Experimental Evidence on Tax Evasion, Corruption and and Incentives to Blow the Whistle

Johannes Buckenmaier\textsuperscript{a}, Eugen Dimant\textsuperscript{b,*}, Luigi Mittone\textsuperscript{c}

\textsuperscript{a}University of Cologne, Germany
\textsuperscript{b}University of Pennsylvania, USA
\textsuperscript{c}University of Trento, Italy

Abstract

We experimentally study how institutional changes that incentivize blowing the whistle through a leniency program affect tax compliance and corruption. In our experimental set up, we nest tax compliance within a corruption framework and systematically change the institutional environments in which participants have to declare taxes. Across four treatments, our design embodies between and within changes of institutions and thus allows us to analyze how deviant behavior and collusion is affected by both the history and the exogenous changes of institutions. We find higher tax compliance in the presence of a leniency program, while the introduction of such a regime causes the opposite reaction. Our results suggest that this effect is vastly different across gender and mainly driven by females. We provide evidence that the effectiveness of new political measures cannot reliably be judged in isolation, but must be considered in view of the actual institutional history, that is, the particular institutional framework in place before the measure is introduced.

Keywords: Corruption, Institutions, Whistleblowing, Tax Evasion

JEL: C91, D03, D73, H26

1. Introduction

Tax evasion is one of the most pervasive forms of illicit behavior. It induces negative externalities on both the economic and societal level [Banerjee, forthcoming; Slemrod].

\*Corresponding author at: Philosophy, Politics and Economics Program, and Behavioral Ethics Lab, University of Pennsylvania, 313 Logan Hall, 249, S. 36th Street, PA 19104, USA.

Email addresses: johannes.buckenmaier@uni-koeln.de (Johannes Buckenmaier), edimant@sas.upenn.edu (Eugen Dimant), luigi.mittone@unitn.it (Luigi Mittone)
Understanding its drivers and implementing suitable institutional measures to curb its severity has been at the center of the past decade’s theoretical, empirical and experimental research. Beyond theoretical exercises, such as the seminal work of Allingham and Sandmo (1972), research that analyzes tax evasion empirically in general and in particular experimentally has been growing, suggesting that individuals pay more taxes than what would be predicted by standard economics (Alm et al., 1992; Torgler, 2003). Tax compliance has also been studied experimentally in the field (Cummings et al., 2009; Hallsworth, 2014). For a general discussion see Andreoni et al. (1998) and Alm (2012).

While most of the economic research has focused on deterrence of income tax evasion or its related variants, other forms of tax evasion, such as trade/import or custom taxes, where taxes are in some way collected through the direct intermediation of a third party (for example custom duties), have received little attention in the experimental literature (Orviska and Hudson, 2003; Bø et al., 2015). This is particularly true for the case of “corruption within tax evasion”.

In contrast to existing experimental studies on this matter (Banuri and Eckel, 2012), we focus on the effectiveness of providing legal immunity to the bribe-giver for blowing the whistle as a measure to deter tax evasion embedded in a setting of collusive bribery, in which tax evasion can only be successful through cooperation among tax payers and public officials. The exchange of bribes is studied in our paper as one explicit collaboration-inducing mechanism, which has previously been found to be effective in sustaining illicit cooperation (Weisel and Shalvi, 2015). Abbink and Wu (2017) study whether rewarding self-reports is effective in reducing collusive bribery. They find this mechanism to be effec-

---

1We chose the environment of the experimental laboratory to ensure the highly controlled environment necessary for our design since this can be rarely achieved in the field, as pointed out by Siemrod and Weber (2012): “with regard to the empirical analysis of tax evasion and the informal economy, the credibility revolution has, for the most part, not yet arrived ... not because of inattention by creative empirical researchers ... [but] because severe measurement problems plague empirical analysis in this context, problems that arise not by chance, but because of the nature of the subject matter.” See also Hallsworth et al. (2014) for a recent study attempting to reduce these measurement difficulties.
tive in some circumstances, especially in a context of repeated interaction. Christöfl et al. (forthcoming) study the possibility to cooperate with the authorities (principal witness) in combination with a leniency policy that offers reduced fines for cooperation in a setup where two bidders compete for a contract. They find a lower number of bribes when a leniency policy is present, while at the same time offering a bribe becomes more profitable for a corrupt bidder. Closely related to our work, Heinemann and Kocher (2013) study the effects of regime changes on tax compliance, however, they focus on changes in the tax rate and consider neither corruption nor reforms that incentivize whistleblowing. By and large, the economics of whistleblowing are understudied and have only recently attracted attention (see Spagnolo, 2004; Apesteguia et al., 2007; Spagnolo, 2006; Heyes and Kapur, 2009; Breuer, 2013; Schmolke and Utikal, 2016). In particular, Butler et al. (2017) study the effectiveness of financial rewards and public scrutiny as triggers to motivate employees to blow the whistle against their managers. Their findings indicate that both financial rewards and public visibility increase the likelihood of whistleblowing (see also Bartuli et al., 2016). The recent surge in cases of whistleblowing and the lack of international institutional uniformity to achieve sufficient protection for whistleblowers renders the importance to further study the economics of whistleblowing (Dyck et al., 2010).

Here, we extend the Butler et al. (2017) study by shedding light on the effectiveness of lenience programs as a means to reduce tax fraud and distort collusive corrupt relationships between public officials and tax payers. In our setting, we introduce such incentives in the form of a safe way out when cooperating with the auditors and blowing the whistle on the corrupt public official, hence resembling the mechanism of a principal witness regulation in which audited tax fraudsters can turn state’s evidence. In many countries, the introduction of a principal witness regulation represents an integral institutional feature aimed at suppressing criminal behavior. In this paper, we are interested in examining the role of the institutional frame of tax-evasion behavior. We compare institutional settings with and without a principal witness regulation.
To the best of our knowledge this is the first experiment studying tax evasion nested in a corruption framework, and, in particular, the influence of institutional changes that embody the described incentives to blow the whistle. We contribute to the corruption and tax evasion literature by shedding light on how tax evasion is affected by the specifics of the strategic interaction of tax payers with an intermediary, a dimension not present in a setting of individual tax evasion. We use a controlled laboratory experiment modeling an income reporting scenario that requires the interaction between two parties, a tax payer and a tax officer, thus opening the door for collusive corruption. Existing research highlights the importance of studying the collaborative roots of deviant behavior due to their inherent negative economic and societal externalities (Weisel and Shalvi, 2015).

Our experimental design employs an extension of the standard tax evasion game (Allingham and Sandmo, 1972). First, we use a scenario of tax evasion nested in a corruption framework (Abbink et al., 2002), in which corrupt tax officers face little to no consequences. The idea was to mimic a situation where tax authorities do not have the means to sufficiently control the tax officers, for example due to the institutional environment rendering enforcement of adequate consequences impossible. Excessive costs of monitoring might be one reason why authorities are not able to detect dishonest officers. However, this could also be due to authority’s need to maintain a reasonable level of consensus among their tax officer’s, which can be undermined by a high level of control. In this basic variant of the game (bribery game) each tax payer receives a fixed income that is taxed at a fixed rate that he is asked to report to the authorities in the form of a tax officer. The first distinct feature of our design is that underreporting requires the cooperation of the tax officer whom the tax payer can offer a bribe as reward for his cooperation. Tax reports are subject to audits with a known probability. Conditional on successful underreporting of

---

2Our focus is on the effectivity of a PWR as a policy intervention, thus we decided to keep a fixed audit probability instead of implementing an endogenously determined audit probability, e.g. by modeling the tax authority as an additional player. For a theoretical analysis of endogenous audit probabilities see Landsberger and Meilijson (1982); Raymond (1999) and for an experimental treatment see Alm et al.
taxes, the incidence of an audit leads to an imposed penalty for the tax payer, but not for the tax officer. Second, we then modify the basic bribery game by adding an additional stage in the spirit of a principal witness regulation. This extended bribery game follows the same rules except that now an underreporting tax payer can avoid the penalty he would face upon detection by denouncing the corrupt tax officer, who in that case must pay a fine instead. The idea of the principal witness regulation was to render the tax officer also formally responsible because a corrupt tax officer now faces the threat of a fine as well.

The goals of this study are twofold: First, we seek to analyze tax compliance and collusive bribery under different regimes where corruption is feasible due to the interaction between tax payer and tax officer. Second, we investigate the effectiveness of a principal witness regulation as a mechanism to increase tax compliance and reduce corruption in settings with and without institutional transitions.

One way to think of the introduction of a principal witness regulation is as a stylized situation where tax authorities decide to invest in establishing control mechanisms that allow for better monitoring of public officials. Hence allowing them to enforce legal consequences not only on tax payers but also on corrupt tax officers, for example via improved monitoring. We mimic transitions of that type by employing not only static treatments, where exactly one regime is present for the whole duration of the experiment, but also dynamic treatments involving a regime change from one to the other. This allows us to study both the effectiveness of either setup in isolation and how subjects react to a change in either direction. In particular, we are interested in whether the transition from one scenario to the other can break collusive behavior established during a previous period were the other institutional frame was implemented, and hence can serve as a tool to reduce tax

3 The basic bribery game can be considered as a game with an asymmetric punishment structure. We refer the reader to Engel et al. (2013) for a detailed analysis of symmetric vs asymmetric punishment schemes. See also Abbink et al. (2014).
evasion in a world where that frame represents the status quo.

Our experimental results highlight the influence of institutional history on the effectiveness of changes in the institutional and economic environments, which is in line with Acemoglu et al. (2005). In particular, our findings highlight two distinct and surprising points: Firstly, when resorting to a between-comparison of institutional environments, that is a world in which a principal witness regulation has always been absent with one in which it has been implemented from the get go, tax compliance is significantly higher in the latter. This effect is mainly driven by a higher tax compliance rate of females and is thus in line with existing literature on gender differences in illicit settings such as corruption or cheating behavior (cf. Croson and Gneezy, 2009; Torgler and Valev, 2010). In contrast, when shedding light on the impact of an institutional shock, that is, a within-comparison of institutional changes, our results paint a substantially different picture. In particular, our results indicate that the introduction of a principal witness regulation into an environment where it has previously been absent leads to significantly less tax compliance. Much to our surprise and both against conventional wisdom and existing literature, we find that these welfare distortions are largely driven by the significantly increased non-compliance of females. In sum, our results provide robust evidence for the effectiveness of principal witness regulations across various institutional settings, which render important policy implications. Our results also add to the growing body of evidence on gender differences within the frame of choice under risk and strategic uncertainty, and provide further evidence to the idea that females are generally more sensitive to the contextual frame (Croson and Gneezy, 2009; Dreber et al., 2013).

The paper is organized as follows: Section 2 describes the experimental design. Section 3 presents the analysis of our empirical results. In Section 4 we discuss our results and related literature. Section 5 concludes.
2. Experimental Design

Both of our institutional setups mimic a scenario where tax evasion is nested within a corruption framework. Taxes are collected through an intermediary, the tax officer. Hence, to successfully evade taxes the tax payer requires the cooperation of the tax officer, for example by “looking the other way.” We now give a detailed description of the two institutional frames used in our experiment.

2.1. Two institutional frames: Bribery Game and Extended Bribery Game

The upper part of Figure 1 illustrates the bribery game (BG). A tax payer (TP) receives an income of 80 Experimental Currency Units (ECU) and has to submit a declaration of his income to the tax authorities. The tax officer (TO), acting as an intermediary, is in charge of processing the tax report. Declared income \( D \) is subject to a tax rate of 50%.\(^4\) The TP can either truthfully declare his full income of 80 or he can evade taxes by declaring a lower income \( D < 80 \). In addition, the TP can try to convince the TO to collude with him in his effort to evade taxes by offering a bribe \( b \) that can range from 0 to 30 ECU. The situation we have in mind is one, where the TP can vastly increase the chance of his false tax declaration not being detected by colluding with the TO, who is in charge of processing the report. For simplicity, we assume that it is impossible for the TP to evade taxes without the TO's support. That is, declaring less than the full income is only possible if the TO accepts the TP’s bribe offer and hereby agrees to collude with the TP, e.g. by manipulating the report.\(^5\) The TO receives the TP’s bribe offer requesting the TO to collude with the TP in evading taxes. She then decides whether to accept or reject the bribe offer. It is important to note that the TO cannot observe the amount of taxes the TP

\(^4\)Subjects were informed that this tax rate is in line, according to a recent study of Confcommercio, with the mean tax burden in Italy.

\(^5\)It is natural to assume that the detection probability of a false report when colluding with the TO is lower than without the TOs help. Our simplification is equivalent to normalizing the latter probability to zero.
wants to evade and hence cannot condition her decision on the amount of taxes evaded.\textsuperscript{6} If the TO rejects a bribe, then she refuses to collude with the TP, which forces the TP to truthfully declare his full income of 80. If the TO accepts the bribe, then the TP is able (with the help of the TO) to file the original (potentially false) tax report.\textsuperscript{7} After the tax report has been submitted the tax report will be audited by the tax authorities with an exogenous probability of 20%. In case of an audit incorrect reports are detected and the TP has to pay both the evaded amount of taxes $0.5(80 - D)$ and an additional fine.

\textsuperscript{6}A reason for this is, for example, that the TO does not know the actual income of the TP, which is only known to the official tax authority conducting the audits.

\textsuperscript{7}Note that in contrast to Abbink and We\textsuperscript{[2017]} we do not allow the tax officer to pocket the bribe without delivering the corrupt favor of colluding with the tax payer.
proportional to the amount of evaded taxes. The fine is set to 25% evaded taxes, hence the maximum fine is 12.5 ECU. Thus, the TP’s payoff is his income minus taxes on the declared income $D$ and potentially the bribe and/or fine paid. The TO’s payoff consists of three components: a fixed wage of 50, a commission of 15% of the taxes collected, and the amount of bribes accepted.

The extended bribery game (EG) is very similar to the bribery game just described, but with one important difference. In the EG we add an additional stage to the BG intended to mimic a situation that a leniency program for blowing the whistle. Decisions in EG are identical to those in BG, however, following detection of an incorrect tax report through an audit the TP now has the opportunity to “blow the whistle” by denouncing the TO. If the TP chooses to denounce, he has to correct the (false) tax report, that is to truthfully declare taxes, but does not incur an additional monetary punishment, that is the fine is waived. A TO that has been denounced, on the other hand, incurs a fine for colluding with the TP to evade taxes. This fine equals the bribe received from the TP plus an additional penalty of 10 ECU. The fine rate of 25% was chosen such that together with the upper bound (of 30) on bribes the TP can never incur a net loss.

2.2. Theoretical Analysis

Consider the bribery game and the extended bribery game described above as one-shot interaction between a TP and a TO, both assumed to be rational in the sense of being risk-neutral expected payoff-maximizers. Assuming the rational model of crime (Allingham and Sandmo, 1972) we now derive theoretical predictions regarding tax compliance and bribe

---

8This is an institutional feature often observed in developed countries (Mittone, 2006).

9The introduction of a commission for the TO mimics something existing in reality. In Italy for example the Tax Authority delegates inspections and audits to a private Organization (Equitalia) and pays Equitalia with a percentage of the money collected.

10In the BG punishment can be viewed as asymmetric as only tax payers are running the risk of being fined, however, in EG leniency shifts, at least partially, this risk towards the tax officer, hence creating a situation that might be perceived as more symmetric. See also Engel et al. (2013) for a discussion of symmetric vs asymmetric punishment regimes.
exchange. Our analysis shows that predicted tax compliance of the TP is the same for both institutional frames. On the other hand, the optimal bribe payment is higher in the EG where the principal witness regulation is present. Moreover, bribe exchange (collusion) is optimal under both regimes. Denote the amount of taxes declared by $D$ and the bribe offered by $b$.

In the BG a rational TO will accept any bribe $b$ that is (weakly) above the expected foregone commission of 15% from the taxes declared, that is 7.5% of $D$. Since the TO does not observe the taxes declared by the TP we assume that she holds a belief $\mu : \{0, \ldots, 80\} \rightarrow [0, 1]$ over $D$. The expected amount of declared taxes given this belief $\mu$ is then $D(\mu) = \sum \mu(D)D$. Hence, the TO will accept a bribe if she believes that the bribe is larger than her foregone commission, that is if and only if

$$b \geq 6 - 0.075D(\mu).$$

The bribe acceptance threshold, which we denote by $b_{BG}(\mu)$, depends only on the TO’s mean belief $D(\mu)$. For example, if the TO expects the TP to declare zero taxes, that is $D(\mu) = 0$, then only bribes of at least 6 are accepted, whereas if she expects the TP to declare half, that is $D(\mu) = 40$, all bribes above 3 are accepted. Note that the threshold is strictly increasing in the mean belief $D(\mu)$. On the other hand, if the TP offers a bribe $b$ and the TO accepts (which is the case for $b \geq b_{BG}(\mu)$), the TP’s expected payoff for reporting an amount of $D$ is

$$\Pi_{TP}(D, b \mid accept) = 55 - b - 0.34375D$$

Note that $\Pi_{TP}$ is decreasing in $D$ and $b$, hence a rational TP will optimally declare an income of $D = 0$ and pay the smallest bribe that is accepted by the TO, which is $b = 6 - 0.075D(\mu)$.

In the EG leniency introduces the possibility for a TP to denounce a corrupted TO
following an audit. In the one-shot scenario it is optimal for the TP to denounce the TO when being audited, in which case the TP now has an expected payoff of

$$\Pi_{TP}(D, b \mid \text{accept, denounce}) = 70 - b - 0.375D$$

This payoff is still decreasing in $D$ and $b$, and thus the TP prefers to declare zero taxes and pay the smallest bribe that is accepted by the TO. However, the bribe threshold in the EG is not the same as in the BG. To see this, suppose the TO anticipates that the TP will always denounce her when audited, then a rational TO will accept a bribe if and only if

$$10 - 0.09375D(\mu) =: b_{EG}(\mu).$$

Intuitively, now the TP has to compensate the TO not only for his forfeited (expected) salary, but also for the risk of being denounced and its consequences.

The game described above is a game of imperfect information (the TO does not observe $D$) and as such it has many Nash equilibria. We use Perfect Bayesian Equilibrium (PBE) as our solution concept of choice. Given a point belief $\mu$ with $\mu(D) = 1$ for some $D \in \{0, \ldots, 80\}$ there is a unique PBE of BG where the TP declares exactly $D$ and offers a bribe $b = b_{BG}(\mu)$ and the TO accepts this bribe. Similarly, for a point belief $\mu$ with $\mu(D) = 1$ there is a unique PBE for EG where the TP declares $D = 0$, offers a bribe $b = b_{EG}(\mu)$ and always denounces the TO when audited, and the TO accepts the bribe $b$. In both, BG and EG, collusion is an equilibrium of the one-shot game. However, since $b_{EG}(\mu) > b_{BG}(\mu)$ for any $\mu$, the bribe acceptance threshold in EG is higher compared to BG. It is important to note that for both games the bribe acceptance threshold is decreasing in the mean of the TO’s belief $\mu$.

From a purely theoretical perspective the predictions for the two games differ only in so far as they predict different acceptance thresholds leading to higher equilibrium bribes in EG. However, this prediction is based on completely rational (in the sense of PBE)
and purely self-interested behavior. In our experiment, we use a loaded framing referring explicitly to tax evasion and corruption, which is likely to trigger moral concerns and strategic considerations that go beyond material gains (Banerjee, 2016). In other words, we are interested whether a principal witness regulation, as considered here, is a viable mechanism to increase tax compliance and deter corruption although it doesn’t provide strong monetary incentives to refrain from collusive bribery. However, we expect that, especially in a repeated environment, leniency can sow mistrust and increase uncertainty between both parties.

2.3. Treatments

In our experiment subjects repeatedly played the bribery game and/or its extended version (with leniency) over the course of a total of 20 rounds. The experiment consists of four treatments. In Treatment 1, participants play the bribery game without leniency for 20 rounds. In Treatment 2 subjects play the extended bribery game instead, also for 20 rounds. These two treatments allow a between-subject comparison of the role that leniency plays with respect to collusive bribery and tax compliance. In addition, these treatments represent a benchmark for Treatment 3 and 4, in which institutional shocks occur. These treatments were designed to study the effects of institutional transitions, since in those treatments institutional changes occur unannounced midway through the experiment after round 10. In particular, in Treatment 3 (noPwr-Pwr) subjects start with the basic bribery game and are then transitioned into an environment in which denouncing the tax officer becomes feasible. Treatment 4 (Pwr-noPwr) captures the same dynamic but in reverse order, that is, first the option to denounce is available and is then abolished after round 10. These two treatments involve a regime change that allows us to analyze the effectiveness of both the introduction and the removal of leniency, in the form of a PWR, relative to a “status quo,” that is, the regime present during the first block of 10 rounds. Table 1 summarizes the four treatments.
Table 1: Overview over the treatments and number of subjects assigned to each treatment.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Round 1-10</th>
<th>Round 11-20</th>
<th>$N_{TO}$</th>
<th>$N_{TP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1 (noPwr)</td>
<td>BG</td>
<td>BG</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>Treatment 2 (Pwr)</td>
<td>EG</td>
<td>EG</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>Treatment 3 (noP-Pwr)</td>
<td>BG</td>
<td>EG</td>
<td>22</td>
<td>66</td>
</tr>
<tr>
<td>Treatment 4 (Pwr-noPwr)</td>
<td>EG</td>
<td>BG</td>
<td>14</td>
<td>42</td>
</tr>
</tbody>
</table>

Note: The principal witness regulation is present in the extended bribery game (EG) but not in the bribery game (BG). $N_{TO}$ and $N_{TP}$ denotes the number of tax officers and tax payers, respectively.

2.4. Behavioral Predictions and Hypotheses

In Section 2.2 we derived the theoretical predictions for one-shot bribery game and the one-shot extended bribery game. Assuming purely self-interested behavior collusion, that is evading taxes and bribing the TO, is optimal for TPs with and without leniency. However, in the extended bribery game the TP always denounces the TO in equilibrium, resulting in a higher bribe acceptance threshold on the side of the TO, which then makes higher bribe offers optimal. Regarding tax compliance there are many equilibria involving different levels of tax compliance, for example there is one equilibrium where the TP declares zero taxes (and the TO correctly anticipates this behavior).

Experimental evidence suggests that subjects’ tax compliance usually is well above the theoretically optimal level, for example due to moral costs of engaging in tax evasion or corruption\footnote{An important feature of our experiment is that tax evasion is nested within a corruption framework that requires collusive behavior for tax evasion to be successful. We believe that this additional layer of interaction is important to help us better understand unethical behavior in situations in which cooperation is necessary. This interaction possibly increases the impact of behavioral factors such as psychological costs and uncertainty on tax compliance and the willingness to engage in collusive bribery. Our main objective is to study the effectiveness of a leniency mechanism as a means to...\footnote{It was shown in \cite{Banerjee2016} that a loaded frame that creates the right sense of entitlement significantly decreases corruption, suggesting that moral costs are indeed at work.}}. An important feature of our experiment is that tax evasion is nested within a corruption framework that requires collusive behavior for tax evasion to be successful. We believe that this additional layer of interaction is important to help us better understand unethical behavior in situations in which cooperation is necessary. This interaction possibly increases the impact of behavioral factors such as psychological costs and uncertainty on tax compliance and the willingness to engage in collusive bribery.

Our main objective is to study the effectiveness of a leniency mechanism as a means to...
increase tax compliance and hinder collusive corruption. The presence of a principal witness regulation effectively reduces the risk the TP faces when evading taxes, while shifting responsibility to the TO and potentially reducing the TP’s psychological costs. Intuitively, in the extended bribery game the possibility to denounce the TO offers the TP a “safe way out”. Thus leniency effectively allows the TP to avoid an additional fine when evading taxes, and if the fine is what is keeping a TP from evading one would expect compliance to decrease. On the other hand, leniency also affects the chances that an attempt to evade taxes (as captured by the amount of taxes declared) is successful. Our design allows for discernment between attempted non-compliance (taxes declared) and actual compliance (taxes paid). Recall that in EG not only the TP faces less risk but it is effectively shifted to the TO as leniency exposes the TO to the possibility of being denounced and fined. This raises the TO’s optimal bribe acceptance threshold. In order to sustain collusion, that is, to ensure the TO’s continued cooperation, the TP has to compensate the TO for this additional risk with higher bribe payments. If a TP fails to acknowledge this increased risk on the TO’s side, collusion fails, resulting in more rejections of bribes by the TO. A TP whose evasion attempts were repeatedly rejected can either react by increasing bribe offers or give up on evasion and declare taxes truthfully. We thus expect higher bribe payments, a higher rejection rate, and overall less collusion in Treatment 2 compared to Treatment 1.

Treatment 3 allows us to study the effect of the introduction of a principal witness regulation into a setting in which corrupt behavior has already been able to thrive in the absence of leniency. We are interested in whether a sudden change of the institutional environment is effective in reducing corruption, for example by breaking up collusive patterns. Introducing leniency might affect the shadow price of corrupt coordination, and as a consequence collusion might fail.

We conjecture that the institutional break is able to distort established collusive relationships and this should be reflected in an increased rate of miscoordination in the second block of Treatment 3. The introduction of a PWR increases the bribe acceptance threshold
of the TO, hence narrows the range of bribe offers that are likely to lead to successful collusion. In addition, a TO might simply dislike or even fear being denounced and because of this reject even more bribes, which will result in higher rejection rates on the side of the TO and thus less collusion overall under in Treatment 3 under leniency.

In Treatment 4 the logic is reversed. Subjects start under a regime with leniency followed by its removal. If collusion is harder to achieve in the presence of leniency, one would expect low acceptance rates and little collusion (exchanged bribes) in the first part when facing the extended bribery game. On the other hand, as the removal of the PWR supposedly facilitates collusion it should lead to higher acceptance rates and a frequent exchange of bribes in rounds 11 to 20.

The optimal bribe payments in equilibrium are higher in the extended bribery game than in the bribery game because the TO’s bribe acceptance threshold is higher. We thus expect the TO to reject higher bribes when facing the extended bribery game, which likely leads to a higher rejection rate, especially in Treatment 2 when TP’s fail to increase their bribe offers accordingly. Intuitively the TO has to be compensated for the additional risk of being denounced and its consequences. When the TPs choose their bribe offers they essentially face an Ultimatum Game-like situation, more precisely we expect that they will offer significantly more than the optimal bribe and that the TO will not accept too small bribes. Notice that the situation is more complex than in a standard UG because here the proposer (the TP) is constrained by a duty and the receiver (the TO) occupies a hierarchical position which should be perceived as higher than the position occupied by the receiver of the standard UG. This is something extraneous to the standard UG setting therefore we do expect that the average offer (bribe) would be higher than the standard average offers in the UG literature.

In Section 2.2 we have seen that the TO’s bribe acceptance threshold depends on his belief about the amount of taxes declared by the TP. Although it is not the main objectives of our study, we think it might be of some interest to see whether TO’s beliefs affect their
acceptance behavior and or the bribe acceptance threshold. To that end, whenever a TO was offered a bribe, we elicited the TO’s beliefs regarding the amount of taxes declared by that TP after she chose to accept or reject that bribe. If beliefs matter, we should observe a correlation of rejection of high bribes and TO’s belief about the income declared being low.

2.5. Experimental procedures

Subjects were randomly assigned either the role of a TP or the role a TO. Participants were randomly matched in groups of 4 consisting of one TO and 3 TPs, that is, a TO was assigned 3 TPs to interact with simultaneously. There was no direct interaction between different TPs in the same group. In each session groups remained fixed throughout the experiment consisting of a total of 20 periods split into two phases of 10 rounds. Subjects were informed that the number of rounds was predetermined, but were not informed about the exact number of rounds. In each period subjects played, depending on the treatment, either the bribery game or the extended bribery game. For Treatment 1 (noPwr) and Treatment 2 (Pwr) no institutional change occurred. In Treatment 3 (noPwr-Pwr) and Treatment 4 (Pwr-noPwr) the participants were informed about a change in the institutional setting after the 10th round via an announcement on screen that provided a detailed description of the new institutional environment. We emphasized that there would be no additional change of the institution until the end of the experiment. Subjects were informed in the instructions that the existing institution may be subject to change but no information regarding the nature of the change was provided. Thus, we use a within variation of the institutional setting to study the effect of leniency on tax compliance. We also elicited the TO’s beliefs about the amount of taxes evaded by each of the TPs offering

---

12. We choose not to announce the number of rounds to avoid potential end-game effects. 13. We provided subjects in all treatments with identical information at the start of the experiment. In particular, participants assigned to Treatment 1 and 2 were informed about the possibility of a change although, ultimately, they would not experience one.
a bribe. Beliefs were elicited in each round after the TO’s decision to accept or reject a bribe offer. Belief elicitation was incentivized.

To make tax evasion more salient in the laboratory setting, we introduce a third party that incurs a monetary damage as a result of tax evasion. All participants were informed that the total tax yield collected would be used to finance future research of doctoral students at the University of Trento. That is, tax evasion in the experimental laboratory translates into an actual social welfare loss outside the lab (see Eckel and Grossman, 1996; Lambsdorff and Frank, 2010).

The experiment was conducted with a total of 268 undergraduate students (46% females) at the Cognitive and Experimental Economics Laboratory at the University of Trento. Table 1 shows the distribution of subjects over the four experimental treatments. Sessions consisted of 20 rounds followed by an incentivized risk-elicitation task (Holt and Laury, 2002) and a demographic questionnaire. The final payoff of each subject was determined as the sum of all earnings over the 20 rounds plus their earnings from the risk-elicitation task, which were then converted to Euro at a rate of 100 ECU = €0.7. All participants were paid their final payoff plus an additional show-up fee of €3 in cash at the end of the experiment. On average, a session lasted about 60 minutes and subjects earned €12 excluding the show-up fee of €3.

3. Results

We structure our analysis in the following way: first, and in line with our overall motivation, we will discuss the effectiveness of leniency in affecting collusive agreements between public officials and tax payers. In a next step, we will break down the behavior of tax payers and public officials individually. Lastly, we discuss our surprising gender results

---

14 This is a common procedure in tax evasion experiments in order to link tax evasion to a negative externality, for example see Fortin et al. (2007); Coricelli et al. (2010).
and provide potential mechanisms that could help to explain it. We employ a very cautious
approach in our data analysis. Following our design, we regard the behavior of one group
(consisting of one public official and three tax payers) averaged over all rounds as one
independent observation. We follow Moffat (2015) and use the bootstrap two-sample t-test
method (hereafter BSM) for paired and unpaired samples with 9999 replications to analyze
mean differences in behavior. This has the advantage that we can retain the rich cardinal
information in the data without making any assumptions about the distribution. Unless
noted otherwise, the use of non-parametric Mann-Whitney-U Tests (hereafter MWU) or
Wilcoxon Signed Rank Tests (hereafter WSR), respectively, yield results that are in line
with the bootstrap approach.

3.1. Collusive Behavior

In line with our principle interest to study the effectiveness of leniency on collusive
arrangements, our experimental design allows us to approach this question from two per-
spectives:

1. Is collusion generally different in environments in which leniency exists?
2. How does an institutional change from an environment with (without) leniency to an
   environment without (with) leniency affect collusive behavior?

To address these two questions, we first compare the average rates of collusion across
treatments. We define collusion as the successful exchange of bribes in return for the avoid-
ance of taxes. Our variable is calculated as the proportion of successful illicit agreements
relative to all rounds in which evading taxes and paying a bribe could take place. With
respect to the first question, as illustrated in Figure 2, our results indicate that collusion
is less frequent when leniency is present, with the differences being highly significant (0.52

\[ \text{An alternative way of defining collusion would be the ratio of successful attempts relative to all attempts. That is, successful bribery conditional on having paid a bribe. Our results are robust to this alternative definition. See Figure B.1 in Appendix B for details.} \]
Interestingly, regarding our second question, our results also indicate that the effectiveness of leniency in distorting collusive relationships is mediated by the history of the institutional setting. In fact, the introduction of leniency, as it is the case in Treatment 3, does not yield any change in collusive agreements (0.43 vs. 0.42, BSM, \( p = 0.61 \)). In contrast, removing leniency, as it is the case in Treatment 4, leads to a highly significant rise in successful collusion (0.38 vs. 0.49, BSM, \( p < 0.01 \)).

Figure 2

These results suggest that the presence of a leniency mechanism indeed deters collusion as the low rate of collusion in Treatment 2 compared to Treatment 1 indicates. However, the introduction of leniency was not able to decrease the level of collusion established in earlier rounds in the absence of such a mechanism (Treatment 3). We conclude that although collusion is less frequent under leniency it is not able to break up already established collusive patterns. Conversely, following the removal of leniency in Treatment 4 there is a jump in collusion indicating that a low rate of collusion established under leniency cannot
be sustained after its removal.\footnote{It is important to note that at this point we do not account for any learning that might take place over the course of multiple rounds. To account for that we also look at the rate of successful collusive agreements over time. Across all periods, our results suggest that successful collusion increases over time, although at different speeds. See Figure B.2 in the Appendix for a graphical illustration. One of three reasons can cause this: for one, the tax payers reduce their bribery attempts over time \[\text{[wenn ich die Grafik mit der alternativen Definition von Collusion mache, sieht es fast genau so aus, also ist diese Antwort eher auszuschlieen und die Grafik zum Kommentar 21. und 22. zeigt deutlich, dass die bribe attempts quasi flat sind]}\]. Alternatively, the tax payers increase the size of bribes offered. Lastly, the public officials reduce their threshold of bribe amount accepted. We will analyze this in more detail in later sections.}

3.2. Decomposed Behavior of Tax Payers and Public Officials

In a next step, we try to understand the drivers of collusion by shedding light separately on the behavior of tax payers and public officials. In particular, we will be looking at bribe payments (average amount of bribes offered conditional on offering a bribe), tax declarations (average amount of taxes declared), frequency of attempted bribe offers (binary variable with 1=bribe offered), and the minimal acceptance threshold “minbribeaccept” (lowest accepted bribe by a TO in that round conditional on having been offered at least 1 bribe by the 3 TPs).

In our setting, a tax payer’s decision on whether to evade taxes and, if so, which amount, goes hand in glove with the decision to pay a bribe and make the tax officer look the other way. Ceteris paribus, higher bribe payments should naturally lead to higher collusion rates.\footnote{As expected, an examination of bribe offers and acceptance decisions across rounds yields a positive relationship. See Figure B.3 in the Appendix.} Hence, in order to understand the drivers of the results previously discussed in Figure 2, we now decompose the tax payers behavior by the amount of taxes declared and bribes paid, respectively in Figures 3, 4, and 5.\footnote{Consistent with our previous approach, we again calculate and test behavior on the group level averaged over rounds.}

A comparison of Treatment 1 and 2 reveals that, in line with our theoretical predictions, bribe payments are significantly higher (48.2\% vs. 55.3\%, BSM, \(p = 0.07\)) and tax declarations are significantly lower (34.7\% vs. 51.4\%, BSM, \(p < 0.01\)) in an institutional environment in which leniency exists. This goes hand in hand with a significantly lower...
frequency of attempted bribe offers (67.7% vs. 55.1%, BSM, $p = 0.03$). Our results further indicate that the introduction of leniency (Treatment 3) results in significantly larger bribes (44.1% vs. 51.3%, BSM, $p = 0.06$) and lower taxes declared (50.4% vs. 37.2%, BSM, $p = 0.02$), whereas the removal of leniency (Treatment 4) does not trigger any significant change of either. Interestingly, in both conditions, the frequency of bribe offers remained invariant to the addition or removal of leniency. Overall, these results suggest that it is not the mere institutional shock that causes significant changes in compliance behavior, but rather the order in which the institutional transition occurs.

As predicted TPs frequently used the option to denounce the TO in order to avoid paying a fine. On average TP’s chose to denounce in 91.4% of all possible cases with no significant difference across treatments. They preferred to go free at the expense of the TO although this might negatively influence collusive cooperation in the future.

In addition, we observe that the minimal bribe acceptance threshold of the tax officers are significantly higher ($p < 0.01$, WSR) when leniency is in place (Treatment 2) as compared when it is not (Treatment 1). This is in line with our theoretical prediction.
that tax officers optimally increase their acceptance threshold from 6 in the bribery game (Treatment 1) to 10 in extended bribery game (Treatment 2). It seems like taxpayers acknowledge the higher risk that public officials have to bear in the presence of a PWR and, at least partially, compensate them with higher bribes.
3.3. Regression Analysis

To investigate the behavior of tax payers and tax officers in more detail we ran a series of logistic panel regressions.

Table 2 compares collusion in Treatment 1 and Treatment 2. We see overall less collusion in Treatment 2 (Pwr) compared to Treatment 1 (noPwr). Male subjects collude more often. Moreover, there is a general learning effect over the course of the experiment, with more collusion in later periods as subjects gain more experience. We find no effect of risk attitudes on collusion behavior. 19

Treatment 3 (NoPwr-Pwr). Table 3 shows the frequency of collusion in Treatment 3. In model 1 there appears to be no effect of leniency the introduction of PWR on collusion. We observe that male subjects collude more frequently than their female counterparts. To control for potential learning effects over the course of the experiment we include Round as

---

19 For tax payers, there were no significant differences in risk attitudes across treatments. For tax officers risk attitudes are marginally different between Treatment 1 and 2 ($p = 0.077$, @Eugen: What test?), but no difference for the other Treatments. We are fairly confident that risk attitudes do not play a major role in explaining our results. Indeed, in the regressions below risk is rarely significant.
Table 2: Treatment 1 vs Treatment 2. Logistic panel regression with random effects and standard errors clustered on the group level.

<table>
<thead>
<tr>
<th>Collusion</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2 (Pwr)</td>
<td>-1.1143***</td>
<td>-0.9028**</td>
<td>-0.9285**</td>
<td>-0.9011**</td>
</tr>
<tr>
<td></td>
<td>(0.4277)</td>
<td>(0.4136)</td>
<td>(0.4278)</td>
<td>(0.4236)</td>
</tr>
<tr>
<td>Male</td>
<td>1.0235***</td>
<td>1.0515***</td>
<td>1.7566</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.3289)</td>
<td>(0.3416)</td>
<td>(1.3005)</td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>-0.0682</td>
<td>-0.0702</td>
<td>-0.0045</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0943)</td>
<td>(0.0974)</td>
<td>(0.1548)</td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td>0.0652***</td>
<td>0.0652***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0173)</td>
<td>(0.0173)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male_Risk</td>
<td></td>
<td>-0.0993</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.1950)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.0647</td>
<td>-0.1605</td>
<td>-0.8478</td>
<td>-1.3300</td>
</tr>
<tr>
<td></td>
<td>(0.2430)</td>
<td>(0.7071)</td>
<td>(0.7173)</td>
<td>(1.0514)</td>
</tr>
<tr>
<td>Observations</td>
<td>1320</td>
<td>1320</td>
<td>1320</td>
<td>1320</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

an additional control (model 2). Indeed, the coefficient on Round is positive and significant, indicating that there is more collusion in later rounds of the experiment. Interestingly, after controlling the increase in collusion due to learning, we find a significant negative effect of the introduction of PWR. We now include the interaction between Male and PWR (model 3) to see whether males and females react differently to the introduction of PWR. The interaction is highly significant, that is, male participants collude less after PWR is introduced, while females show no change in behavior. These results are unaffected if we allow for different learning effects before and after the introduction of PWR (model 4).

Collusion requires the cooperation of both, the tax payer and the tax officer. In order to pin down whether the effects on collusion are mainly driven by the tax payer or the tax officer. We first consider the tax payer. Table 4 shows the results of a logistic panel regression with BribeInc, denoting whether a bribe was offered or not, as a dependent variable for Treatment 3. Overall bribe offers are more frequent in the presence of leniency,
however, this joint effect is a result of two counterveiling effects that differ between gender, females offer bribes significantly more often under leniency, whereas the number of incidences where a bribe is offered by a male participant is lower. This can partially explain why we find a decrease of collusion in Treatment 3 under leniency for males, but not for females.

We now consider how the TO’s behavior is affected by leniency. Table 5 shows regression results of the bribe acceptance rate (BribeAccRate), that is the fraction of bribe offers received by the TO that she accepted in a given round. The TO accepts significantly less bribes when leniency is in place, which may be a result of both, the higher number of bribe offers as well as the potential risk of incurring a fine when being denounced. That is, the TO rejects more bribe offers after leniency is introduced, which might explain why we do not observe an increase in collusion between TOs and female TPs although these
Table 4: Treatment 3. Logistic panel regression with random effects and standard errors clustered on the group level.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>BribInc</td>
<td>0.6036***</td>
<td>1.1104***</td>
<td>1.5914***</td>
<td>2.9736***</td>
</tr>
<tr>
<td></td>
<td>(0.1721)</td>
<td>(0.2519)</td>
<td>(0.3426)</td>
<td>(0.6888)</td>
</tr>
<tr>
<td>PWR</td>
<td>-0.4665</td>
<td>-0.4681</td>
<td>-0.0815</td>
<td>-0.0751</td>
</tr>
<tr>
<td></td>
<td>(0.4995)</td>
<td>(0.5013)</td>
<td>(0.5344)</td>
<td>(0.5359)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.1456</td>
<td>-0.1462</td>
<td>-0.1485</td>
<td>-0.1497</td>
</tr>
<tr>
<td></td>
<td>(0.1131)</td>
<td>(0.1134)</td>
<td>(0.1171)</td>
<td>(0.1180)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.0504***</td>
<td>-0.0508***</td>
<td>0.0080</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0186)</td>
<td>(0.0188)</td>
<td>(0.0289)</td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td>-0.8382**</td>
<td>-0.8545**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.3674)</td>
<td>(0.3717)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male_PWR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round_PWR</td>
<td>-0.1258***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.0480)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.6559**</td>
<td>1.9394**</td>
<td>1.7485**</td>
<td>1.4283*</td>
</tr>
<tr>
<td></td>
<td>(0.8116)</td>
<td>(0.7807)</td>
<td>(0.7731)</td>
<td>(0.7549)</td>
</tr>
<tr>
<td>Observations</td>
<td>1320</td>
<td>1320</td>
<td>1320</td>
<td>1320</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

offer bribes more frequently. Interestingly, we also see some evidence for an increase in successful collusive cooperation over time in the absence of leniency, while under leniency we see no such effect. This is in line with the idea that leniency indeed makes it more difficult coordinate on a collusive agreement, that is honored by both parties.
Table 5: Treatment 3. GLS panel regression with random effects and standard errors clustered on the group level.

<table>
<thead>
<tr>
<th>BribeAccRate</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWR</td>
<td>-0.1006*</td>
<td>-0.2821***</td>
<td>-0.3546**</td>
</tr>
<tr>
<td></td>
<td>(0.0547)</td>
<td>(0.0821)</td>
<td>(0.1453)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.0043</td>
<td>-0.0027</td>
<td>-0.0028</td>
</tr>
<tr>
<td></td>
<td>(0.0384)</td>
<td>(0.0372)</td>
<td>(0.0374)</td>
</tr>
<tr>
<td>Round</td>
<td>0.0181***</td>
<td>0.0146*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0058)</td>
<td>(0.0082)</td>
<td></td>
</tr>
<tr>
<td>Round,PWR</td>
<td></td>
<td>0.0070</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0115)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.7637***</td>
<td>0.6547***</td>
<td>0.6753***</td>
</tr>
<tr>
<td></td>
<td>(0.2469)</td>
<td>(0.2412)</td>
<td>(0.2497)</td>
</tr>
<tr>
<td>Observations</td>
<td>421</td>
<td>421</td>
<td>421</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

Treatment 4 (PWR-NoPWR). *** To be done! (JB)
Table 6: Treatment 4. Logistic panel regression with random effects and standard errors clustered on the group level.

<table>
<thead>
<tr>
<th>Collusion</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWR</td>
<td>-0.6295**</td>
<td>0.2357</td>
<td>0.3383</td>
<td>0.2834</td>
</tr>
<tr>
<td></td>
<td>(0.2540)</td>
<td>(0.3714)</td>
<td>(0.3382)</td>
<td>(0.9458)</td>
</tr>
<tr>
<td>Male</td>
<td>0.2854</td>
<td>0.2890</td>
<td>0.3760</td>
<td>0.3759</td>
</tr>
<tr>
<td></td>
<td>(0.4461)</td>
<td>(0.4529)</td>
<td>(0.4599)</td>
<td>(0.4591)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.1485</td>
<td>-0.1504</td>
<td>-0.1509</td>
<td>-0.1509</td>
</tr>
<tr>
<td></td>
<td>(0.0954)</td>
<td>(0.0966)</td>
<td>(0.0970)</td>
<td>(0.0969)</td>
</tr>
<tr>
<td>Round</td>
<td>0.0874***</td>
<td>0.0874***</td>
<td>0.0848*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0235)</td>
<td>(0.0235)</td>
<td>(0.0508)</td>
<td></td>
</tr>
<tr>
<td>Male_PWR</td>
<td>-0.1809</td>
<td>-0.1806</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.2717)</td>
<td>(0.2729)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round_PWR</td>
<td>0.0052</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0864)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.6005</td>
<td>-0.7459</td>
<td>-0.7921</td>
<td>-0.7520</td>
</tr>
<tr>
<td></td>
<td>(0.5966)</td>
<td>(0.7323)</td>
<td>(0.7144)</td>
<td>(1.0021)</td>
</tr>
<tr>
<td>Observations</td>
<td>840</td>
<td>840</td>
<td>840</td>
<td>840</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01
Table 7: Treatment 4. Logistic panel regression with random effects and standard errors clustered on the group level.

<table>
<thead>
<tr>
<th>BribeInc</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWR</td>
<td>0.5431**</td>
<td>1.0668***</td>
<td>1.4812***</td>
<td>2.1097***</td>
</tr>
<tr>
<td></td>
<td>(0.2187)</td>
<td>(0.3130)</td>
<td>(0.3689)</td>
<td>(0.6319)</td>
</tr>
<tr>
<td>Male</td>
<td>0.6178</td>
<td>0.6200</td>
<td>1.0124</td>
<td>1.0148</td>
</tr>
<tr>
<td></td>
<td>(0.6201)</td>
<td>(0.6230)</td>
<td>(0.6770)</td>
<td>(0.6789)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.0672</td>
<td>-0.0675</td>
<td>-0.0653</td>
<td>-0.0654</td>
</tr>
<tr>
<td></td>
<td>(0.1220)</td>
<td>(0.1225)</td>
<td>(0.1215)</td>
<td>(0.1216)</td>
</tr>
<tr>
<td>Round</td>
<td>0.0521**</td>
<td>0.0526**</td>
<td>0.0821*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0244)</td>
<td>(0.0245)</td>
<td>(0.0424)</td>
<td></td>
</tr>
<tr>
<td>Male_PWR</td>
<td>-0.8349**</td>
<td>-0.8371**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.4085)</td>
<td>(0.4095)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round_PWR</td>
<td>-0.0610</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0486)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>1.0646</td>
<td>0.2607</td>
<td>0.0492</td>
<td>-0.4055</td>
</tr>
<tr>
<td></td>
<td>(0.8124)</td>
<td>(0.9123)</td>
<td>(0.8760)</td>
<td>(1.0763)</td>
</tr>
</tbody>
</table>

Observations: 840 840 840 840

Standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

Table 8: Treatment 4. GLS panel regression with random effects and standard errors clustered on the group level.

<table>
<thead>
<tr>
<th>BribeAccRate</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWR</td>
<td>-0.2317***</td>
<td>-0.0677</td>
<td>-0.1827</td>
</tr>
<tr>
<td></td>
<td>(0.0693)</td>
<td>(0.1155)</td>
<td>(0.2481)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.0198</td>
<td>-0.0197</td>
<td>-0.0197</td>
</tr>
<tr>
<td></td>
<td>(0.0232)</td>
<td>(0.0233)</td>
<td>(0.0233)</td>
</tr>
<tr>
<td>Round</td>
<td>0.0165**</td>
<td>0.0109</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0068)</td>
<td>(0.0134)</td>
<td></td>
</tr>
<tr>
<td>Round_PWR</td>
<td></td>
<td>0.0108</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0201)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.8805***</td>
<td>0.6247***</td>
<td>0.7110***</td>
</tr>
<tr>
<td></td>
<td>(0.1727)</td>
<td>(0.1865)</td>
<td>(0.2621)</td>
</tr>
</tbody>
</table>

Observations: 267 267 267

Standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01
3.4. A Note on the Surprising Gender Effect

Note that there were no differences in gender composition across treatments for tax officers, and only a weakly significant difference for tax payers between Treatment 1 and 2 ($p = 0.064$, **@Eugen: What test?).

The review by Croson and Gneezy (2009) suggests that there is ample gender heterogeneity with respect to both risk taking in general and particularly engaging in risky unethical behavior within contexts of or similar to tax evasion. Existing research indicates that males have a tendency to be less risk-averse and engage in illicit behavior more often than females (cf. Torgler and Valev, 2010; Banuri and Eckel, 2012). However, it was not among the specific aims of this paper to investigate these well-established differences across gender with respect to the general propensity to evade taxes or engage in unethical behavior (as for example corruption). In particular, aside from these standard observations demonstrating a level difference between women and men, we did not expect a different reaction between males and females in relative behavior across the different institutional environments. However, our experiment yields surprising results with respect to how differently male and female participants respond to the introduction of a principal witness regulation. In this section, we will briefly discuss these surprising results regarding differences in relative behavior across gender. Notably, we find a strong heterogeneity in behavioral reactions to the introduction, but not the removal, of a principal witness regulation across gender. We want to emphasize that our experiment was not designed to study gender effects, in particular the fraction of females is not constant across treatments, ***Eugen, may be you can add here the male/female distribution and your tests for differences between treatments. and we had no male-only or female-only groups.

*** To be done: Discuss differences in collusion in Treatment 3 and 4 across gender! (JB)

*** Consider removing or shortening discussion on gender differences in taxes declared in favor of a discussion of collusion.
Figure 8 shows that the difference in tax compliance found in Section ?? is mainly driven by female participants, who show a significantly higher tax compliance in Treatment 2 compared to Treatment 1, where PWR was absent ($p < 0.01$, Wilcoxon rank-sum test). On the other hand, male participants barely change their tax compliance behavior as the small and insignificant increase in the amount of taxes declared from 30.9 in Treatment 1 to 34.7 in Treatment 2 demonstrates.20

We find that the observed effects regarding tax compliance are heavily driven by gender differences. When comparing gender behavior in treatments 1 and 2, that is where PWR was either present (Pwr) or not present (noPwr) over the full course of the experiment, we observe that tax compliance is significantly higher for females when PWR is present.

---

20We use robust random-effects GLS and Tobit estimations with clustered standard errors at the individual level to shed light on the individual drivers of tax compliance. The being controlled for tax compliance in the previous round leads to lower (higher) tax compliance when the principal witness regulation is absent (present) in the subsequent round. In addition, when PWR is in place, having to pay a fine in the previous round leads to higher tax compliance in the subsequent round. The estimations for all treatments are available from the authors upon request.
Figure 7: Mean collusion in Treatment 4 split by gender.

$(p < 0.01$, Wilcoxon rank-sum test). We do not observe significant changes in behavior for males. Results are presented in Figure 8.

Figure 9 illustrates the gender heterogeneity and the significant change of behavior on the side of females. We find that in Treatment 3 females evade significantly more taxes when PWR is introduced $(p < 0.01$, Wilcoxon rank-sum test), as opposed to males who do not exhibit any significant reaction to the institutional change $(p = 0.21$, Wilcoxon rank-sum test).
Figure 8: Difference in average tax compliance between a static framework without PWR (Treatment 1) and a static setup with PWR (Treatment 2) by gender.

Figure 9: Difference in taxes declared with and without PWR across treatments and by gender.
4. Discussion

To do: Update discussion (JB)

We provide empirical evidence that judging a political measure in isolation, thus disregarding the reference point provided by the pre-reform system, can lead to a flawed assessment of its effectiveness. It is therefore crucial to consider the whole history of political or legal systems in order to decide upon means to combat tax evasion and corruption. The classical economic model of tax evasion does not consider the fact that individuals are “born into” a certain legal system, but exactly this status quo might determine whether a potential reform is effective or not. Taking this evidence into account will be crucial in understanding why sometimes reforms are highly effective in a certain country or cultural environment, while they are ineffective in others. This might be related to the echo effect found in Mittone (2006), that is a change in the audit sequence affects behavior because subjects “learn” to be risk-averse or risk-seeking through experiencing early or late first audits. As a consequence past experience can create some sort of reference behavior that cannot easily be “unlearned,” and hence might hinder the effectiveness of a subsequent reform. Following that line of argument reforms can often be a one-way street, once implemented it cannot be undone as easily. Hence rolling out reforms is a process that ought to be taken with great caution by policy makers.

We have seen that taxpayers made use of the possibility to denounce almost to the full extent with an overall average propensity to denounce the TO of about 91.4%. Denouncing was most frequently used in Treatment 4 (98.6%), but no significantly different from Treatment 3 (87.4%) and Treatment 2 (90.0%). We thus do not find any evidence for reciprocity among TPs and TOs, which may partially be attributed to the fact that in our setting TPs who chose to denounce were granted partial anonymity. TOs were only informed that and by how many TPs they were denounced, but not exactly by whom. Hence, unless a TO was denounced by all (evading) TPs it was not possible to determine whether a particular TP chose to denounce or not, limiting the scope for retaliation (for example withholding
future cooperation). In contrast it has been argued that betrayal, such as denouncing, is associated with a moral or psychological cost (see also Coricelli et al., 2010).

We find surprising differences in behavioral reactions to the introduction, but not the removal, of a principal witness regulation across gender. Gender differences have been repeatedly demonstrated in various domains such as risk preferences, social preferences, lying behavior (Childs, 2012), and honesty (Muehlheusser et al., 2015). For example, Hasseldine and Hite (2003) study framing effects in tax compliance and find a significant frame by gender interaction indicating a stronger reaction to changes in framing for females. Although, aside from the standard observation that females are found to be less inclined to be corrupt or evade taxes (see for example, Kastlunger et al., 2010; Torgler and Valev, 2010), we did not expect such strong gender differences. In what follows, we will briefly discuss potential drivers of our surprising results. The experimental design adopted in our study included two main factors that potentially play a role in explaining female participants reactions to the implemented institutional change. The first ingredient is risk (to be fined) and the second one is the particular institutional setting adopted. For one, a general difference in risk attitudes across gender could potentially explain the significant drop in female tax compliance after the introduction of a PWR in Treatment 3, since by design the PWR sharply reduces the risk of deviant behavior. However, our results survive and remain highly significant when controlling for individual risk aversion attitudes (see the regression in Figure D.3 provided in Appendix A). For another, Lighthall et al. (2009) study how stress affects decision making under risk and find that overall males take more risk than females, but interestingly stress increases risk-taking for males, whereas females become more risk-averse. Assuming that a regime with PWR is perceived as less stressful for TPs as it offers a “safe way out”, the observed different reactions across gender in Treatment 3 might be explained by such an “inverted stress effect”. Unfortunately, we did not collect

\[\text{This result holds using both random-effects Tobit and GLS estimations with robust standard errors clustered on the individual level. See Figure D.3 in Appendix A for further details.}\]
any physical measure of stress (such as cortisol levels or heart rate) that would allow us to test this explanation, but we view this as an interesting avenue for future research.

An alternative explanation would be that the sudden institutional change affects females more strongly than males, which is in line with (Croson and Gneezy, 2009; Dreber et al., 2013) who argued that females are more sensitive to the contextual frame. There is a growing body of evidence on gender differences within the frame of choice under risk and strategic uncertainty arguing that females are generally more sensitive to the contextual frame. Hasseldine and Hite (2003) study framing effects in tax compliance and find a significant frame by gender interaction indicating a stronger reaction to changes in framing for females. In our context, the introduction of a PWR renders the TO formally responsible, hence creating a situation where the responsibility (and risk) is shared among TP and TO. Our evidence indicates that females strongly react to this new situation by a drastic drop in compliance, whereas males appear to be unaffected. It is important to note that the same does not hold for the removal of a principal witness regulation as this effect is not present in Treatment 4. We interpret this as evidence in support of the idea that gender effects might oftentimes stem from a higher sensitivity of females to the institutional environment as proposed by Croson and Gneezy (2009).

5. Conclusion and Policy Implications

To Do: Update Conclusion (JB)

Our work represents, to our knowledge, the first attempt to shed light on the effectiveness of the principal witness regulation utilizing a controlled laboratory setting. We nest a general tax evasion setting within a corruption framework, therefore adding a dimension of strategic interaction that allows us to capture a broader spectrum of tax evasion contexts such as custom duties. We provide evidence that in such a setting the effectiveness of new political measures heavily depends on history, that is the particular institutional frame in place before a change is introduced. Comparing a static setting without PWR (Treatment
1) to a static setting where PWR is available (Treatment 2), we find significantly higher
tax compliance by TPs under PWR. Had we not run additional dynamic settings involving
a regime change (Treatments 3 and 4), one might have easily come to the conclusion
that a regime change towards a system with PWR is recommendable. The data from our
dynamic treatments, however, shows a different picture. Here the introduction of such a
policy measure, given a history where it was absent, has a negative short-term effect of
decreased average tax compliance, but at the same time induces a reversal in the dynamic
adjustment over time. Before PWR was introduced tax compliance was at a steady decline,
whereas we observe a significant upwards trend following the regime change. One might
interpret this as a hint upon a potential positive long-term effect of a principal witness
regulation on tax compliance. Nevertheless we want to emphasize that the large effect one
might have expected from the difference between the static settings does not carry over
to the dynamic setting. This is also important from a normative point of view and poses
a question for future research: How can we improve tax compliance without introducing
negative consequences? We show that not only the final outcome/system one wishes to
achieve is relevant, but also the sequence of, potentially minor, intermediate reforms that
are implemented on the way should be taken into serious consideration. In particular, a
reform that proves to be successful in a particular country or institutional frame, will not
necessarily yield the same positive result when introduced in another environment where
the status quo/history of reforms differs dramatically. Interpreted on a more general level,
our results suggest that the success of political reforms may, at least to some extent, also
hinge upon the particular path taken and hence should be taken into account in policy
decision-making.

Appendix A. Summary statistics

Table [A.1] and [A.2] provides some average results on tax compliance, both attempted
and effective, the frequency of successful bribe exchange (collusion), the amount of bribes
paid, the smallest bribe accepted by a TO, the proportion of bribes accepted by the TO, and the propensity of taxpayers to denounce tax officers when given the chance.

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1 (noPwr)</th>
<th>Treatment 2 (Pwr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TaxDeclared</td>
<td>34.7</td>
<td>51.4</td>
</tr>
<tr>
<td>TaxPaid</td>
<td>50.2</td>
<td>69.8</td>
</tr>
<tr>
<td>BribesExchanged</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BribesOffered</td>
<td>47.9</td>
<td>54.4</td>
</tr>
<tr>
<td>MinBribeAccept</td>
<td>49.0</td>
<td>54.6</td>
</tr>
<tr>
<td>AcceptanceRate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denounce</td>
<td>-</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Table A.1: Summary statistics in percentages for treatments 1 and 2. The following terms are referred to as follows: taxes declared (TaxDeclared), taxes actually paid by the tax payer (TaxPaid), frequency of successful bribe exchanges that were offered and accepted (BribeExchange), bribes offered (BribesOffered), minimal bribes accepted by the public official (MinBribeAccept), and the frequency (in %) of denounce decisions made by the tax payer conditional on having been offered the chance to denounce the respective tax officer (Denounce).

Table A.1 indicates that mean declared taxes in Treatment 1 are 34.7 compared to 51.4 in Treatment 2 where a PWR was present. This difference is highly statistically significant ($p < 0.01$, Wilcoxon rank-sum test). Our results show that tax compliance is higher in a world with PWR compared to a world without.

---

22 In order to allow for easier interpretation, we use percentage values throughout the paper. Both for tax compliance and bribe payments, percentage values indicate the amount as a proportion of the tax required to pay (40 ECU) and the maximum amount of bribes that can be paid (30 ECU).
<table>
<thead>
<tr>
<th></th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rounds</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(noPwr-Pwr)</td>
<td>(Pwr-noPwr)</td>
</tr>
<tr>
<td>Rounds</td>
<td>1-10</td>
<td>11-20</td>
</tr>
<tr>
<td>TaxDeclared</td>
<td>50.4</td>
<td>37.2</td>
</tr>
<tr>
<td>TaxPaid</td>
<td>65.2</td>
<td>62.0</td>
</tr>
<tr>
<td>BribesExchanged</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BribesOffered</td>
<td>43.3</td>
<td>49.2</td>
</tr>
<tr>
<td>MinBribeAccept</td>
<td>48.3</td>
<td>54.3</td>
</tr>
<tr>
<td>AcceptanceRate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denounce</td>
<td>-</td>
<td>87.4</td>
</tr>
</tbody>
</table>

Table A.2: Summary statistics in percentages for treatments 3 and 4. The following terms are referred to as follows: taxes declared (TaxDeclared), taxes actually paid by the tax payer (TaxPaid), frequency of successful bribe exchanges that were offered and accepted (BribeExchange), bribes offered (BribesOffered), minimal bribes accepted by the public official (MinBribeAccept), and the frequency (in %) of denounce decisions made by the tax payer conditional on having been offered the chance to denounce the respective tax officer (Denounce).

**Appendix B. Further graphs**
Figure B.1

Figure B.2
Figure B.3
Appendix C. Examination of the evolution of behavior across periods

Appendix C.1. Evolution of tax compliance

Figure C.4: Trends in average taxes declared by tax payer without PWR (graph top left, solid line) and with PWR (graph top right, solid line), as well as average taxes paid by tax payer without PWR (top left, dashed line) and with PWR (top right, dashed line) across treatments. We also provide the average bribe acceptance rate by tax officer with PWR (graph bottom left) and without PWR (graph bottom right).

The dashed lines in Figure C.4 show the evolution of the amount of taxes actually collected over the course of the experiment. We observe a significant decrease in tax yield collected over time in both static treatments independent of whether PWR is present ($p < 0.01$, Wald test). Since our tax evasion experiment is nested within a corruption framework, the amount of taxes actually paid is the result of tax payers as well as tax officers decision. On the one hand, higher acceptance rates for bribes (and hence flawed tax reports) on the side of the TO will lead to less taxes being collected. On the other hand, lower tax declarations by the TP might also explain why tax yield collected is at a constant decline. We find that in treatments 1 and 2 acceptance rates are increasing over

---

23 We examine all differences in trends using post-hoc estimation Wald tests.
time, while tax compliance is decreasing as Figure C.4 illustrates. Hence the combination of higher acceptance rates and lower compliance leads to a constant decline in efficiency.

Appendix C.2. Evolution of tax compliance by gender

Examining the evolution of tax (non-) compliance by gender, Figure C.5 illustrates tax compliance over rounds for the two treatments involving a regime change separately for males and females. In Treatment 3, we observe a highly significant change in slopes for females (\( p < 0.01, \) Wald test) and males (\( p < 0.01, \) Wald test) respectively. Here, the change in slopes also suggests that females and males react differently to the change in the institutional frame: females resort to a stationary high tax evasion behavior, while males gradually converge towards higher tax compliance. In Treatment 4, we find no difference in trends for both, females (\( p = 0.44, \) Wald test) and males (\( p = 0.85, \) Wald test).

*Figure C.5: Trends in average taxes declared by tax payer (solid line) and average taxes paid by tax payer (dashed line) for Treatment 3 (top) and Treatment 3 (bottom) broken down by gender. The vertical line indicates the regime change between round 10 and 11.*

Appendix C.3. Institutional history and evolution of tax compliance

Figure C.6 shows the evolution of tax compliance behavior over time in the four different treatments. Treatment 1 (noPwr) and Treatment 2 (Pwr) allow us to study the effect of a PWR on tax compliance when there is no institutional change. We observe a constant
decline of tax compliance in Treatments 1 and 2 that did not feature an institutional change. Surprisingly, the picture is drastically different for Treatment 3 (noPwr-Pwr). Here, we see a steady decline in tax compliance prior to the introduction of a PWR, and then observe a drastic and highly significant break and change in slopes following the institutional change ($p < 0.01$, Wald test). In addition, the drop in tax compliance following the institutional shock is highly significant ($p < 0.01$), indicating that the institutional transition to PWR breeds tax evasion. Surprisingly and in contrast to existing literature, we find that this drop in compliance is driven by female participants. This is also insofar surprising, as that we, in contrast, do not observe a similar and/or significant reversal in trends in Treatment 4 ($p = 0.75$, Wald test), where the order of institutional change is reversed. That is, an institutional environment with PWR is replaced by an institutional environment in which PWR has been abandoned. Instead, we merely observe a restart effect that results in an initial jump of taxes declared but the same gradual decline in compliance as in the first 10 rounds with PWR.

Overall, these results suggest that it is not the mere institutional shock that causes significant changes in compliance behavior, but rather the order in which the institutional transition occurs.

Appendix C.4. Institutional history and evolution of bribe acceptance

Appendix C.4.1. Behavior of Tax Officers

Figure C.6 also shows the evolution of the proportion of bribes accepted by the TO over time in treatments 3 and 4. For a fixed institutional environment the slopes are positive and significantly different from 0 ($p < 0.01$, Wald test) in both treatments indicating that collusion increases over time.

Interestingly, although collusion increases constantly for a fixed institutional frame, the

\[24\text{Restart effects have been observed in various repeated environments such as repeated public good games (Andreoni, 1988).}\]
Figure C.6: The interpretation of this graph and the solid or dashed lines is identical to the interpretation in Figure C.4. The vertical line indicates the regime change between round 10 and 11. Treatment 3 (4) depicts the behavior without PWR (with PWR) left and the behavior with PWR (without PWR) to the right of the vertical line.

Figure C.7: Proportion of bribes accepted by the TOs.
findings in Treatment 3 indicate that the average acceptance rate of bribes drops from 66% to 55% following the transition to an institutional environment with existing PWR. While this difference is not statistically significant ($p = 0.20$, Wilcoxon rank-sum test) due to lack of statistical power resulting from a limited amount of tax officers in our experiment, we indeed find a significantly lower acceptance rate in the short run, that is during the first five rounds with PWR ($p = 0.01$, Wilcoxon rank-sum test). This result suggests that the introduction of a PWR might, at least temporary, dampen corruption by decreasing the frequency of successful corrupt collaboration. This partially supports that the introduction of a leniency mechanism can hinder (at least temporary) collusion, as the higher rejection rates indicate.

*** Do a similar test for collusion (last 5 rounds, first 5 rounds)? In your line graph for collusion we see a drop in Treatment 2 although the averages with and without PWR are the same.

Again, the reversed pattern cannot be observed in Treatment 4 (Pwr-NoPwr). Although acceptance rates are increasing the removal of the PWR does not result in a significant upwards shift in acceptance rate. We interpret this as evidence indicating that this effect is not driven by the presence of a PWR per se, but crucially depends on the initial system that is in place before the reform is implemented.

The next graph shows the acceptance rate conditional of the size of offered bribe for each Treatment. We see that higher bribes are accepted more frequently across all four Treatments.

*** Potentially include minimal bribe acceptance threshold
Appendix D. Regression Analysis

*** Old
Our dependent variable TaxDeclared depicts the amount of declared taxes. Our independent variables are: BRIBE = amount of bribe paid by the tax payer; Sex = 1 if male; HL\textsubscript{Value} = Holt and Laury risk aversion measure, with a higher number indicating a later switching point and thus more risk-aversion; l1\textsubscript{ACCEPTED} = 1 if bribe was accepted by PO in the previous round; l1\textsubscript{CONTROL} = 1 if TP was audited in previous round; l1\textsubscript{FINE} = fine paid by TP in the previous round.

**Acknowledgements**

We gratefully acknowledge helpful comments and suggestions by Carlos Alós-Ferrer, Simon Gächter, Heike-Hennig Schmidt, and participants at the Workshop for Experimental Psychology at the University of Pennsylvania and the 7th International Conference of the French Association of Experimental Economics. We would like to thank Marco Tecilla for the technical support.
References


