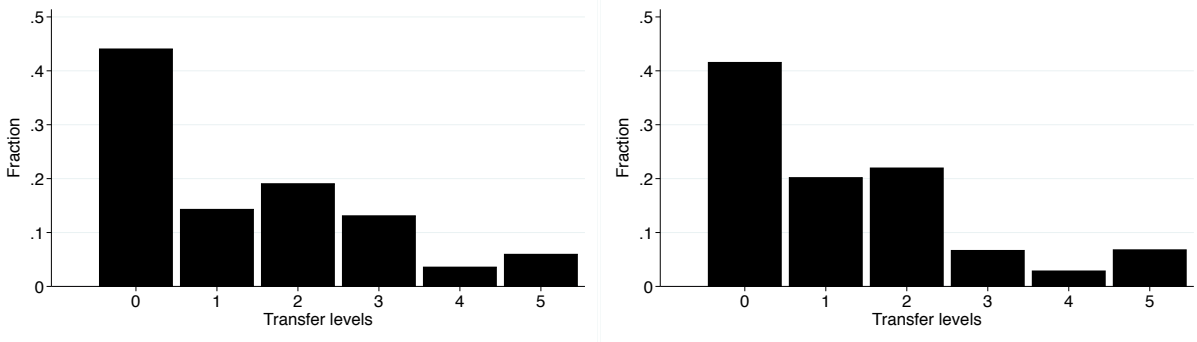


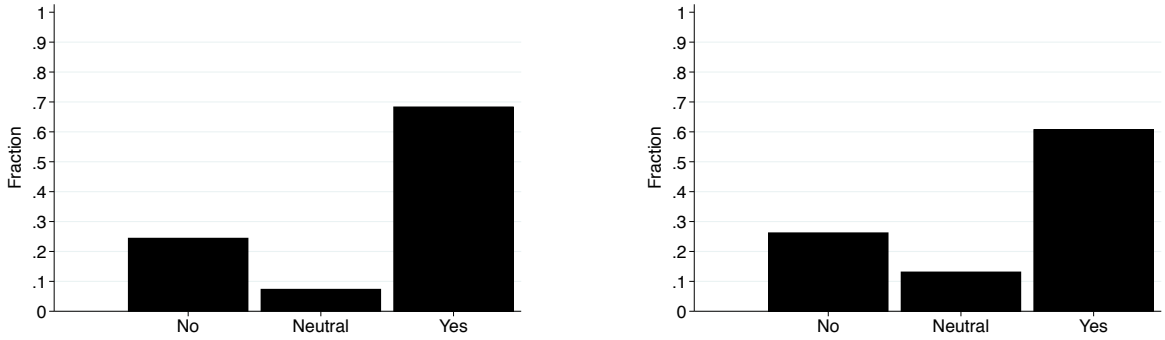
Figure S3: Distributions of transfers



(a) Period 1

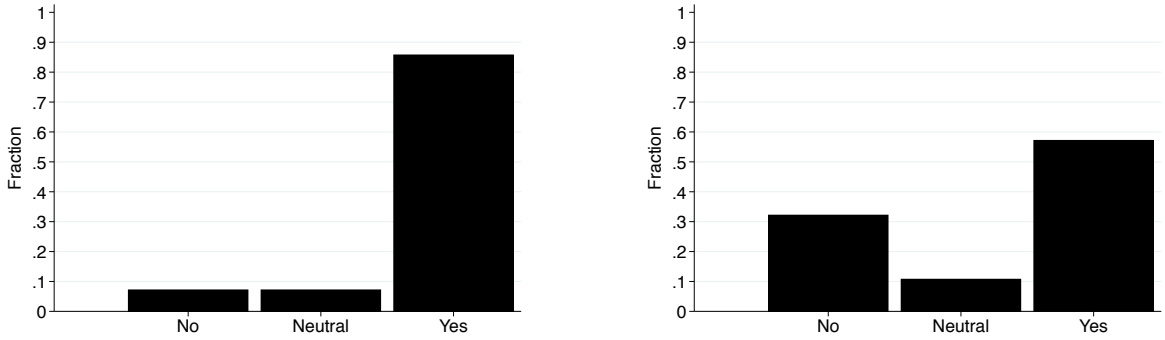
(b) Periods 1-20

Figure S4: Citizens (panel a and b) and monitors' beliefs (panel c and d)



(a) Citizens' beliefs

(b) Citizens' beliefs

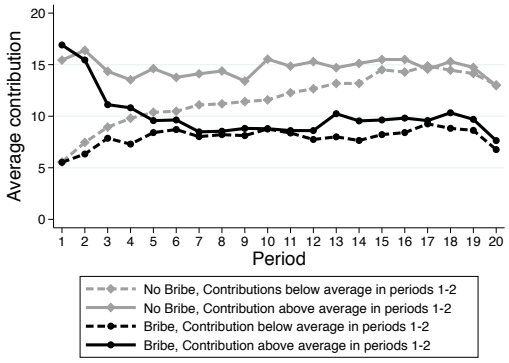


(c) Monitors' beliefs

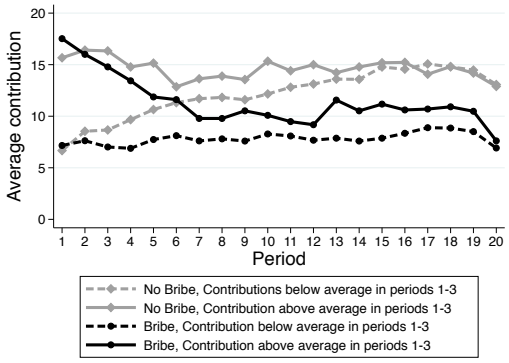
(d) Monitors' beliefs

Panel (a) and (c): Subjects' agreement with the statements "It is normal that participant B *[the monitor]* assigns fewer deduction points to a participant A *[a citizen]* who transfers tokens". Panel (b) and (d): Subjects' agreement with the statement "Participants A *[Citizens]* transfer tokens to participant B *[the monitor]* in the hope of receiving fewer deduction points". Answers range from 1 ("I fully disagree") to 7 ("I fully agree") and are collapsed into three categories: "No" (1-3), "Neutral" (4) and "Yes" (5-7). Words in bracket were not displayed in the original questionnaire.

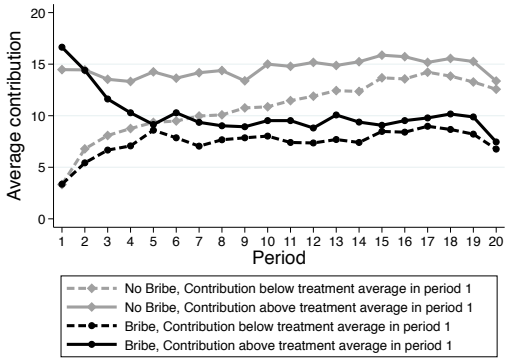
Figure S5: Robustness checks for the definition of initially high contributors.



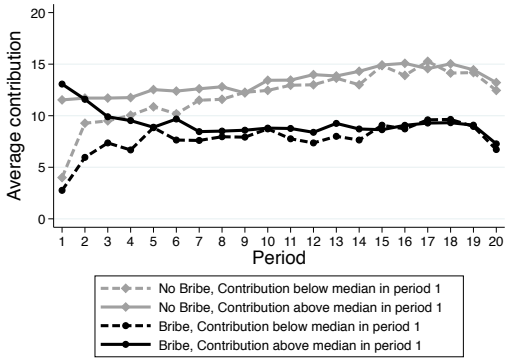
(a) Initially high contributor defined as a citizen who contributed more than the avg contribution of his group in periods 1 and 2.



(b) Initially high contributor defined as a citizen who contributed more than the avg contribution of his group in periods 1, 2 and 3.



(c) Initially high contributor defined as a citizen who contributed more than the avg contribution of the treatment in period 1.



(d) Initially high contributor defined as a citizen who contributed more than the median contribution in period 1.

Figure S6: Overall efficiency (average payoff of the monitor and the 3 citizens)